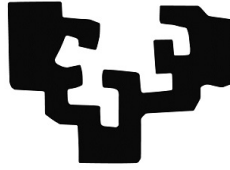


eman ta zabal zazu



Universidad del País Vasco Euskal Herriko Unibertsitatea

Konputazio Zientzia eta Adimen Artifiziala saila
Informatika Fakultatea

**Multimedia edukien ulerpen semantikorako
ekarpen metodologikoak: irudien behe-mailako
analisitik bideoen ekintzen sailkapenera**

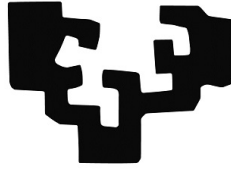
Naiara Aginako Bengoa

Zuzendariak:

Basilio Sierra Araujo
Julian Florez Esnal

Donostia, 2017

eman ta zabal zazu



Universidad del País Vasco Euskal Herriko Unibertsitatea

Konputazio Zientzia eta Adimen Artifiziala saila
Informatika Fakultatea

**Multimedia edukien ulerpen semantikorako
ekarpen metodologikoak: irudien behe-mailako
analisitik bideoen ekintzen sailkapenera**

Naiara Aginako Bengoak Basilio Sierra
Araujo eta Julian Florez Esnalen
zuzendaritzapean egindako
ikerketa-txostena, Euskal Herriko
Unibertsitatean Informatikan Doktore
titulua eskuratzeko aurkeztua

Donostia, 2017ko Martxoa

Naiara Aginako Bengoa

*Multimedia edukien ulerpen semantikorako ekarpen metodologikoak: irudien behe-mailako
analisitik bideoen ekintzen sailkapenera*

Zuzendariak: Basilio Sierra Araujo eta Julian Florez Esnal

Euskal Herriko Unibertsitatea (UPV/EHU)

Konputazio Zientzia eta Adimen Artifiziala saila

RSAIT (Robotika eta Sistema Autonomoen Ikerketa Taldea)

Donostia - San Sebastián

Laburpena

Azken urtetan munduan sortzen den multimedia eduki digital kopurua izugarri hazi da. Eduki mota hauen artean irudiak eta bideoak dira nagusi. Azken hauek dira batez ere neurrigabeko hazkunde bat izaten ari direnak. Gaur egun, komunikazio digitalen oinarritzko unitate bihurtu dira eta hazkunde esponentzial hau hurrengo urteetan mantenduko dela aurreikusten dute gai honetako adituek.

Ondorioz, eduki guzti horien metatze, kudeaketa eta analisirako metodo eraginkorrak behar dira. Ikerketa lan honek irudien eta bideoen analisia ahalbidetzen duten metodoen ikasketa eta garapena du helburu nagusitzat. Metodo hauek irudi eta bideoen ulerpen semantikorako beharrezkoak diren oinarriak finkatzen dituzte. Edukien ulerpen semantiko honek etiketatu automatikoa ahalbidetzen du eta honek garrantzia handia dauka bi ikuspuntutatik. Batetik, edukien sortzaileen ikuspuntutik, edukiak sortzen diren neurrian automatikoki anotatu badaitezke, metaketa prozesua etiketa hauetan oinarrituz egitea posible izango litzateke eta ondorioz, metaketa eta bilaketa prozesu eraginkorragoak sortzeko aukera egongo litzateke. Bestetik, edukien erabiltzaileen ikuspuntutik, edukien bilaketa kontzeptu semantikoetan egiteko aukera egongo litzateke eta honek asko normalizatzen du konputagailu-pertsona komunikazioa; askoz naturalagoa bilakatzen da bien arteko elkarrekintza.

Gaur egun, multimedia edukien gordailu handiena Internet dela argi dago. Bertan pilatzen dira munduan sortzen diren eduki gehienak. Eduki hauen berezitasun nagusia hauen heterogeneotasuna da. Edukiak edozein alorretakoak izan daitezke, hau da, ez dira domeinu finko baten barnean kokatzen. Ondorioz, ezin dira domeinura egokitutako estrategiak erabili eta honek asko zailtzen du analisia. Bestalde badaude, domeinu zehatz baten barruan kokatutako multimedia edukiak eta azken izan dira ikerketa honen helburu.

Txosten honetan deskribatzen den ikerketa lana konputagailu bidezko ikusmenaren barruan kokatzen da. Aurretik aipatu bezala, egindako ikerketak, domeinu finko eta itxi bateko edukiaren azterketarako metodoen garapenean oinarritzen dira. Hiru izan dira ikerketa lerro nagusiak. Lehenengoa, irudien ulerpen semantikorako behe-mailako deskriptoreen erabilera eta etiketatu semantikora bideratutako metodoen garapenera bideratutako lerroa izan da. Bigarren lerroa, lehenengoaren helburu

berdina izanik, ikasketa automatikoko metodoen erabilera aztertzeraz bideratu da. Sailkatzaileen erabilerak, izugarri zabaltzen du aztertu daitezkeen arazoaren espektroa. Jadanik, ateratako kontzeptu horiek ez dira hain sinpleak izan behar eta irudi mota desberdin gehiagorekin lan egiteko kapazitatea gehitzen diote garatutako sistemari. Azkenik, hirugarren lerroa, bideoen analitikara bideratu da. Bideoen analitika honetan dago gaur egun erroka handiena. Dokumentu honetan bideoen ekintzen ezagutzarako irudien behe-mailako ezaugarrietan eta sailkatzaileetan oinarritutako metodologia berri bat aurkezten da.

Irudi eta bideoen ulermen semantiko automatikorako bidea luzea da oraindik baina ikerketa honetan urrats esangarri batzuk eman dira. Horregatik, etorkizunean egin beharreko lana emandako urrats hauetatik abiatuko da.

Abstract

In the last years, the amount of digital content that is produced worldwide has grown exponentially. Concretely, digital images and videos are the main representatives of this growth. They have become the core unit of all the digital communications and the experts in the field foretell that this trend will continue in the next few years.

Therefore, there is a real need of effective and stable methods for the storage, management and analysis of this huge volume of content. The presented dissertation work focuses on the research and development of analysis methods for images and videos. The main objective is to establish the pillars for the semantic understanding of images and videos. Remark that understanding the analysed content permits the automatic labelling of it, which incurs in benefits from different perspectives. On the one hand, content producers could automatically label all their information while they are generating it. This fact permits the adoption of more effective storage strategies and content retrieval systems. On the other hand, content consumers gain the possibility of searching content using semantic concepts, ergo, natural language terms. In consequence, Human-Computer Interaction (HCI) becomes more natural.

Internet is nowadays the biggest warehouse of the digital content. Most of the content that is everyday produced in the world is housed there and the main characteristic of this content is its heterogeneity. Content can proceed from very different fields; in other words, they don't belong to a unique domain. Hence, domain specific strategies can't be applied and this hinders the analysis of the content. In order to tackle this, lot of research has been done in effective methods for the domain recognition. However, these methods are commonly based on the analysis of the content itself which ends in a Vicious circle.

Computer Vision drives all the research results presented in this dissertation work. Main contributions rely on the fact that designed and developed methods are domain specific; even though, some domain-agnostic methods has also been studied.

Main contributions can be divided into three principal research lines. First research line includes the analysis and development of methods based on low-level descriptors

for the inference of simple image semantic concepts. Methods for the automatic labelling of images are also the target. Second line focuses on the addition of machine learning strategies for image classification, recognition and understanding. The inclusion of classification algorithms permits broadening the spectrum of the resolved issues. The semantic concepts that can be extracted from the images are more complex, or high-level, and developed solutions can deal with a greater variety of images.

Third and last research line is focused on video analytics. This line benefits from all the work accomplished in the previous lines. Even though previous developments can't be directly applied as timing of videos is a highly relevant variable, learned conclusions are very relevant when designing and developing the adapted solutions. Nowadays, video analytics is one of the most challenging tasks within Computer Vision. This dissertation work presents a new methodology based on low-level descriptors and classifiers for action recognition in videos.

Automatic image and video understanding is still an open issue within research community. Still there is a lot of work to do but some steps has been made in order to overcome the actual computer vision challenges. Future work will imply to continue in the path already started.

Eskerrak

Dokumentu honetan idatzitako lerro gehienak buruarekin idatziko ditut beraz kapitulu hau bihotzarekin idatziko dut. Eta nire bihotzean lehena zu zaude Ane. Amatxoren azkenengo txanpa ilusioz betetzeko heldu zinen eta zure ondoan igarotzen dudan segundo bakoitzean lortzen duzu zure helburua. Eta hori, amaren energia xurgatzen duzula. Marisorgin halakoa!

Tesiaren prozesu luze honetan jende pila eduki dut ondoan, batzuk lagun besteak kide baina denak lagundu naute aurrera egiten modu batean edo bestean. Lehenengo aipamena Vicomtecheko jendearentzat, bertan eman baitzituen nire ikerketa lanak bere lehen pausuak. Bertako lankideak lagun bihurtu ziren eta orain nire bizitzako parte garrantzitsu bat dira. Zerrenda ederra daukat guztien izenekin, baina espero dut ZUEK badakizuela nortzuk zareten (Eii Tiburones). Haiei esker sozializatu nuen Donostian, niretzako arrotza zen hiri batean eta kriston esperientziak bizi izan ditut. Oraindik ez dut danborradarekin ateratzeko aukerarik izan baina helduko da eguna. Eider, musu bero bat zutzat, beti Vicomtechekoekin nahastuta zabilta baina oraingoa espresuki bidali nahi dizut zuri musua.

Badaude *tiburoi* horietatik aparte Vicomtecheko beste lagun asko: Gorka, Aiala, Jon Haitz, Igor, Juan, Iñaki, Mariate, Marco, Montse, Mikel M., Mikel Z., eta abar luze bat. Nire ikerketaren lehen etxea utzi eta unibertsitatara egin nuenean salto pertsona bikain gehiago ezagutzeko aukera izan nuen, bereziki nire Matematika Aplikatu Saileko kideak. Milesker zuei ere hain harrera ona egitegatik. Espero luzerako hemen geratzea.

Eta nola ez, prozesu honetan ez ezik bizitza osoan alboan izan ditudan familiako kideei: Aita, Ioritz, izeba, osaba, Amaia eta Mikel; Lorea, Iker, Nerea, Luisma, Esteban eta Mari Jose. Ni naizena banaiz zuengatik da ere. Hori ez inoiz ahaztu. Gasteizko nire lagunentzat, betiko koadrilarentzat, ere aipamen berezia eta musu erraldoi bat. Urrun zaudete baina egunero zaudete nire ondoan.

Milesker zuri ere Julian. Beti hor egotegatik, zure izateko era gertuagatik eta bizitzan balio duten gauzengatik arduratzeagatik. Basi, zu ezagutzeak eta zure laguntzak eman zidan azkenengo bultzada. Zure animoak eta zure lanaren par-

te nintzela sentiaraztea aurrera egitera animatu ninduen. Milesker bioi bihotz bihotzez.

Eta nola ez, ZU. Ikerketa honen bidean ezagutu zintudan. Bide horren hasieran, nahiko eztabaidatu genuen arau trigonometrikoei buruz (ze burugogorra zinen). Beranduago beste gauza askori buruz eta oraindik eztabaidatzen jarraitzen dugu (burugogor izatearena ez zaizu pasa). Hala ere, irribarrez betetzen didazu bizitza. Zu eta Ane gure ohean elkarrekin ikustea da egunero ohetik altxatzera bultzatzen nauena. Aurrera jarraitzera. Maite zaitut.

Azkenik amatxo, azkenengo urtean ikasi dut zer den nigatik sentitzen duzuna, zer den kondiziorik gabeko maitasuna, betirako hor izango dudana. Pila eskertzen dizut nigatik egunero egiten duzuna. Mila bizi biziko banitu ere zu aukeratuko zintuzket beti.

*Mila esker bihotzez,
Naiara*

*Si alguna vez no te dan la sonrisa esperada,
sé generoso y da la tuya,
porque nadie necesita una sonrisa
como aquél que no sabe sonreír a los demás.*

Aurkibidea

1	SARRERA	1
1.1	Motibazioa	1
1.2	Testuingurua	4
1.2.1	Vicomtech-IK4	4
1.2.2	Robotika eta Sistema Autonomoen Ikerketa Taldea	6
1.3	Ikerketaren oinarriak	7
1.3.1	Konputagailu bidezko ikusmena (<i>Computer Vision</i>)	7
1.3.2	Ikasketa automatikoa (<i>Machine Learning</i>)	10
1.4	Ikerketa lerro nagusiak	12
1.4.1	Irudien ulerpen semantikorako behe-mailako analisia	13
1.4.2	Irudien ulerpenerako ikasketa automatikoko metodoen erabilera	14
1.4.3	Bideo analitika	15
1.4.4	Ikerketaren ekarpen nagusiak	17
1.5	Argitalpenak	20
1.6	Txostenaren antolaketa	22
2	IKERKETAREN EMAITZAK	25
2.1	Irudien ulerpen semantikorako behe-mailako analisia	25
2.1.1	Web-Based Supervised Thematic Mapping	29
2.1.2	Visual Processing of Geographic and Environmental Information in the Basque Country: Two Basque Case Studies	43
2.1.3	Weather analysis system based on sky images taken from the Earth	55
2.1.4	The COST292 experimental framework for RUSHES task in TRECVID 2008	63
2.2	Irudien ulerpenerako ikasketa automatikoko metodoen erabilera	71
2.2.1	Iris Matching by means of Machine Learning Paradigms: a new Approach to Dissimilarity Computation	75
2.2.2	Periocular and iris local descriptors for identity verification in mobile applications	83
2.2.3	Identification of plant species on large botanical image datasets	95
2.3	Bideo analitika	105
2.3.1	Method of detection and recognition of logos in a video data stream	109

2.3.2	Method for detecting the point of impact of a ball in sports events	121
2.3.3	Accurate ball trajectory tracking and 3D visualization for computer-assisted sports broadcast	131
2.3.4	Machine Learning for Video Action Recognition: a Computer Vision Approach	157
3	ONDORIOAK ETA ETORKIZUNERAKO LANA	175
3.1	Ondorioak	175
3.2	Etorkizunerako lana	176
	ERANSKINAK	178
A	I+G proiektuak	181
A.1	RUSHES	181
A.2	SIAM	182
A.3	GRAFEMA	183
A.4	SKEYE	184
A.5	SIRA	185
A.6	CAPER	185
A.7	P-REACT	186
A.8	BEGIRA	188
B	Beste argitalpen batzuk	189
	Bibliografia	195

Irudien zerrenda

1.1	Minutuero munduan sortzen den datu kopurua [Jos16]	2
1.2	Teledetekzio irudi baten kontzeptu semantikoen sailkapena	8
1.3	Irudien ezaugarri lokalen lorpena	9
1.4	<i>Imagnet</i> datu basearen irudien ez-gainbegiratutako sailkapena	11
1.5	Irudi baten segmentazio semantikoa [Sha]	14
1.6	Bideo zaintza merkaturen aurreikuspenak 2021. urterako [RM16]	16
1.7	Aurkeztutako ikerketa lanaren ekarpen nagusiak	17
2.1	Teledetekzio irudien sailkapen tematikoa	26
2.2	Iris irudien behe-mailako deskriptoreak ateratzeko prozesua	72
2.3	Landareen sailkapenerako garatutako fusio metodologia	73
2.4	Pilotaren segmentazioa eta jarraipena lurrarekiko ukitze puntua kalku- latzeko	106
2.5	Pilotaren ibilbidearen berreraiketa	106
A.1	<i>RUSHES</i> plataformaren prozesuen fluxua	182
A.2	<i>Skeye</i> proiektuan lortutako zero irudien segmentazioa	184
A.3	P-REACT plataforma	187

SARRERA

1.1 Motibazioa

Ikerketa lan hau hasi zenetik munduan sortzen den eduki digitalaren kopurua esponontzialki hazi da. Honetaz ohartzeko, nahikoa da minutuero munduan zehar sortzen den datu kopuruan erreparatzea (ikus 1.1 irudia). Honen ondorioz, hauen kudeaketa eraginkorra erronka bat bilakatu da bai industria bai ikerketa mundurako. Azken urteetan hainbeste aditzera ematen ari den Big Data terminoa datu hazkunde honetatik sortu zen.

Datu hauen barruan irudiak eta bideoak aurkitzen dira. Azken hauek gero eta garrantzia handiagoa hartzen ari dira, internet, sare sozial eta gaur egungo gailu mugikorrek eskeintzen dituzten aukerengatik. Ciscok aurreratzen duenez, interneteko trafikoaren %80a bideoak izango dira 2019an [Car15]. Beraz argi dago, eduki multimedia guzti hau prozesatzeko teknikak beharrezkoak direla.

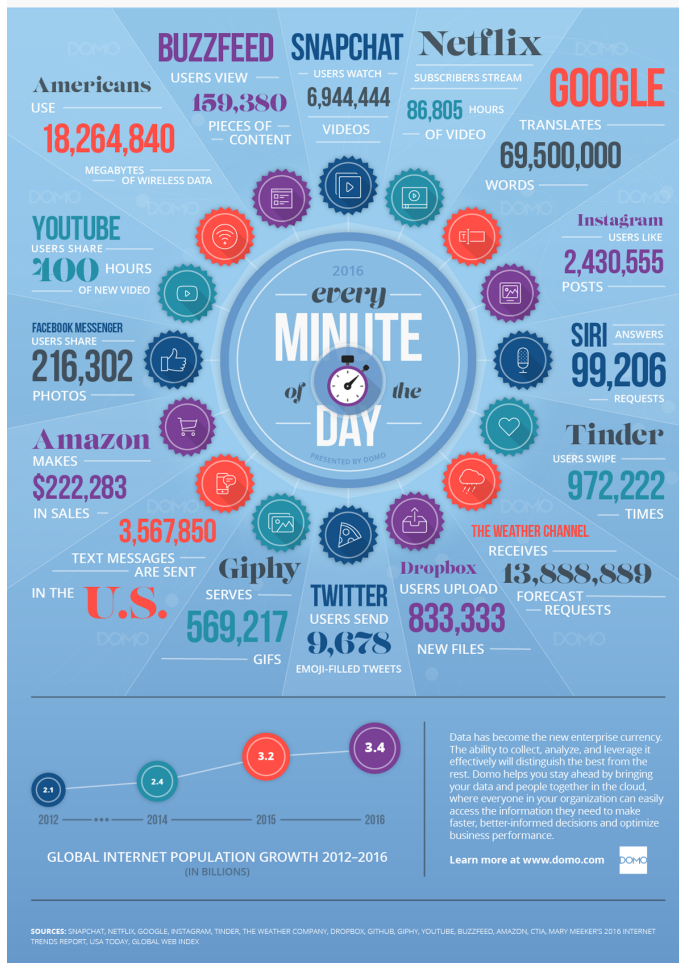
Konputagailu bidezko ikusmenaren (Computer Vision) zientziak irudi eta bideoen ulerpenera du helburutzat, halaber, burmuinak betsareak jasotako informazioarekin egiten duena. Hau da, irudiaren analisisa egin ondoren irudi horretatik ezagutza ateratzeko tresnak biltzen dituen zientzia da. Tresna hauen artean irudi prozesamenduko metodoak aurkitzen dira. Horretaz gain, ikasketa automatikoko zientziaren barruan aurkitzen diren algoritmoak ere erabiltzen dira ikusmenaren arazoei irtenbidea emateko. Bi zientzia hauen arteko elkarrekintza oso aberasgarria suertatzen da ulerpen automatikorako.

Teknologiaren aurrerapenengatik eta sortutako beharengatik, konputagailu bidezko ikusmenaren helburuak asko aldatu dira azken 50 urteetan [Sze10],. Azkenengo hamarkadan, irudien ulerpen semantikoaren ikerketa izan da batez ere hazkunderik handiena izan duen ataza. Orain dela 30 urte Treisman [TG80], Marr [Aco85] eta Biederman-en [Bie87] lanek kontzeptu honek irudien informazioaren prozesamenduan zuen lekua zehazten ahalegindu ziren. Hauen ustez, irudien ulerpenaren prozesua bi noranzko prozesua da. Batetik, behetik-gorako prozesua, hau da, irudian hasi eta honen behe-mailako analisisa egin adierazgarriak diren ezaugarriak ateratzeko. Eta bestetik, goitik-beherako prozesua, zeinetan behe-mailako informazioari forma eta zentzua emateko arauak eta ezagutzak ezartzen diren. Hau da hain zuzen



DATA NEVER SLEEPS 4.0

How much data is generated every minute? In the fourth annual edition of Data Never Sleeps, newcomers like Giphy and Facebook Messenger illustrate the rise of our multimedia messaging obsession, while veterans like YouTube and Snapchat highlight our insatiable appetite for video. Just how many GIFs, videos, and emoji-filled Tweets flood the Internet every minute? See for yourself below.



Irudia 1.1.: Minutuero munduan sortzen den datu kopurua [Jos16]

ere goi-mailako prozesamendu bezala ezagutzen dena. Bi noranzko hauek, bretxa semantikoa edo semantic-gap delakoa, saihestea dute helburu irudien ulerpena lortu ahal izateko.

Behe-mailako prozesamendua hasieratik argi zegoen kontzeptua izan da eta ikerketa lan guztien oinarria zen. 80. hamarkadako argitalpen hauetan adierazten den bezala, [Der90], [VVV88] eta [NL84], pixelen behe-mailako analisia da ikusmenaren muina. Ikertzaileek bazekiten zertan zetzan pixelen azterketa; goi-mailako prozesamendua berriz, kontzeptu askoz nahasiagoa izan da. Nahasmen hori ordenatzeko intentzioarekin, domeinuen ezagutza ezarri zen, goi-mailako ezagutza domeinu bakoitzean aditua izateko intentzioarekin. Liu-k [Liu+07] bere argitalpenean laburbiltzen duen bezala, goi-mailako analisia domeinuarekiko dependentzia duten metodoetan oinarritu izan da orokorrean. Horregatik, gaur egun mediku-irudiak, irudi infragorriak, teledetekzioko irudiak, radar irudiak, eta abar luze batean banatzen da irudien prozesaketaren arazoa. Modu honetan, irudiaren ulerpen semantikoa erraztu egiten da domeinu batera moldatutako metodoak baitira.

Baina domeinu finko batean kategorizatu daitezkeen irudiak asko izan arren, gaur egun, Internet dago. Aurretik komentatu den bezala, erabiltzaileek Internetera igotzen dituzten irudi eta bideo kopurua izugarri hazi da, momentu honetan hazten ari da eta hazten jarraituko du. Gehienetan, eduki guzti horren kontestua ezezaguna eta beraz, hain garrantzitsua den domeinuaren ezaguera galdu egiten da. Honen ondorioz, irudien ulerpen semantikoa lortzea askoz prozesu zailagoa da.

Domeinuaren garrantzia ikusirik, ikerketa asko domeinu hori finkatzeko metodoen garapenera bideratu dira. Adibidez, Salemi [SHS10] eta Lazebnik [LSP06a] egindako lanetan, domeinua ezartzeko metodoak aurkezten dira. Hala ere, bi lan hauetan, domeinu orokor batetako irudietatik domeinu zehatzago baten definizio egiten dute. Lehenengo kasuan, kirolletako irudietan oinarritzen dira, kirol konkretu batetan kokatzeko irudia eta bigarrenaren kasuan, naturako irudietara moldatutako soluzioa aurkezten dute. Olaizola [Ola+14] eta Bergamo-k [BT10] aurkeztutako lanetan ere domeinua zehazteko metodo eraginkorrak azaltzen dira. Azpimarratu badirela aurreko ezagutzarik gabe irudiaren eszena klasifikatzen dituzten metodoak; hauen artean aipagarrienak Torralba [OT01] eta Boschek [BZM06] aurkeztutako ikerketa lanak dira.

Guzti honekin lotuta beste kontzeptu berri bat: ingelesez CBIR (Content Based Image Retrieval) eta euskaraz edukian oinarritutako irudien berreskurapena bezala ezagutzen dena. Multimedia datu-base handietan gordetzen diren irudien berreskurapenerako behar diren ordenagailu bidezko ikusmenaren teknikak garatzea du helburuetako bat. Baina betiere berreskurapena, irudiaren berezko edukiaren azterketatik abiatuz. Termino hau 90.eko hamarkadan jaio zen eta nahiz eta ikerketa

asko egin den honen inguruan, lortutako emaitzek ez dute aurrerapauso handirik aurkeztu. Era honetako sistema gehienak irudien kolorea, barneko formak, testura, banaketa espaziala, eta abarren azterketan oinarritzen dira. [Jin+03], [MKS03], [ZH03] eta [Qia+16] argitalpenetan aurkezten diren sistemak berriz, erabiltzailearen berrelikaduran oinarritzen dira ikasteko eta emaitzen doiketa egin ahal izateko. Beste askok berriz, [Scl+99], [PG17], [MAJ13], irudiaren informazioa ez ezik, honekin batera metatzen den testuaren azterketa gehitzen diote analisiari irudien berreskurapen doitasuna hobetzeko.

Laburbilduz, irudi eta bideoen semantika ulertzeko bidean lan handia dago egiteke eta honetarako informazioaren behe-mailako azterketa nahitaezkoa da. Txosten honetan deskribatzen den lana gai horretan egindako ikerketa zorrotza aditzera ematen du. Lan sakona da, gaiak eskatzen duen modukoa, baina etorkizunerako geratzen den lana oraindik lan sakonagoa dakarrelako ondoriora eraman gaitu egindako ikerketak.

1.2 Testuingurua

Txosten honetako ikerketa lana, Vicomtech-IK4 ikerketa aplikatuko zentroan gartua izan da EHUko Robotika eta Sistema Autonomoen Ikerketa Taldearen (RSAIT) laguntzarekin. Aurrera eramandako ikerketak industriaren beharrekin batera eboluzionatu du; honek zituen eta gaur egun oraindik dituen arazoei aurre egiteko teknologiak garatzeko asmoz.

Ikerketa aplikatuko zentro batean garatutako ikerketa izatearen ondorioz, orokorrean, soluzio zehatz eta teknologikoki inplementagarria bilatu da. Hala ere, soluzio horien bilaketa prozesuak, oinarritzko ikerketarekin harreman estuago bat zeuzkaten ikerketetan sakontzeko beharra ekarri du. Bi bide horien eragina argi ikus daiteke lan honetan aurkezten diren publikazioetan: batetik patenteak, zeinak proiektu industrialetan aurrera eramandako soluzioen jabetza intelektuala babesteko idatzi diren eta bestetik, bai konferentzietan edota aldizkarietan aurkeztutako argitalpen zientifikoak, zeinetan aurrera eramandako ikerketaren emaitzak aditzera ematen diren.

1.2.1 Vicomtech-IK4

Dokumentu honetan deskribatutako ikerketa lan gehienak Vicomtech-IK4 zentroan egin dira. Vicomtech-IK4 Donostiako Parke Teknologikoan kokatuta dagoen ikerketa aplikatuko zentroa da, hau da, oinarritzko ikerketa eta ikerketa aplikatua bateratzen

ditu garatutako teknologiak industriara transferitzeko. Enpresen eta erakundeen berrikuntza-beharrei erantzutea da bere helburua nagusia.

Orotar, Vicomtech-IK4-k ikus-elkarrekintzako eta komunikazioko multimedia-teknologiak garatzen ditu. Industriekin, unibertsitateekin eta beste zentro teknologiko batzuekin du lankidetzeta estua, horien osagarri delarik.

Urtean ehunka ikerketa-proiektu gauzatzen ditu (industriekin eta tokiko, estatuko eta Europako administrazioekin), 100 ikertzaile inguru dauzka eta 7 milioi euro fakturatzen ditu urtero. Gainera, ikerketetan lortutako emaitzak garrantzizko publikazioetan eta patenteen bidez aditzera ematen ditu. Gaur egun, Vicomtech-IK4 sei sail desberdinetan banatzen da, hauetako bakoitza dagokion sektoreko industriari erantzunak emateko sortua:

- Industry and Advanced Manufacturing
- Digital Media
- Ahotsaren eta Lengoia Naturalaren Prozesamendurako Teknologiak
- Smart Environment & Energy
- Garraio-sistema adimendunak eta ingeniarietza
- eOsasuna eta biomedikuntza aplikazioak

Lan honetan aurkezten den lana gaur egun *Digital Media - Euskarri Interaktiboak, Ikus-Entzunezko Ekoizpenak eta Multimedia Edukien Kudeaketa* bezala ezagutzen den sailaren barruan burutu da. Sail honen egungo helburua ikus-entzunezkoen eta multimediaren sektorean lan egiten duten eragileei irtenbide teknologikoak eskaintzea da. Hurrengo ildo-nagusiak ditu:

- **Komunikazio Teknologia Elkarreragileak.** Informazio Gizartean gailu mugikorretarako joera ikaragarri hazi da azken urteotan, eta ondorioz multimedia-fluxuak eta elkarrekintza-gaitasunak erraztasunez integratzen dituzten irtenbide teknologikoak behar dira. Zerbitzu hauek erabiltzaileen esperientzia aberastu eta hobetzeko balio dute.
- **Ekoizpeneko Teknologi Aurreratuak.** Eduki digitalen ekoizpen prozesuek azkar barneratu behar dituzte kontsumo elektronikan eta mundu profesionalean ematen diren aurrerapen teknologikoak. Egun ekoizpen katean ematen ari

diren aurrerakadek teknologia eskaera berriak dakartzate eta Vicomtech-IK4k bere I+G jarduera horretan garatzen du, bereziki antzeko edukien sorrera, atzemate, kudeaketa, prozesatze/manipulazio, igorpen, jasotze eta erreproduzio teknologien garapenerako.

- **Multimedia-Edukien Analisia eta Kudeaketa.** Gaur egungo baliabide digitalak gero eta heterogeneoak dira eta gainera gero eta lotura estuagoak dituzte. Horregatik, baliabide digitalen kudeaketa funtsezko erronka da eta horretarako multimedia-edukien analisi automatikoa ezinbestekoa bihurtu da.

Txosten honetan aurkeztutako ikerketa, multimedia-edukien, eta zehatzago irudi eta bideoen analisisan jorratu baina beste bi ildoekin ere harreman estua dauka. Edukiaren ezagutza izateak erabiltzaileei emandako zerbitzua aberasteko edo ekoizpen prozesuetan laguntzeko oso baliagarria izan baitaiteke.

Lan hau sailean azken urteetan aurrera eraman diren oinarrizko ikerketako proiektu eta proiektu industrialen emaitza da. Gainera, sailean aurkeztutako [Ola12] A FRAMEWORK FOR CONTENT BASED SEMANTIC INFORMATION EXTRACTION FROM MULTIMEDIA CONTENTS tesiaren jarraipena da. Ikerketa lerro honen jarraitasunak, Vicomtech-IK4-ren barruan multimedia edukien ikerketak duen garrantzia eta pisua azpimarratzen du.

1.2.2 Robotika eta Sistema Autonomoen Ikerketa Taldea

Azkenengo urteetako ikerketa Robotika eta Sistema Autonomoen Ikerketa Taldeko (RSAIT) kideekin batera egin da. Talde hau Euskal Herriko Unibertsitateko Informatika Fakultatean jaio zen eta bertan egiten du ikerkuntza. Taldearen ikerkuntzaren lerro nagusia robotika mugikorra da. Robotika mugikorrean aplikatzen dira estatistika eta ikasketa automatikoko teknikak, roboten autonomia maila areagotzeko.

RSAIT taldeak honako gaiak uztartzen ditu:

- Robot autonomoen esplorazioa eta nabigazioa
- Ikasketa automatikoa (Machine Learning)
- Gizaki eta roboten arteko elkarrekintza
- Ordenagailu bidezko ikusmena (Computer Vision)
- Estatistika

Ordenagailu bidezko ikusmenaren gaia, roboten irudien analisisira bideratzen da. Horretaz gain, eta ikerketa lan honen ildoan, irisaren detekzio eta ezagupenarekin lotuta dauden ikerketak garatu dira. Taldeak Ikasketa Automatikoko metodoei buruz duen ezagutzari esker, sailkapenarekin lotura duten ikerketa proiektuetan Vicomtech-*IK4*-rekin harreman estua izan du. Partekatutako ikerketa hauen ondorioz, ikusmenaren arazo askoren irtenbidea aurkitzea posiblea izan da.

1.3 Ikerketaren oinarriak

Aurrera eraman den ikerketaren lerro nagusiak konputagailu bidezko ikusmenaren zientziaren barruan aurkitzen dira. Zientzia guztiekin gertatzen den bezala, konputagailu bidezko ikusmenaren zientziak ere elkarrekintza handia dauka beste zientzia batzuekin. Kasu honetan, ikasketa automatikoko zientziarekin hain zuzen ere. Ondorioz, ikerketa honen barruan interakzio horren adierazle diren estrategiak erabiliko dira ikusmeneko arazoei aurre egiteko. Irakurlea ikerketaren esparruaren oinarritzko nozioak ezagutu ditzan, jarraian bi zientzia hauen sarrera txiki bat agertzen da. Ikusmenaren erronkak zeintzuk diren aipatu eta horiei erantzuna emateko aurkezten diren ikerketa lerro nagusiak aztertuko dira. Ikerketa lerro horiek izan dira aurrera eraman den ikerketaren muina.

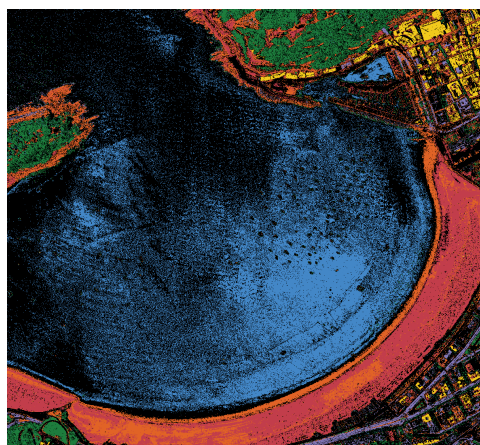
1.3.1 Konputagailu bidezko ikusmena (*Computer Vision*)

Aurretik aipatu den bezala, txosten honetan aurkezten den ikerketaren ekarpenak konputagailu bidezko ikusmenaren zientziaren barruan kokatzen dira. Baina zer da zehazki konputagailu bidezko ikusmena? Ordenagailu bat edo makina bat *ikusteko* gaitasunarekin hornitzea da. *Ikusi* esaten denean ez da soilik aurrean dagoen irudiaren datuen bilketa, baizik eta ikusmenaren prozesu guztia adierazten da. Hau da, irudia jaso, prozesatu eta interpretatzeko behar diren pausu guztiak hartzen dira kontutan.

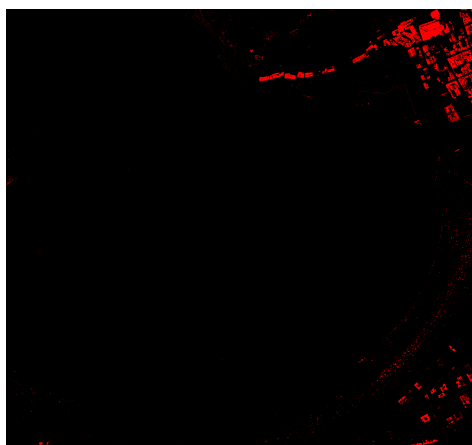
Beraz, ikusmenaren prozesua adimenarekin elkarrekintza handia duen prozesu bat da. Ondorioz, konputagailu bidezko ikusmenaren zientziak beste zientzia askorekin puntu amankomunak ditu: optika, argazkilaritza, matematika, estatistika, neurozientzia, psikologia, etabar. Gainera, alor honetan egiten diren ikerketen helburuei erreparatuz, arazoari irtenbidea bilatzeko bi ikuspuntu nagusi daude: konputagailu bidezko ikusmena ingenieritza bat bezala edo adimen artifizialerako bide bezala ulertzea [Eri11].

Ingeniaritzaren ikuspuntuaren kasuan, helburu nagusia mundu errealean baliagarriak diren aplikazioen garapena da. Beraz, arazo errealei irtenbidea ematea bilatzen

da eta horregatik, garatutako metodoak azkarrak, fidagarriak eta irmoak izan behar dute. Askotan metodo hauek jadanik egonkortuta dauden teknologietan oinarrituko dira. Orokorrean, kasu hauetan garatutako irtenbideak, arazoari doituta daude eta ez dira beste domeinu batean aplikagarriak. Adibidez, teledetekzio irudietan (ikusi 1.2 irudia) ezaugarri konkretuak dituen eskualde bat aurkitzeko erabilitako metodoa ez da baliagarria izango zerura ateratako argazkietan hodeiak identifikatu ahal izateko.



(a) Kontzeptu semantikoen sailkapena



(b) Eraikin kontzeptu semantikoko pixelen aukeraketa

Irudia 1.2.: Teledetekzio irudi baten kontzeptu semantikoen sailkapena

Garatutako metodoen helburua edozein sistema adimen artifizialarekin hornitzea bada berriz, arazo orokorragoei irtenbidea emango dioten metodoak dira beharrezkoak. Honek burmuinaren ulertzeko era aztertzea eskatzen du, eta ondorioz, arazoaren konplexutasuna asko handitzen da. Konplexutasun horren barruan, irudiaren behe-mailako ikasketa, goi-mailako ikasketa eta testuinguruaren detekzioa azpimarratu behar dira. Adibide gisa, hainbeste ikerketa talde egun konpontzen ahalegintzen ari diren argazki batean dauden pertsonen ezagutzaren arazoa jar daiteke.

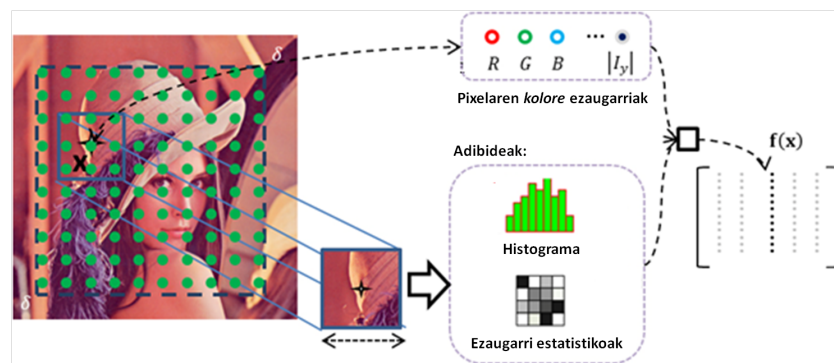
Aurretik esan den bezala, konputagailu bidezko ikusmenaren zientziak, irudietatik edo irudi sekuentzia batetik, bideotik halaber, era automatikoan erabilgarria den informazioa atera, aztertu eta ulertzeko balio dituzten erremintak garatzeko metodoak biltzen ditu. Orokorrean, metodo hauek oso helburu desberdinak izan ditzakete: irudien eskuratzea, irudien aurre-prozesaketa, irudien prozesaketa, irudien ezaugarrien ateratzea, eskualde baten segmentazioa, objektuen segmentazioa, objektu baten errekonozimendua, mugimendu analisia edota irudiaren ulerpene.

Zerrendatutako helburu guzti horiek irudiaren behe-mailako analisitik abiatzen dira. Irudi baten behe-mailako analisia irudiaren berezko informazioa aztertzean datza [dAn+14]. Horretarako, irudiaren oinarritzko informazioa den pixela hartzen da unitate bezala. Pixelen informazioa, honen auzoko pixel multzoen informazioa

eta haien arteko osaketa espaziala erabiltzen dira irudien ezaugarriak ateratzeko. Ezaugarri hauen helburua irudiaren errepresentazio gisa erabiltzen den informazio kopurua murriztea da, horregatik, irudiaren ordezkari zehatzena izatea da komenigarriena.

Orokorrean, bi multzo nagusitan banatzen dira irudiak aztertzeko erabiltzen diren behe-mailako ezaugarriak: ezaugarri lokalak eta ezaugarri globalak. Ezaugarri globalak irudia bere osotasunean deskribatzen dute, lokalek berriz, irudiaren adabaki edo zati bat. Irudiaren adabaki horiek aukeratzeko, irudiaren *keypointak* edo gako puntuak bilatzen dira, hauek irudiaren puntu esanguratsuenak kontsideratzen direlarik. Hots, informazio kantitate handiko puntuak. Ezaugarri globalen artean kolorea, ingerada edo silueta, forma edota testura aztertzen dira. Forma matrizeak, Momentu inbariantek (Hu, Zerinke), Orientatutako Gradienten histograma (HOG) [ASC13], [DT05] eta Co-HOG dira hauen adibide batzuk. Orokorrean, irudien berreskuratzean eta objektuen detekzioan erabiltzen dira batez ere.

Ezaugarri lokalak deskriptore bezala ere ezagutzen dira. Ezaugarri lokalak [TM08] berriz, irudiaren zati baten informazioa konprimitzen dute, zehatzago, pixel baten inguruko testura irudikatzen dute orokorrean (ikus 1.3 irudia). Hauen artean, SIFT [Low99], [HTZ08], SURF, LBP [GZZ10], BRISK, MSER eta FREAK dira aipagarrienak. Hauek, objektuen identifikaziorako erabiltzen dira batik bat. Azpimarratu behar da, planteatutako arazoaren arabera eta eskuragarri dauden datuen arabera erabili beharreko ezaugarrien aukeraketa prozesua oso desberdina izan daitekeela eta askotan aukeraketa hori izan daiteke arazoaren irtenbidea aurkitzeko gakoa.



Irudia 1.3.: Irudien ezaugarri lokalen lorpena

Konputagailu bidezko ikusmenak etorkizun erakargarria du aurretik, batez ere irudi eta bideoen ulermen automatikoaren norabidean. Nahiz eta azken bost hamarkadetan modan egon den zientzia izan, oraindik gaztea kontsideratzen da. Szeliski-k [Sze10] azpimarratzen duen bezala, "ordenagailu batek bi urtetako ume batek irudi batean dauden objektu guztiak aurkitu eta errekonozitzeko duen gaitasun berdina izateko urte asko falta dira oraindik". Gaur egun zientzia honen barruan dagoen erronkarik

nagusiena irudi eta bideoen ulerpena da, adimen artifizialaren bidean lagunduko duena.

Erronka honi aurre egiteko, beharrezkoa da beste azpi-erronka batzuei irtenbidea aurkitzen joatea. Azpi-erronka hauek ikusmenaren barruan aurkitzen diren metodoek betetzen dituzten funtzioen arabera banatzen dira gehienetan. Ikusmenaren funtzioen artean hauek dira aipagarrienak: irudiaren ezaugarrien eskuratzea, irudiaren aurre-prozesaketa eta prozesaketa, irudi segmentazioa, objektu segmentazioa eta identifikazioa, eszenen kategoria banaketa, 3D berreraiketa, eta abar. Funtzio hauetako asko bideoen analisisian ere erabiltzen dira. Bideo analitikaren berezkoak diren funtzioen artean ekintzen sailkapena eta errekonozimendua, pertsona eta objektuen tracking edo jarraipena, mugimendu analisia, eszenen banaketa, eta abar aipatu daitezke. Funtzio guzti hauek irudien eta bideoen ulerpena lortzeko tresnak dira baina gainera beraiek bakarrik konputagailu bidezko ikusmenak konpondu ditzakeen aplikazio zehatzei irtenbidea emateko giltza dira.

Ikerketa lan hau, hurrengo funtzioen analisisian oinarritu da: irudiari dagokionez, irudiaren ezaugarrien eskuratzea, irudiaren aurre-prozesaketa eta prozesaketa, irudi segmentazioa, objektu segmentazioa eta identifikazioan. Bideoen kasuan berriz, ekintzen sailkapena eta errekonozimendua eta pertsona eta objektuen tracking edo jarraipenaren barruan kokatzen dira egindako ekarpen nagusiak.

1.3.2 Ikasketa automatikoa (*Machine Learning*)

Arthur Samuel-ek [Sam59]-n definitzen duen bezala, ikasketa automatikoa explizituki programatu gabe ordenagailuei ikasteko abilezia ematen dien zientzia da. Beranduago Tom Michell-ek [Mit97] beste definizio bat gehitu zuen: datuetan modeloak, patroiak eta erregulartasunak aurkitzeko metodoetan ikertzen duen konputazio zientzia da. Ikasketa automatikoa zientzia numerikoak bere kasa ebatzi ezin dituen problemak ebazteko kapaza da.

Bi multzo handietan banatzen dira ikasketa automatikorako erabiltzen diren teknikak:

- **Gainbegiraturako ikasketa:** sistema trebatzeko erabiltzen diren datuak klasifikatuta daude, hau da, klase etiketa bat egokituta dute.
- **Ez-gainbegiraturako ikasketa (Clustering):** sistema trebatzeko erabiltzen diren datuak ez daude etiketatuta.



Irudia 1.4.: *Imaginet* datu basearen irudien ez-gainbegiratutako sailkapena

Gainbegiratutako ikasketaren helburua normalean egokitutako iragarle bat garatzea da. *Ikasketa* prozesua, arazo konkretu baterako doituta dagoen sistemak erabiliko dituen algoritmo matematiko konplexuen optimizazioan datza. Horregatik, prozesua normalean bi pausutan egiten da: lehenengo sistema trebatu egiten da eta ondoren testatu. Sistema entrenatzeko beharrezkoak dira etiketatuta dauden datuak erabiltzea, hau da, aldez aurretik sailkatuta dauden datuak. Datu hauek sistemaren optimizazio prozesuan erabiliko dira. Sistema optimizatu denean, testerako datuak sartzen dira eta hauen sailkapena egingo da emaitza bezala probabilitate banaketa bat lortuz.

Aurretik etiketatuta ez dauden datuak erabiltzen direnean, ez-gainbegiratutako ikasketa metodoak erabiltzen dira (ikus 1.4 irudia). Metodo hauen helburua datuen arteko harremanak bilatu eta datuen barnean ager daitezkeen patroiak identifikatzea da.

Azkenengo urteetan bi hauen arteko erdibide bat ere agertu da: **Erdi-gainbegiratutako ikasketa**. Normalean, etiketatu gabeko datu multzo handiak dauden esparruetarako aplikazioak garatzeko erabiltzen dira metodo hauek. Adibidez, irudien prozesaketa, irudien berreskuratze eta bioinformatikan. Datu kantitatea oso handia denez, multzo horretako batzuk bakarrik daude etiketatuta.

Konputagailu bidezko ikusmenak erronka handiko arazoak aurkezten dizkio ikasketa automatikoari. Era berean, ikasketa automatikoak konputagailu bidezko ikusmenean erabilgarriak diren algoritmo malgu eta sendoen garapenerako ahalmena dauka. Ikasketa prozesuetan oinarritutako irtenbideek domeinu batetarako garatutako ikusmen sistema bat beste domeinu batetara egokitzeko aukera ematen dute, *ikasiz* lortutako ezagupena berrerabili bait daiteke.

Gaur egun, ikusmenaren barruan dauden ikasketa automatikoko aplikazioak asko dira: segmentazioa eta ezaugarrien ateratzea, formen errepresentazioa, ikasketa arauak, patroien ikasketarako algoritmoak, eta abar luze bat. Laburbilduz, ikasketa automatikoak konputagailu bidezko ikusmenaren zientziari mundu erreala hobeto ulertzeko erremintak eskaintzen dizkio eta hau ezinbestekoa da adimen artifiziala lortu ahal izateko.

1.4 Ikerketa lerro nagusiak

Ikerketa lan honen eboluzioa bide natural bat izan da, irudi prozesamendu soiletik, ikasketa automatikoaren metodoen ezarpena eta azkenik, aurreko bietan lortutako ezagutza erabiliz, bideo analitika. Hasieran, irudi analisirako behe-mailako deskriptoreak soilik erabili dira ikusmenaren arazoei aurre egiteko. Gero, ikerketa jorratu den proiektuen beharrak bultzatuta, ikasketa automatikoko metodoak gehitu dira. Jakina denez, irudiaren behe-mailako analisiarekin bakarrik arazo asko irtenbiderik gabe geratzen dira; orduan, lortutako emaitzak hobetu ahal izateko, beharrezkoak dira sailkatzaileek erabiltzen dituzten ikasketa metodo konplexuak.

Azkenik, munduan egunero sortzen diren bideoen hazkuntzak ikerketa bideoaren azterketarantz bideratzea ekarri du. Honek analisia beste ikuspuntu batetik egitea eskatzen du baina aurreko bi alorretan aplikatutako irudi analisirako eta ikasketa automatikorako ezagutza anitz berrerabil daitezke.

Konputagailu bidezko ikusmenaren erronka handiena irudien eta bideoen ulerpena da eta hau lortzeko egun aplikazio zehatzei irtenbidea emateko erabiltzen diren funtzio askok helburu hori lortzeko erremintak eskaintzen dituzte. Funtzio horien artean, irudi prozesamendua, irudiaren eskuratzea, irudiaren aurre-prozesaketa eta prozesaketa, irudi segmentazioa, objektu segmentazioa eta identifikazioa, 3D berreraiketa, ekintzen sailkapena eta errekonozimendua, pertsona eta objektuen tracking edo jarraipena, mugimendu analisia, eszenen banaketa aurkitzen dira.

Lan honen ikerketa irudien prozesaketa, segmentazio eta sailkapenean eta bideoen ekintzen sailkapenean ardaztu da. Zehatzago, bereziki hiru ikerketa lerro hauetan lortu dira emaitza adierazgarriak: irudien ulerpen semantikoa lortzeko behe-mailako deskriptoreen analisi eta erabilera, irudi analisirako deskriptoreez gain sailkapen metodoen ezarpena eta bideoen analitika.

1.4.1 Irudien ulerpen semantikorako behe-mailako analisia

Aurretik azaldu den bezala, irudien behe-mailako analisia ezinbestekoa da irudiaren semantika atera ahal izateko. Ikerketa azpi-lerro honen muina irudietatik bereizgarriak diren behe-mailako ezaugarriak ateratzeko metodoetan datza. Ezaugarri hauek irudien informazioa era murriztuan adierazteko erabiltzen dira, era honetan informazioaren dimentsionaltasuna murrizten da eta irudi kopuru handiak prozesatzeko kapazitatea handitu egiten da.

Ezaugarri hauen aukeraketa ez da ataza erraza, irudiaren edukiaren arabera hauen egokitasuna alda daiteke. Horregatik, garrantzitsua da arazoa ondo identifikatu eta atributu egokienak aukeratzea. Gehienetan, ezaugarri global eta lokalak bateratzen dira, konbinaketa honek emaitzen zehaztasuna hobetzen duelako. [Wan+11]-n lanean, ezaugarrien konbinaketa erabiltzen da antzeko bi irudi antzemateko. Wu [WW09] eta Eisa-k [Eis14] berriz forman oinarritutako irudien berreskupenerako metodoak proposatzen dituzte. [And+12] ikerketan bi ezaugarri motak erabiltzen dira irudi eta bideoaren berreskurapenerako.

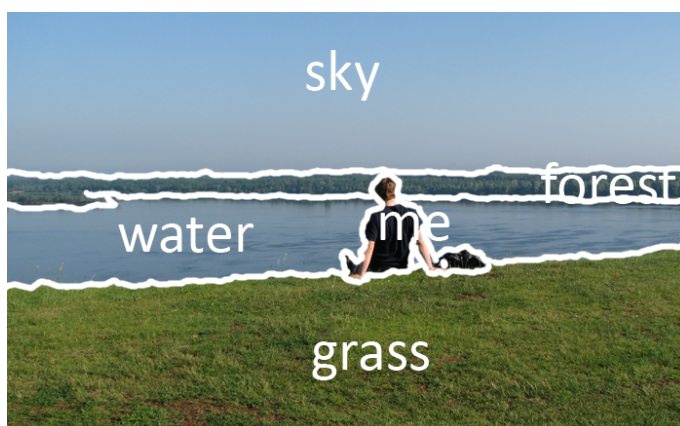
Hala ere, gaur egun, ezaugarri edo deskriptore lokalak indarra hartzen ari dira bereizgarriak baitira, hau da, transformazio geometriko askotara inbariantek dira. Deskriptore lokalak asko erabili dira konputagailu ikusmenaren arloan besteak beste irudien berreskupenerako [SM97], [TV04] [DKN08], [PV13], [YYD15] eta [DKN04]; irudi errekonozimenduan [Low04], [BMP02], [Tao+07] eta [Guo+14]; objektuen sailkapenerako [MLS05] eta [FPZ03] eta bideoen eszenen detekzioan [SSZ06].

Azkenengo hamarkadan, ingelesez *Bag of Features (BoF)* [OD11] bezala ezagutzen den kontzeptu berri bat agertu da eta ordenik gabeko ezaugarri lokalen bildura bat egitean datza. Izena testu berreskurapenean erabiltzen den Bag of Words errepresentaziotik dator. Metodo hauen oinarria hiztegi bisual bat da. Hiztegi hori sortzeko, entrenamendurako erabiltzen diren irudietatik ateratako ezaugarriak multzokatzen dira. Multzokatze hau beharrezkoa da sortutako hiztegiaren tamaina murrizteko, alegia, irudietatik ateratako milioika ezaugarrien laginketa bat egiteko. Irudi berri bat sailkatu nahi denean, ezaugarriak atera, multzo bakoitzera duen distantzia kalkulatu eta gertuko multzoaren barruan sartzen da. Metodo hau aplikatuz, oso emaitza onak lortu dira bai irudi berreskurapenean bai irudi sailkapenean [NJT06], [LSP06b], [ZZZ16], [RV16].

1.4.2 Irudien ulerpenerako ikasketa automatikoko metodoen erabilera

Aurretik aipatu den bezala, irudi prozesamendua soilik ez da nahikoa konputagailu bidezko ikusmenaren inguruan agertzen ari diren arazoei aurre egiteko. Ikasketa automatikoko metodoak gehitzean aurrerapauso handi bat ematen da ikusmen automatiko adimentsuaren norabidean, batez ere moldaketa beharrezkoa den arazoan aurrean. Ikasketa automatikoko metodoek arrazoiketa estatistikoa erabiltzen dute ahalik eta soluzio hobereana aurkitzeko eta analisi estatistiko honek ematen dituen emaitzak erdi-maila edo goi-mailako ulerpen semantikoa lortzeko emaitza adierazgarriagoak eskaintzen ditu. Ikasketa automatikoko algoritmoek bi helburu nagusi dituzte konputagailu bidezko ikusmeneko sistemetan: irudiaren ingurua ulertzeko baliabideak eskaini eta inguru horren barne irudikapenaren eta ezagutzaren arteko zubia eraikitzea.

Irudi edo bideoen ulerpeneren esparruan ikasketa metodoen erabilerearen aplikazioak oso desberdinak dira. Adibidez, irudien segmentazio semantikoa egiteko erabiltzen dira. Askotan, irudian bertan dagoen objektu bat identifikatzea baino askoz interesgarriagoa da irudi guztian dauden kontzeptu semantikoak segmentatzea (ikus 1.5 irudia). Kasu honetan, pixel mailako sailkapena oso erabilgarria da. Gainera kontutan izan behar da, entrenamendurako erabilgarri dauden datuak asko direla pixelak baitaute etiketatuta. Lattner-ek [LMH04] lanean paisai irudien pixel multzoak etiketatzeko sailkatzaile desberdinen konparazioa aurkezten du. Loui-k [LK06] patentean berriz, clustering eta gainbegiratutako metodoak konbinatzen ditu irudi baten eskualdeen sailkapen semantikoa egiteko. Irudien ulerpeneren bidean, [KLH02] lanak naturaren irudien eskualdeak sailkatzeko metodo berri bat proposatzen du irudiaren kontestua ezartzeko aurre-pauso bezala.



Irudia 1.5.: Irudi baten segmentazio semantikoaren erakustea [Sha]

Eskuizkribu-ezagutze automatikoko aplikazioetan ere asko erabiltzen dira sailkapen metodoak. Aplikazio hauetan, erabilitako irudi prozesamendua eta sailkapen metodoak harreman estua izaten dute, hau da, ikasketarako erabiliko diren metodoen arabera hautatzen dira irudietatik aterako diren ezaugarriak; edota ezaugarri onenak zeintzuk diren jakinda sailkapen metodo bat edo beste erabiltzen da. [Kne+98] eta [Oli+95] argitalpenetan, karaktereen segmentazio inplizitu baten ondoren hauen errekonozimendurako, Markov-en Ezkutuko Ereduak erabiltzen dira. [SDN16], [Zam+14] eta [Pha+14] lanetan berriz, neurona sareetan oinarritutako sailkatzailak erabiltzen dira helburu berdinarekin.

Azkenengo 50 urteetan sakonki aztertu den alorra, irudietan agertzen diren aurpegiaren errekonozimendua da. Identifikazio prozesu honek lau pausu nagusi ditu: aurpegiaren kokapena aurkitu, aurpegia aurkitzen den irudi adabakiaren normalizazioa (irudien konparaketa egin ahal izateko), ezaugarrien ateratzea eta berezko errekonozimendua. Hiru lehenengo pausotan irudi prozesaketako metodoak erabiltzen dira, azkenengoan berriz sailkapen metodoak. Orokorrean, sailkapen metodo hauen artean, bektore-euskarridun makina [Jon+02], [Phi+98], [Wan+16] eta neurona sareak [Hu+15], [RBK98], [Law+97] irudien ezaugarri globalak ateratzen direnean erabiltzen dira eta Markov-en Ezkutuko Ereduak [SH94], [NH98], [LC03], [BCM03] berriz, ezaugarri lokalekin lan egiten denean.

Ikasketa automatikoko metodoak ere bideoen ulerpenerako sistemetan aplikatzen dira. Bideoen segmentazioa egiteko, hau da, bideoa esanahi semantikoa duten unitatetetan banatzeko, [ZL02] eta [LMP04] lanetan Markov-en Ezkutuko Ereduak erabiltzen dira adibidez. Clustering metodoak ere askotan erabiltzen dira antzeko ezaugarriak dituzten irudi edo frameak multzokatzea baita helburua [YYL98], [GFT98]. [Yin+07], [TZ04] eta horretan dira iaioak clustering metodoak hain zuzen ere.

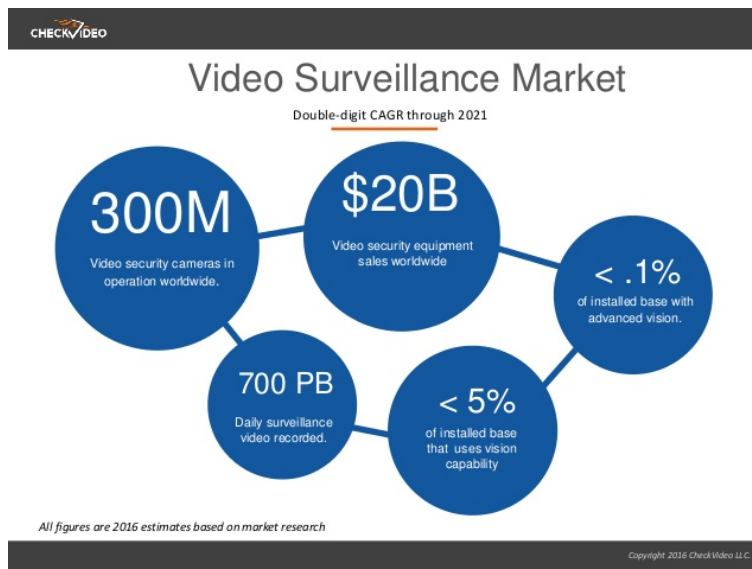
Hemen azaldutakoa, ikasketa metodoek ikusmenaren barruan duten aplikazioen lagin txiki bat da soilik. Irudi eta bideoen ulerpenaren bidean argi dago ezinbestekoak direla metodo hauek eta ondorioz, beharrezkoa da hauen erabilpena aztertzea estrategiarik onena aukeratu ahal izateko.

1.4.3 Bideo analitika

Bideo analitika konputagailu bidezko ikusmenaren zientziaren barnean kokatzen da. Bideo bat bata bestearen jarraian dauden irudiak izanik, naturala da irudien azterketatik bideoen azterketara salto egin izana komunitate zientifikoak. Gainera, gaur egun munduan sortzen den eduki gehiena bideoak direla jakinik, beharrezkoa da irudien inguruan egindako hausnarketa bideoetara luzatzea. Irudiaren ulerpenaren

barruan bezala helburu edota aplikazio aunitz biltzen ditu: objektuen ezagutzea, eszenen sailkapena, eszenen semantikaren inferentzia, mugimendu detekzioa, objektuen jarraipena, ekintzen sailkapena, etabar luze bat.

Txosten honen hasieran aipatu den legez, 2019rako sarean dagoen edukiaren %80a bideoak izango direla aurreikusten da. Horregatik azken urteetan bideo analitikaren inguruan egiten den ikerketa zientifikoa pila hazi da. 2016ko Global Video Analytics Market Report delakoan Bideo Analitikaren merkatua 1,69 mila milloi dolarretatik 2016an, 4,23 mila milloi dolarretara haziko dela estimatzen da [RM16]. Baina badago beste arrazoi nagusi bat hazkunde honetarako: segurtasun aldetik agertzen diren mehatxuaren aitzakiaz zaintza adimentsua bezala ezagutzen den kontzeptua. Gero eta kamara gehiago kokatzen dira segurtasuna bermatzeko eta hauek egunero milaka terabyte grabatzen dituzte (ikus 1.6 irudia. Informazio guzti hori gizakiok aztertzea ez da bideragarria eta ondorioz automatizatutako irtenbideak beharrezkoak dira. Hala ere, gaur egun, oraindik soluzio erreal batetik urrun gaude.



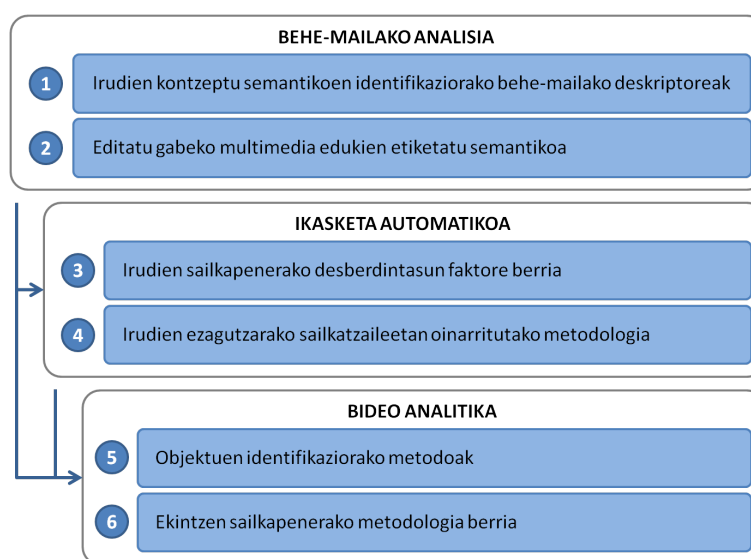
Irudia 1.6.: Bideo zaintza merkaturen aurreikuspenak 2021. urterako [RM16]

Bideoen ulerpenaren bidean, gizakien ekintzen sailkapen eta errekonozimendua, atazarik garrantzitsuenetakoa bilakatu da azkenengo hamarkadetan. Hau da sistema adimentsuei eskatzen zaien lehenengo funtzionalitatea. Gizakion ekintzen errekonozimenduak aplikazio asko ditu: adibidez, edukian oinarritutako bideo analitika [GN08], [Pet00]; pertsona-ordenagailu interakzioa [JS07], [RA15], [Cow+01]; zaintza bisuala [HVL08], [Hu+04], [Kim+10], [PW12]; bideoen indexazioa [SZ94]; eta abar. Aplikazio guzti hauen helburu nagusia bideoan gertatzen dena ulertzea da.

Lan guzti hauek lortutako emaitzak etorkizun handikoak dira, baina irudien kasuan oraindik gizakiek eman dezaketan erantzunetik urrun gaudela esaten bada, bideoen kasuan distantzia hori oraindik nabariagoa da arazoaren konplexutasun mailagatik.

1.4.4 Ikerketaren ekarpen nagusiak

Aurretik deskribatu den bezala ikerketa honen ekarpen nagusiak hiru lerrotan banatzen dira (ikus 1.7 irudia): irudien ulerpen semantikorako behe-mailako deskriptoreen analisi eta erabilera, helburu berdinarekin ikasketa automatikoen metodoen erabilera eta azkenengoa, bideo analitika.



Irudia 1.7.: Aurkeztutako ikerketa lanaren ekarpen nagusiak

Irudien ulerpen semantikoa lortzeko behe-mailako analisiari dagokionez, ikerketa lan honen ekarpen nagusiak hurrengoak dira:

- 1 Irudi barruan dauden kontzeptu semantikoen identifikaziorako behe-mailako deskriptoreak:** irudi baten barruan dauden pixelen sailkapena egiteko irudien behe-mailako deskriptoreetan oinarritutako metodo berriak definitu dira. Algoritmo hauek irudi bakoitzaren barruan agertzen diren kontzeptu semantiko edo klaseak identifikatzea ahalbidetzen dute.
- 2 Editatu gabeko multimedia edukien etiketatu semantikoak:** irudien behe-mailako deskriptoreak erabiliz, irudi bakoitzean dauden objektuen identifikaziorako metodoak. Metodo hauek oinarri bezala bideoetako irudiak hartzen dituzte baina ez dute bideo baten tenporaltasuna kontutan hartzen analisirako.

Lortutako objektuak kontutan hartuta eta arau sinpleak erabiliz, goi-mailako etiketak jartzen zaizkie *shot* edo hartze bezala ezagutzen diren bideo zatiei.

Ikasketa automatikoko metodoen ezarpenari dagokionez, ikerketa lan honen ekarpen nagusiak hurrengoak dira:

- 3 **Irudien sailkapenerako desberdintasun faktore berri baten definizioa:** bi irudien arteko desberdintasuna neurtu ahal izateko, sailkatzaileen emaitza bezala lortzen den probabilitate banaketan distantzian oinarritutako neurri berri bat definitu da.
- 4 **Irudien ezagutzarako sailkatzaileetan oinarritutako metodologia:** irudien sailkapena egiteko hauen behe-mailako deskriptoreak erabiliz, sailkatzaile ezberdinen egokitasuna aztertu da datu base ezberdinetan. Horretaz gain, sailkatzaileen emaitzen fusiorako metodologia berri bat aurkeztu da ere.

Bideo analitikari dagokionez, ikerketa lan honen ekarpen nagusiak honako hauek dira:

- 5 **Objektuen identifikaziorako metodoak:** bideo baten barruan dauden objektu desberdinen segmentazio, tracking eta identifikazioa egiteko algoritmo berrien garapena. Metodo hauek aplikazio zehatzetarako doitu daude eta ondorioz, arazo konkretu bati aurre egiteko diseinatuta. Hala ere, diseinu honek algoritmoak antzeko arazoetara erraz moldatzea ahalbidetzen du.
- 6 **Ekintzen sailkapenerako metodologia:** bideo batetan agertzen diren ekintzen sailkapena egiteko metodologia berri bat garatu da. Hau, irudietatik hartutako behe-mailako deskriptore globaletan, bideoaren konpresioa irudi batean, sailkatzaileetan eta atributu egokien aukeraketan oinarritzen da.

Jarraian ikerketa lan honen ekarpen bakoitzaren emaitza bezala lortutako argitalpen nagusiak aipatzen dira:

1. Irudi barruan dauden kontzeptu semantikoen identifikaziorako behe-mailako deskriptoreak

- [Loz+15a] J. Lozano, N. Aginako, M. Quartulli, I. G. Olaizola, E. Zulueta. *Web-Based Supervised Thematic Mapping*. In: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8.5 (May 2015), pp. 2165–2176. Inpaktu-faktorea: 2,145 (2015). Q1 kuartila. (ikusi 2.1.1)

- [Seg+09] A. Segura, A. Moreno, I. G. Olaizola, N. Aginako, M. Labayen, J. Posada, J.A. Aranda, R. Garcia de Andoin. *Visual Processing of Geographic and Environmental Information in the Basque Country: Two Basque Case Studies*. In: *GeoSpatial Visual Analytics: Geographical Information Processing and Visual Analytics for Environmental Security*. Ed. by Raffaele De Amicis, Radovan Stojanovic, and Giuseppe Conti. Dordrecht: Springer Netherlands, 2009, pp. 199–207. (ikusi 2.1.2)
- [LAO08] M. Labayen, N. Aginako, I. G. Olaizola. *Weather analysis system based on sky images taken from the Earth*. In: *IET Conference Proceedings*. Institution of Engineering and Technology, Jan. 2008, 146–151(5). (ikusi 2.1.3)

2. Editatu gabeko multimedia edukien etiketatu semantikoa

- [Nac+08] S. U. Naci, U. Damjanovic, B. Mansencal, et al. *The COST292 Experimental Framework for Rushes Summarization Task in TRECVID 2008*. In: *Proceedings of the 2Nd ACM TRECVID Video Summarization Workshop*. TVS '08. Vancouver, British Columbia, Canada: ACM, 2008, pp. 40–44. (ikusi 2.1.4)

3. Irudien sailkapenerako disimilitate faktore berri baten definizioa

- [Agi+17b] N. Aginako, G. Echegaray, J. Martínez-Otzerta, I. Rodriguez, E. Lazkano, B. Sierra. *Iris matching by means of machine learning paradigms: a new approach to dissimilarity computation*. In: *Pattern Recognition Letters (2017)*. Inpaktu-faktorea: 1,586 (2015). Q2 kuartila. (ikusi 2.2.1)

4. Sailkatzaileetan oinarritutako irudien ezagutzarako metodologiak

- [Agi+17c] N. Aginako, J. Martínez-Otzerta, B. Sierra, M. Castrillón-Santana, J. Lorenzo-Navarro. *Periocular and iris local descriptors for identity verification in mobile applications*. In: *Pattern Recognition Letters (2017)*. Inpaktu-faktorea: 1,586 (2015). Q2 kuartila. (ikusi 2.2.2)
- [Agi+14] N. Aginako, J. Lozano, M. Quartulli, B. Sierra, I. G. Olaizola. *Identification of plant species on large botanical image datasets*. In: *1st International Workshop on Environmental Multimedia Retrieval co-located with ACM International Conference on Multimedia Retrieval (ICMR 2014)*. Vol. 1222. 2014, pp. 38–44. (ikusi 2.2.3)

5. Objektuen segmentazio, tracking eta identifikazioa

- [OAL12] I. G. Olaizola, N. Aginako, and M. Labayen. *Method of detection and recognition of logos in a video stream*. European Patent (PCT). ES2395448 (T3). 2013-02-12. (ikusi 2.3.1)
- [Ola+08] I. G. Olaizola, J. Flórez, J.C. San Román, N. Aginako, M. Labayen. *Method for detecting the point of impact of a ball in sports events*. European Patent (PCT). ES2402728 (T3). 2013-05-08. (ikusi 2.3.2)
- [Lab+14] M. Labayen, I. G. Olaizola, N. Aginako, J. Flórez. *Accurate ball trajectory tracking and 3D visualization for computer-assisted sports broadcast*. In: *Multimedia tools and applications* 73.3 (2014), pp. 1819–1842. Inpaktu-faktorea: 1,331 (2015). Q2 kuartila. (ikusi 2.3.3)

6. Ekintzen sailkapena

- [Agi+17a] N. Aginako, G. Echegaray, I. G. Olaizola, J. Flórez, B.Sierra. *Machine Learning for Video Action Recognition: a Computer Vision Approach*. In: *Machine Vision and Applications (AURKEZTUTA)*. Inpaktu-faktorea: 1,272 (2015). Q2 kuartila. (ikusi 2.3.4)

1.5 Argitalpenak

Jarraian, dokumentu honetan deskribatzen den ikerketaren ekarpen nagusiak argitaratutako bidearen arabera banatuta azaltzen dira. Eranskineko **Beste argitalpen batzuk** atalean ikus daitezke ikerketa honen emaitza bezala argitaratutako beste lan batzuk.

ALDIZKARIAK/LIBURUAK:

1. [Loz+15a] J. Lozano, N. Aginako, M. Quartulli, I. G. Olaizola, E. Zulueta. *Web-Based Supervised Thematic Mapping*. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8.5 (May 2015), pp. 2165–2176. Inpaktu-faktorea: 2,145 (2015). Q1 kuartila.
2. [Agi+17b] N. Aginako, G. Echegaray, J. Martínez-Otzerta, I. Rodríguez, E. Lazkano, B. Sierra. *Iris matching by means of machine learning paradigms: a new approach to dissimilarity computation*. In: *Pattern Recognition Letters* (2017). Inpaktu-faktorea: 1,586 (2015). Q2 kuartila.

3. [Agi+17c] N. Aginako, J. Martínez-Otzerta, B. Sierra, M. Castrillón-Santana, J. Lorenzo-Navarro. *Periocular and iris local descriptors for identity verification in mobile applications*. In: Pattern Recognition Letters (2017). Inpaktu-faktorea: 1,586 (2015). Q2 kuartila.
4. [Lab+14] M. Labayen, I. G Olaizola, N. Aginako, Julián Flórez. *Accurate ball trajectory tracking and 3D visualization for computer-assisted sports broadcast*. In: Multimedia Tools and Applications 73.3 (2014), pp. 1819–1842. Inpaktu-faktorea: 1,346 (2014). Q2 kuartila.
5. [Seg+09] A. Segura, A. Moreno, I. G. Olaizola, N. Aginako, M. Labayen, J. Posada, J. A. Aranda, R. García De Andoin. *Visual Processing of Geographic and Environmental Information in the Basque Country: Two Basque Case Studies*. In: GeoSpatial Visual Analytics: Geographical Information Processing and Visual Analytics for Environmental Security. Ed. by Raffaele De Amicis, Radovan Stojanovic, and Giuseppe Conti. Dordrecht: Springer Netherlands (2009), pp. 199–207.
6. [Agi+17a] N. Aginako, G. Echegaray, I. G. Olaizola, J. Flórez, B.Sierra. *Machine Learning for Video Action Recognition: a Computer Vision Approach*. In: Machine Vision and Applications (*AURKEZTUTA*). Inpaktu-faktorea: 1,272 (2015). Q2 kuartila.

PATENTEAK:

1. [OAL12] I. G. Olaizola, N. Aginako, M. Labayen. *Method of detection and recognition of logos in a video stream*. European Patent(PCT). ES2395448 (T3). 2013-02-12.
2. [Ola+08] I. G. Olaizola, J. Flórez, J.C. San Román, N. Aginako, M. Labayen. *Method for detecting the point of impact of a ball in sports events*. European Patent(PCT). ES2402728 (T3). 2013-05-08.

KONFERENTZIAK:

1. [Nac+08] S. U. Naci, U. Damnjanovic, B. Mansencal, J. Benois-Pineau, C. Kaes, M. Corvaglia, E. Rossi, N. Aginako. *The COST292 Experimental Framework for Rushes Summarization Task in TRECVID 2008*. In: Proceedings of the 2nd ACM TRECVID Video Summarization Workshop. TVS '08. Vancouver, British Columbia, Canada: ACM, 2008, pp. 40–44.

2. [LAO08] M. Labayen, N. Aginako, I. G. Olaizola. *Weather analysis system based on sky images taken from the Earth*. In: IET Conference Proceedings. Institution of Engineering and Technology, Jan. 2008, 146–151(5).
3. [Agi+14] N. Aginako, J. Lozano, M. Quartulli, B. Sierra, I. G. Olaizola. *Identification of plant species on large botanical image datasets*. In: 1st International Workshop on Environmental Multimedia Retrieval co-located with ACM International Conference on Multimedia Retrieval (ICMR 2014). Vol. 1222, pp. 38–44.

1.6 Txostenaren antolaketa

Txosten honek aurrera eraman den ikerketa lanaren testuingurua azaldu, lortutako emaitzak laburbildu eta argitaratutako lan nagusiak biltzea du helburu. Hiru atal nagusietan banatuta dago:

1. atala **SARRERA** da eta esku artean duzun kapitulu honek osatzen du. Lehenik, ikerketa lan honen motibazioa azaltzen da, lan hau aurrera eramatera zerk bultzatu zuen eta zeintzuk ziren, eta diren, komunitate zientifikoan multimedia edukia ulerpenaren aurrean dauden beharrak eta erronkak. Ondoren, lan honen testuingurua laburbiltzen da, ikerketaren nondik norakoak ulertzeko balioko duena. Jarraian ikerketaren oinarri izan diren alor orokorren aurkezpen txiki bat egiten da eta ondoren ikerketa lanean jorratu diren lerro garrantzitsuenen azalpen sakonago bat egiten da. Honekin batera, ikerketaren ekarpen nagusiak azpimarratzen dira, lerro bakoitzaren barruan dauden argitalpen nagusiak aipatuz. Bukatzeko, ikerketa honen emaitza izan diren argitalpen garrantzitsuenak era laburrean aurkezten dira.

IKERKETAREN EMAITZAK txosten honetako 2. atala da. Bertan, ikerketaren lerro bakoitzean egindako ekarpenen eta lortutako emaitzen deskribapen sakonago bat egin ondoren, lerro bakoitzari dagozkion argitalpenak agertzen dira. Argitalpen hauek argitaratu ziren formatu berean agertzen dira eta ondorioz, ingelesez. Atal hau aurrera eraman den ikerketaren emaitzen adierazgarri nagusia da.

3. atalean, **ONDORIOAK ETA ETORKIZUNERAKO LANA** ikerketaren ondorio nagusiak biltzen dira. Honekin batera, ikerketaren inguruan zabalik geratu diren eta ikerketaren hurrengo pausu bezala proposatzen diren etorkizunerako hainbat bide aipatzen dira.

Txosten honen azkeneko atala, **Eranskinak** da. Atal honen barruan bi azpiatal aurkitzen dira. A. atala **I+G proiektuak** da eta bertan deskribatzen dira ikerketa lan

honen oinarri izan diren I+G proiektuak. Atal honetan, proiektuen oinarrizko datuak agertzeaz gain, proiektu bakoitzeko deskribapena eta helburu nagusiak laburbiltzen dira. Horretaz gain, proiektu bakoitzean ikerketari lotuta dauden lorpenak zehazten dira.

Eta azkenik, B. atalean, **Beste argitalpen batzuk** atalean, ikerketarekin hain lotura zuzena ez duten argitalpenak biltzen dira. Nahiz eta ikerketaren ardatz ez izan, saihetseko ikerketa hauetan egindako lanak, ikerketan aurkeztutako arazoaren ikuspegi sakonago bat izatera lagundu du, betiere, prozesu honen aberastasuna bultzatuz.

IKERKETAREN EMAITZAK

Atal honetan ikerketa lerro bakoitzeko egin diren ekarpen nagusiak deskribatzen dira. Dokumentu honen sarreran aipatu den bezala, gehienetan aurrera eraman diren proiektuek mugatu dituzte ikerketa honen norabidea eta lortutako emaitzak. Horregatik, ikerketa lerro bakoitzean zer proiektu garatu diren aditzera ematen da.

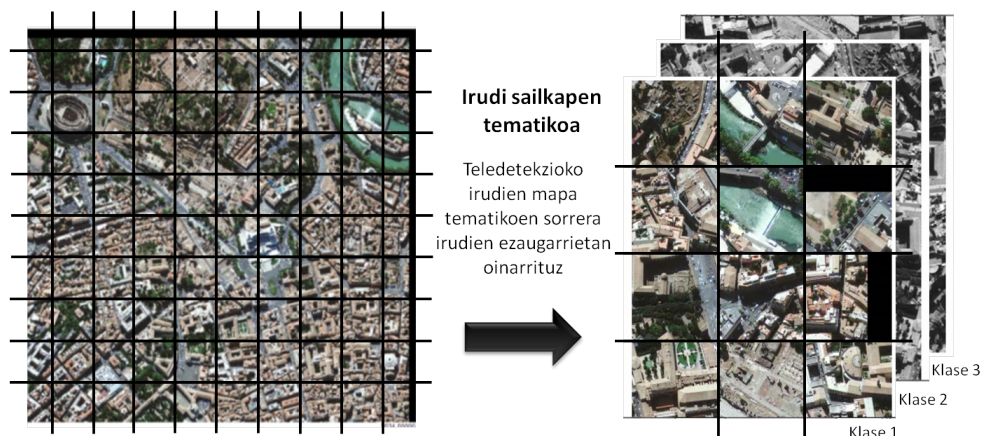
2.1 Irudien ulerpen semantikorako behe-mailako analisia

Irudien ulerpenerako bidean beharrezkoak dira irudien barne edukia aztertzekeo tresnak. Askotan, barne eduki honen azterketa nahikoa da erdi edo goi-mailako informazio bezala ulertu daitekeen ezagutza lortzeko. Alor honen barruan egindako ikerketaren ekarpenak bi multzotan banatzen dira:

1. **Irudian aurkitzen diren kontzeptu semantikoaren identifikaziorako behe-mailako deskriptoreen erabilerari** dagokionez, irudien ezaugarri globaletan oinarrituz, irudien pixel guztiak klase desberdinetan sailkatzeko metodoak garatu dira. Pixelen sailkapen honek, irudiaren barruan dauden kontzeptu semantikoak finkatzea ahalbidetzen du eta ondorioz, irudia bera sailkatzeko informazioa lortzen da (ikusi 2.1.1, 2.1.2 eta 2.1.3 argitalpenak).
2. **Editatu gabeko multimedia edukien etiketatu semantikoaren** kasuan berriaz, irudiaren ezaugarri lokalak eta globalak erabiltzen dira irudiaren etiketatu semantikoaren egiteko. Kasu honetan, bi ezaugarri moten konbinaketa erabiltzeak etiketatu horren aberastasun semantikoaren ekarri du (ikusi 2.1.4 argitalpena).

Irudian aurkitzen diren kontzeptu semantikoaren identifikaziorako behe-mailako deskriptoreak, teledetekzio eta meteorologian baliagarriak diren aplikazioak garatzeko erabili dira. [Loz+15b] **Web-Based Supervised Thematic Mapping** argitalpenean teledetekzio irudien mapa tematikoak sortzeko metodologia erdi-automatiko bat aurkezten da. Pixelen behe-mailako ezaugarrietan oinarrituta, irudi bakoitzaren barruan dauden kontzeptu semantiko desberdinak (eraikina, ura, gunee berdea)

sailkatzen dira eta honek era berean, irudien sailkapen tematikoa egiteko aukera ematen du (ikusi 2.1 irudia). Zehatzago, irudietako ezaugarri globalak erabiliz, kasu honetan, kolore eta testurako ezaugarriak, pixel bakoitzari klase bat egokitzen zaio. Sailkapen honek, batetik, klase zehatz bati dagozkion pixelak bakarrik aukeratzea ahalbidetzen du eta bestetik, irudia barnean dituen klaseak kontutan izanda irudiak etiketaketa errazten du. Horrela, irudi hauen bilaketa eta berreskurapena berauek egokituta dauzkaten kontzeptuak erabiliz egin daiteke.



Irudia 2.1.: Teledetekzio irudien sailkapen tematikoa

Meteorologiari dagokionez, [Seg+09] **Visual Processing of Geographic and Environmental Information in the Basque Country: Two Basque Case Studies** eta [LAO08] **Weather analysis system based on sky images taken from the Earth** lanetan aurkeztutako metodoak, zeruko irudietatik abiatuta, laino-estaldura faktorea automatikoki kalkulatu du. Horretarako, irudiaren behe-mailako ezaugarri globalak erabiliz, irudiaren pixelak lau klasetan banatzen dira: eguzkia, zerua, lurra (zerua ez dena) eta lainoak. Zeru eta laino klaseen arteko pixel kopuruaren ehunekoa kontutan izanda, laino-estalduraren balioa kalkulatu da. Horretaz gain, artikulu honetan norabide zehatz batean behe-laino dagoen jakiteko algoritmoa aurkeztu da. Horretarako, zeruertza detektatu da irudiaren analisisa eginez.

Editatu gabeko multimedia edukien etiketatu semantikoari dagokionez, irudien behe-mailako analisisik irudien etiketatu semantikoa egiteko metodoa aurkeztu da. Ez da aurreko lanean bezala irudiaren barruan dauden kontzeptu semantikoetan oinarritzen baizik eta ezaugarriak atera eta hauen azterketatik zuzenean ondorioztatzen da erdi-mailako ezagutza. Kasu honetan, bideoak aztertu dira baina hauen denbora ezaugarria kontutan hartu ez denez, irudi analisiaren barruan kokatzen den aplikaziotzat hartu da.

[Nac+08] **The COST292 Experimental Framework for Rushes Summarization Task in TRECVID 2008** argitalpenean TRECVID 2008 [NIS] lehiaketako bideoen laburpen automatikoa egiteko metodo bat deskribatu da. Horretarako, bideoa

eszenetan banatzeko algoritmoa garatu da eta eszenei garrantzia maila bat esleitzeko bideoen behe-mailako ezaugarrien egokitzapena aztertu da; hau da, eszena horien etiketatu semantiko bat egiten da. Horrela, desiragarriak ez diren irudiak ezeztatzen dira (adibidez barra diagrama irudiak, irudi zuriak, eta abar) eta aldaketa gehien dituzten eszenak laburpenari atxikitzeke aukeratzen dira.

Lortutako emaitzek adierazten duten bezala, irudi prozesamendu soilak, domeinu itxi eta konkretu baten barruan aurkitzen diren arazoei aurre egiteko tresnak eskaintzen ditu. Aplikazio askotan, nahikoak dira irudietatik ateratako behe-mailako ezaugarriak eta ezarritako arau simple batzuk irudiaren kontzeptu semantikoak ondorioztatu ahal izateko.

Irudien behe-mailako analisiaren inguruan egindako ikerketa Vicomtech-IK4-en garatu diren I+G proiektuen hauen barruan egin da (eranskinean proiektu hauen deskribapen sakonago bat eta bakoitzean lortutako emaitzak adierazten dira):

- RUSHES- Retrieval of multimedia semantic units for enhanced reusability (Proiektu Europarra) (A.1)
- SIAM- Diseño y Desarrollo de un Sistema de Análisis Multimedia de Contenido Audiovisual en Plataformas Web Colaborativas (Proiektu Autonomikoa) (A.2)
- GRAFEMA- Multimedia edukien kudeaketarako sistema (Proiektu Autonomikoa) (A.3)
- SKEYE- Sistema de análisis meteorológico basado en imágenes del cielo tomadas desde tierra (Proiektu Autonomikoa) (A.4)

2.1.1 Web-Based Supervised Thematic Mapping

- **Izenburua:** Web-Based Supervised Thematic Mapping
- **Egileak:** Javier Lozano, Naiara Aginako, Marco Quartulli, Igor G. Olaizola, Ekaitz Zulueta
- **Aldizkaria:** IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing
- **Argitaletxea:** IEEE
- **Zenbakia (Orrialdeak):** Vol.: 8, Issue: 5 (2165-2176)
- **Inpaktu-faktorea (urtea):** 2,145 (2015)
- **Kuartila:** Q1
- **Urtea:** 2015
- **DOI:** <https://doi.org/10.1109/JSTARS.2015.2438034>

Web-Based Supervised Thematic Mapping

Javier Lozano Silva, *Student Member, IEEE*, Naiara Aginako Bengoa,
Marco Quartulli, *Senior Member, IEEE*, Igor G. Olaizola, *Senior Member, IEEE*, and Ekaitz Zulueta

Abstract—We introduce a methodology for semiautomatic thematic map generation from remotely sensed Earth Observation raster image data based on user-selected examples. The methodology is based on a probabilistic k-nearest neighbor supervised classification algorithm. Efficient operation is attained by exploiting data structures for high-dimensional indexing. The methodology is integrated in a Web-mapping server that is coupled to an HTML supervision interface that supports interactive navigation as well as model training and tuning. Quantitative classification quality and performance measurements are extracted for real optical data with 0.25 m resolution on a highly diverse training area.

Index Terms—Remote sensing, thematic mapping, Web-based mapping systems.

I. INTRODUCTION

EARTH Observation (EO) data mining systems are the subject of significant research and development efforts [1]. Petabyte-scale raster data archive volumes are growing at rates of about 10 GB per day and about 95% of their contents have never been accessed by a human observer [2]. Metadata search needs to be complemented by efficient content-based mining tools, for instance, to provide large-scale thematic mapping capabilities and similarity search based on user examples. This implies the development of new strategies and algorithms that are able to characterize and search required detailed objects/concepts. A basis for such tools is represented by semantic labeling algorithms that build upon supervised classification machine learning methods.

Such systems imply a potential expansion of the practice EO data analysis and exploitation from remote-sensing scientists and technology practitioners to application domain experts in multiple sectors. Applications can be served in environmental resource management, agronomy, ecology, risk management, and transport. The resulting applicative transition represents an evolution from Web-based cartography for casual users to thematic cover map generation based on the needs specified interactively by experts of different domains. This entails a basic set of drivers for a system implementing this concept: ease of use, scalability, and effectiveness. These three drivers can be addressed by creating high-quality supervised classification systems integrating them in scalable Web-server architectures

Manuscript received September 23, 2014; revised April 09, 2015; accepted May 11, 2015. Date of publication June 16, 2015; date of current version July 20, 2015.

J. Lozano, N. Aginako, M. Quartulli, and I. G. Olaizola are with Digital Television and Multimedia Services, Vicomtech, Donostia 20009, Spain (e-mail: jlozano@vicomtech.org).

E. Zulueta is with the Department of System Engineering and Automation, EHU/UPV, VitoriaGasteiz 01006, Spain.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/JSTARS.2015.2438034

made accessible by simple-to-use interfaces that combine a geospatial navigation and a supervision component.

For what pertains to the actual core classification algorithm, approaches to the analysis of images acquired by remote-sensing systems include object-based, pixel-based, and hybrid methods. Classical solutions for pixel-based approaches include clustering and classification. Schröder *et al.* [3] use the input provided by the user in a Bayesian learning framework for supervised classification of a currently open image and also for finding most relevant images (images with large extents of the trained class) across an archive. Costa *et al.* [4] present a supervised per-pixel classification followed by a postclassification processing with image segmentation and semantic map generalization. The results show that segmentation of high spatial resolution images and semantic map generalization can be used in an operational context to automatically produce land-cover maps. A hybrid example can be found in [5], with the use of images from QuickBird over Arizona to compare object-based and pixel-based approaches. Their study demonstrates that the object-based classifier improves significantly over classical per-pixel results. An example of combination of techniques can be found in the work developed by Maulik and Sarkar [6]. The authors propose a scalable parallel clustering algorithm of multispectral remote-sensing imagery using point symmetry-based distance. They use a K-d tree-based nearest neighbor search algorithm to compute this distance.

Visual interaction applied to remote-sensing technologies is another topic that is becoming of great interest for the scientific community. Ho *et al.* [7] introduce a framework and class library to shorten the time and effort needed to develop Web applications for visual geospatial analytics and provides a collection of geographical and information visualization representations. Keel [8] introduces an environment for the support of remote-collaborative sense-making activities. The system has computational agents that infer relationships among information items by the analysis of their spatial and temporal organization. Within this kind of work, user interaction, user friendly environments and user interface design are issues that acquire great relevance.

The present contribution has two main aspects: 1) an algorithmic one and 2) a system development one. On the algorithmic side, we extend a probabilistic k-nearest neighbor classification developed in the database domain to the generation of multi-class thematic layers from user supervision. To the best of our knowledge, this is the first time that this algorithm is considered and evaluated for large-scale thematic map generation from submetric resolution remote-sensing imagery. Furthermore, we address the issue of performance in the case of large multi-dimensional datasets by introducing efficient data structures

based on K-d trees. On the side of system implementation, we develop a preoperational prototype that includes a Web-based user interface and both a Web-mapping and a supervised classification server. A user interface (UI) allows domain experts to navigate large volumes of geospatial data and provide training to the machine learning components on the server. The machine learning server computes thematic layers by supervised classification based on the efficient data structures that it maintains.

This paper is organized as follows. Section II introduces theoretical concepts used in the development of the presented prototype. Section III describes system architecture, including a description of the workflow process. Section IV presents the created evaluation framework. In Section V, measured performance values are presented and analyzed. Finally, conclusion is presented.

II. METHODOLOGICAL APPROACH

Unlike [9], which considers unsupervised classification approaches, we focus on including the user in the training process. An interactive learning scheme allows a supervisor to define examples by interacting with a Web-based geovisualization interface. Interaction events directly influence a probabilistic model of the thematic class of interest that is built on top of an indexing structure.

This supervised thematic mapping involves uncertainties in the form of noise in the data and of uncertainties in the training provided by the users. The principled management of these uncertainties requires probabilistic classification algorithms, while operational efficiency requires that such algorithms are implemented on top of efficient data structures for large N-dimensional datasets. This section presents the probabilistic k-nearest neighbor algorithm selected for the thematic mapping service, and describes its efficient implementation in terms of K-d trees.

A. Supervised Probabilistic Classification Algorithm

The specific version of the classification algorithm that we employ is closely related to the probability threshold k-nearest neighbor (T-k-PNN) in [10], which is designed to return a most probable set S from D for a given data point o_i such that

$$S|S \subseteq D \wedge |S| = k \text{ and } p(S) \geq T, \text{ where } T \in [0, 1].$$

In this, the qualification probability $p(S)$ of a k -subset S is computed as

$$p(S) = \sum_{o_i \in S} \int_0^{+\infty} d_i(r) \prod_{o_j \in S - \{o_i\}} D_j(r) \prod_{o_h \in D - S} (1 - D_h(r)) dr \quad (1)$$

where the distance pdf of uncertain training pixel o_i is denoted by $d_i(r)$ while its cumulative density function (cdf) is denoted by $D_i(r)$, $r \in \mathfrak{R}$ being a value taken on by the absolute distance $r_i = |o_i - q|$ to the query point q , and where pdfs are estimated

TABLE I
SYMBOLS FOR (1) DESCRIBING THE PROBABILISTIC k -NEAREST NEIGHBOR SUPERVISED CLASSIFICATION ALGORITHM

Symbol	Meaning
S	Point class—image region class
D	Uncertain database—image regions to classify
k	Number of points
$p(S)$	Quantification probability of S
T	Probability threshold
o_i	Uncertain object of $D(i = 1, \dots, D)$ —training image region
q	Query point—image region with unknown class
r_i	$ o_i - q $ Distance from current region to training
$d_i(r)$	PDF of r_i (distance Probability Density Function)
$D_i(r)$	CDF of r_i (distance Cumulative density Function) or Distance relative to ownership of the class.
$D_h(r)$	CDF of r_h (distance Cumulative Density Function) or Distance relative to the other classes

by kernel-based estimation and numerically integrated in cdfs. Table I summarizes the symbols in (1).

The merging process by (1) can be carried out based on estimating and minimizing a per-pixel distance $r_i(o_i, q)$ to the nearest training element in either the feature or the geographic space, resulting in a pure multiclass classification or in a multiclass classification with a significant segmentation component related to the spatial dimension.

The authors of the methodology [10] observe that (1) can be understood by considering that in order for S to be a query answer, the distance of any object o_h (where $o_h \notin S$) from q must be greater than that of o_i where $o_i \in S$. At distance r , the pdf that object $o_i \in S$ has the k th shortest distance from q is the product of the following factors:

- 1) the pdf that o_i has a distance of r from q , i.e., $d_i(r)$;
- 2) the probability that all objects in S other than o_i have shorter distances than r , i.e., $\prod_{o_j \in S \wedge o_j \neq o_i} D_j(r)$; and
- 3) the probability that objects in $D - S$ have longer distances than r , i.e., $\prod_{o_h \in D - S} (1 - D_h(r))$.

The integration function in (1) is essentially the product of the above three factors. By integrating this function over $(0, +\infty)$, we obtain the probability that S contains the k nearest neighbors with o_i as the k th nearest neighbor. Finally, by summing up this probability value for all objects $o_i \in S$, (1) is obtained.

B. K-d Tree-Based Implementation

The authors observe in their contribution that (1) is inefficient to evaluate, requiring as it does the computation of the distance pdf and cdf of each object, a costly numerical integration over a large range.

The exploitation of efficient data structures such as K-d trees allows the developed system to perform efficiently to the point of supporting efficient queries across a network environment.

Pixel-based approaches [3] require processing very large data volumes. In this sense, to get an efficient response to the queries, data organization is critical. In particular, nearest neighbor search can benefit from hierarchical indexing structures. K-d trees are space partitioning data structures for point organization in k-dimensional Euclidean spaces. They are based on sets of hyperplanes each perpendicular to one of the axes of the coordinate system. All nodes in the tree, including root and leaves, store a point and a space-dividing hyperplane.

To efficiently find the nearest neighbors, it is necessary to define a local search scope which is accomplished by the K-d tree. The key benefit is the reduction in the computational cost to find the nearest neighbor from $O(n)$ to $O(\log(n))$ in the average case. This significantly improves the performance when dealing with large data archives. The tree construction algorithm used is described in Maneewongvatana and Mount [11].

In the thematic layer generation process, the user selects different training regions. As we have seen, the result of this selection is modeled as a combination of random variables in a feature space with an associated pdf. This requires the use of proper kernel estimation techniques to go from training histograms obtained by selecting areas of interest to full pdf estimations.

This approach uses different generalization radius parameters, for each of the training points. The final results are obtained by operations on the generated K-d trees.

If a multiclass problem is considered, a situation in which a same pixel is classified in different classes needs to be solved by (1).

C. Implementation Details

In an actual optimized algorithm execution flow, (1) needs to be computed repeatedly for all the pixel regions to be classified. A caching mechanism based on memorized functions is used in order to avoid recomputing results, as in dynamic programming schemes. In order to reduce processing costs, the integral is computed as a quantized sum over the space of distances. The processing cost is further reduced by only computing the distances for couples that are nearby according to batched queries to a K-d tree instantiated based on feature values for user-provided supervision training areas.

D. Data Characteristics and Primitive Features

Advances in remote-sensing technology have improved in quality and quantity of the images that we have available. Until recently, the decametric resolution of this kind of images has limited observable classes to urban areas, forest, agricultural areas, bare soil areas, and water bodies. With metric resolution images, the development of new strategies and methodologies is necessary.

In this contribution, we consider data available in the Open Data Euskadi repository.¹ In this specific work, we considered multiple test sites, with a composite size of about 25 000 × 5000 pixels, with a resolution corresponding to 25 cm in each

¹[Online]. Available: <http://opendata.euskadi.net/>

TABLE II
PRIMITIVE RADIOMETRIC AND GEOMETRIC DESCRIPTORS WITH
EXTRACTION PARAMETERS

Descriptor name	Analysis region size	Angular quantization	Reference
HSV	1 × 1	None	[26]
Histograms of oriented Gradients (HOG)	12 × 12	8	[27]
Local binary Patterns (LBP)	24 × 24	8	[27]
Right-angle Detector (LSD)	12 × 12	4	[28]
Edge density	12 × 12	None	[29]
Sift density	24 × 24	None	[28]

direction. As is typical of image acquisition systems with very high geometric resolution, the radiometric resolution of the acquired data is limited to a limited number of channels.

CBIR literature typically devotes significant efforts to the careful choice and implementation of image content descriptors [12].

The literature of image analysis includes a diverse gamut of content-based primitive feature extractors, ranging from pixel-based descriptors like color to geometrical ones such as texture [13]–[15]. The use of combinations of these features is also usual [1], [16].

Global and image-level descriptors are often complemented by local ones. While the former ones have properties desirable for the discrimination of the semantic context of the scene, the latter ones enable the characterization and recognition of specific elements of the scene. The proper composition of discrimination strategies at the semantic context and at the object level is the subject of a large corpus of research [1], [17], [18].

As indicated by state of the art results in metric resolution classification for remote-sensing applications [19], except the HSV color-based descriptor the considered primitive features are region-based: histograms of oriented gradients (HOGs), local binary patterns (LBP), a right-angle/line segment detector (LSD), edge density, and SIFT.

In [20] land cover changes of the last 40 years in the country of Mali are analyzed. Object-based feature extraction and supervised (maximum likelihood) and unsupervised (ISODATA) classification are used to this end on high resolution panchromatic and multispectral remote-sensing imagery.

A framework is presented on [21] for building extraction from visible band images. Combining supervised and unsupervised classification, accurate rooftop extraction is achieved using a Higher order Conditional Random Field.

Another framework is presented in [22], where weakly supervised learning and high-level feature learning layers are combined: SIFT descriptors are clustered by K -means and fed to deep Boltzmann machines to capture structural and spatial patterns.

In [23], the problem of learning high-level features from a limited labeled subset in a large amount of unlabeled data is addressed using semisupervised ensemble projection (SSEP). The proposed method represents an image by projecting it

TABLE III
CHARACTERISTICS OF THE TEST SITES USED IN THE EVALUATION

Site id	Site name	Center (Lat/Lon)	Site description	Ground cover classes
1	La Concha	43.3190, -1.9923	Bay area	Bare soil, beach, buildings, pasture, roads, sea, woodland and urban area.
2	Arratz-Erreka	43.0056, -2.4737	Mountain area	Bare soil, beach, buildings, fields, pasture, roads, scrubs, and, woodland
3	Barakaldo	43.2986, -3.0004	Industrial area	Buildings, industrial, roads, urban and water
4	Vitoria Gasteiz	42.8505, -2.6690	Mixed urban area	Buildings, roads, and urban
5	Urdaibai	43.3837, -2.6905	Estuary natural reservoir	Buildings, fields, pasture, roads, urban, water and woodland.

The five test sites represent a significant degree of contextual diversity as well as a significant number of specific ground cover classes. Rectangular bounding boxes are given as the latitude and longitude of the upper-left (north western-most) and lower-right (south eastern-most) points. Each of the five test sites has a geographical extension of 1.2×1.2 km², which corresponds to about 23.6 Mpixels for each of the five input regions.

onto an ensemble of weak training (WT) sets sampled from a Gaussian approximation of multiple feature spaces.

The feature extraction approach presented in [24] consists on five steps: 1) feature extraction; 2) feature learning; 3) feature encoding; 4) feature pooling; and 5) classification. The process starts with low-level feature extraction by, e.g., SIFT. Then, a set of normalized basis functions is computed by unsupervised learning. Orthogonal matching pursuit is used for coding the basic function set. Finally, the sparse features are pooled to create the final representation to be fed to a support vector machine classifier.

In [25], an extensive evaluation of SIFT local invariant features, is conducted for the retrieval of land cover classes in high-resolution aerial imagery, with a comparison with standard features such as color and texture.

We establish an extraction process that defines a common grid among the extracted descriptors, so as to allow the subsequent data fusion procedure. This requires a spatial resolution rescaling that we implement as a nearest-neighbor interpolation for lower-resolution descriptors. Descriptions for the primitive features, with corresponding extraction parameters including region sizes, are reported in Table II.

E. Training Strategy

As in [19], the training supervision is provided to the system in terms of polygon-bounded regions manually defined over specific single-class coverage areas in the input image, see Table IV.













Sampling without replacement is used to extract an equal number of samples (usually in the order of the tens of thousands) for all training classes. The extracted sample sets are used for estimating pdfs for the class-specific distributions via kernel-based methods.

III. PROTOTYPE IMPLEMENTATION

While Section II describes the methodological basis of the implemented service, this section details its implementation strategy and its architecture in the form of collaborating Web-based services. Proposed system architecture is presented in Fig. 1.

As stated in the paragraphs regarding motivation in the introduction, we consider a Web-based architecture for reasons of accessibility and horizontal scalability.

TABLE IV
TRAINING PATCHES BY CLASS WITH GLOBAL TRAINING SAMPLE SIZE

Class	Training patch	Class	Training patch
Bare soil		Roads	
Beach		Scrubs	
Buildings		Sea	
Fields		Urban	
Industrial		Water	
Pasture		Woodland	

The actual pixels for training are sampled without replacement in a number of 1024 from the above global training polygons which fall within the areas identified for the class in the ground-truth map. The total size of the training sample used is therefore of $1024 \times 12 = 12\,288$ pixels, about 0.01% of the total size of about 118 Mpixels for the whole pixel size of the evaluation test site set.

A. Architecture

A map server module manages the imagery to be used by the system, both in terms of quick-looks for representation and training, and as output created thematic layer tiles. It is based on TileStache, a Python-based server application that can serve map tiles based on rendered geographic data,² Mapnik, a free toolkit for developing mapping applications³ and GDAL, a translator library for raster and vector geospatial data formats.⁴ Zoom, drag, and drop operations are available. The system generates thematic tiling at different resolution levels based on input by the user. The classification system is coupled with a tiling service for optimizing time-to-display.

The Web-server module is based on the Flask micro framework.⁵

The processing server module is in charge of image processing and classification. It computes the needed K-d trees and

²[Online]. Available: <http://tilestache.org/>

³[Online]. Available: <http://mapnik.org/>

⁴[Online]. Available: <http://www.gdal.org/>

⁵[Online]. Available: <http://flask.pocoo.org/>

performs the actual classification based on the pixels selected by the user. The processing server receives processing requests from the client module, processes them effectively and provides the resulting tiles to the Map Server. The server also provides an identification number to the client that allows to request the created thematic tiles to the map server.

The client side is a Web-based graphical user interface. This interface, with a screenshot depicted in Fig. 4, is built around an interactive map view that supports supervised training according to the semantics of the thematic class of interest. A configuration panel presents a description of the training itself and allows the user to interactively manipulate some parameters of the learned model. Interaction is managed by the event handlers of the jQuery library.⁶ Workflow process scheme is presented in Fig. 2.

B. Processing Flow

In a Web-based environment, optimizing performance issues related to data communication and memory footprint in the client is of foremost importance. In the case in which the data has a volume that allows to store it in the memory of a single server [30], [31], static K-d trees can be computed. In the case in which layer data volume hinders agile management, a dynamic strategy is needed.

The developed solution tries to be simple and effective, creating only the needed K-d trees. The created layer is limited to the available area around the visible map in the browser. This strategy requires more communication between client and server, for the server to create and process the necessary K-d trees. As the user navigates the map, the client sends to the server the information related to the visualization area.

With this information the server is able to create the K-d trees related with the navigation. Click-and-drag operations in the client move the map view port as is typical of Web-based geospatial interfaces. Events that impose an extension or a recomputation of the live area under analysis are handled by spawning new processing requests to the server. The system configuration aims at reducing these requests to a minimum, while avoiding an excessive load on the client memory.

The supervisor is free to define a semantic class based on a probabilistic composition of simpler components. each one represented by a different K-d tree instance.

An example sequence diagram is represented in Fig. 3, from client request to the creation of a thematic map in the user interface. The description of the steps is as follows:

- 1) If they are not available, the server creates tiles corresponding to the current predefined active area.
- 2) After this, K-d tree indexes from tiles are calculated for training. At this point the server is ready.
- 3) The client requests a Web page from the Web server.
- 4) The Web server receives this request and responds with an HTML Web page with the information needed to create the map.
- 5) The client requests the required map tiles from the map server.

⁶[Online]. Available: <http://www.jquery.org>

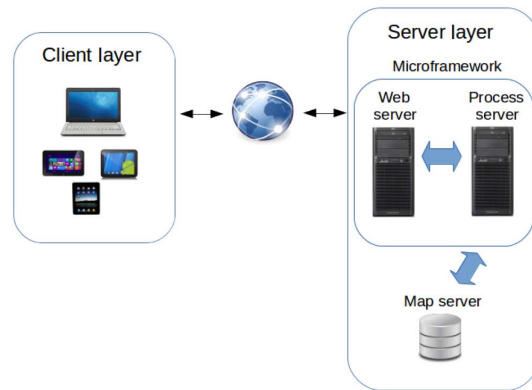


Fig. 1. Proposed system architecture in the form of different modular layers.

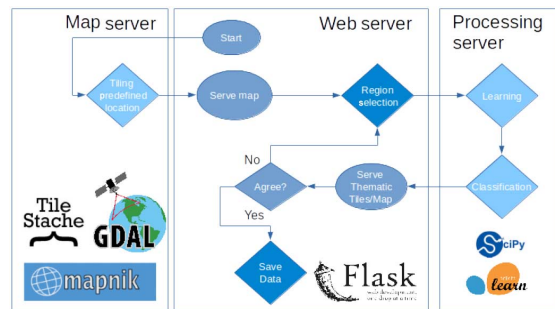


Fig. 2. Workflow process scheme for server-side operation: the map server component serves primitive feature tiles to the classification Web server that operates based on training supervision by the user. The processing server component performs the classification and serves the resulting thematic map tiles to the user for optimizing time-to-display. A complete sequence diagram including client-side operation is represented in Fig. 3.

- 6) The map server returns the needed tiles. Now the client is ready.
- 7) The user navigates around the map searching for instances of the target class.
- 8) The user selects pixels according to the semantics of the search, and can subsequently tune configuration parameters of the model.
- 9) Once the client completes the pixel selection phase, it requests the new thematic layer of the active area. Training and model configuration data are sent by AJAX asynchronous requests, allowing the application not to wait the end of response data transmission.
- 10) When the processing module receives the data, it checks if needed K-d trees are created or not, to request any required tiles to the map server.
- 11) Then needed K-d trees are processed with user selected pixel data. This process creates a tile with the nearest neighbor pixel class corresponding to the training pixels.
- 12) The tiles are stored in the map server.
- 13) Once all thematic tiles are created, an identification number is returned by the processing server to the client.

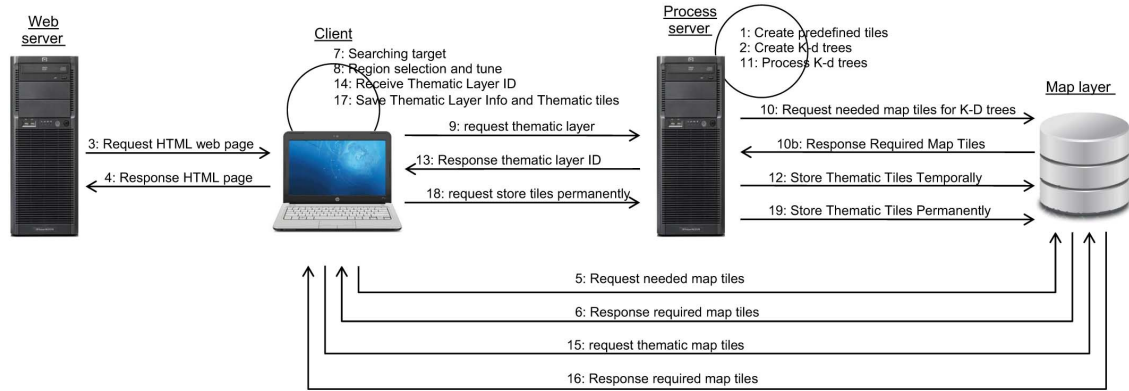


Fig. 3. Client/server thematic map creation sequence diagram.

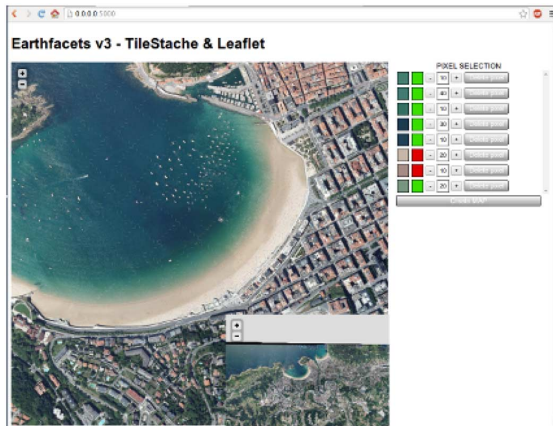


Fig. 4. User interface screenshot. An interactive map viewer supports supervised training and output presentation. A configuration panel to the right allows the user to interactively manipulate the parameters of the learned model.

- 14) The client receives this identification number and is able to activate or deactivate the created thematic layer.
- 15) The client requests new thematic tiles.
- 16) The map server returns them.
- 17) The client is able to save the created thematic layer. It locally saves the data needed to create the thematic tiles.
- 18) The client potentially creates a save request.
- 19) The processing server receives a save request and requests to the map server to save the new tiles.

IV. EVALUATION METHODOLOGY

The end-to-end validation of the system naturally focuses on the performance evaluation of the implemented classification system, since its operation involves all subsystems in the prototype. This performance evaluation is conducted as is customary by analyzing the quality of thematic map images produced based on a well known input.

The analysis is carried out on five separate test sites located in the Basque country region Fig. 5, each with an extension of

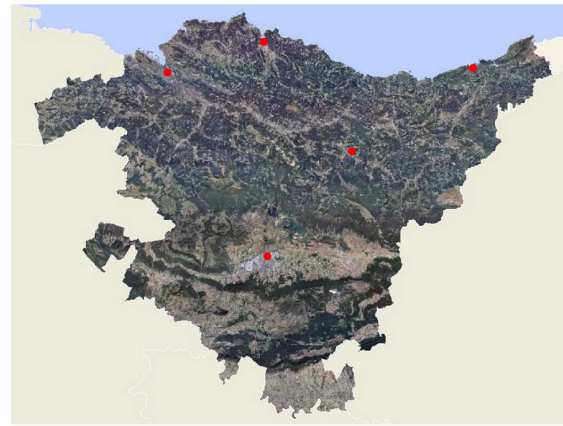


Fig. 5. Geographical location of test sites in the Basque country. From left to right, top to bottom: Bilbao industrial site, Urdaibai estuary protected area site, La Concha bay site in Donostia San Sebastian, Arratz Erreka mountain area site and Vitoria Gasteiz urban area site. These areas include the 12 different ground cover classes considered: beach, buildings, fields, industrial area, bare soil, pasture, scrubs, sea, urban area, urban roads, water. See Table III for the characteristics of the sites.

4864 × 4864 pixels (about 1.2 km × 1.2 km each). A composition is shown in Fig. 6(a).

The sites include the 12 ground cover classes considered and corresponding to layers in reference geographical maps extracted from the Open Data Euskadi repository⁷ managed by the Basque regional government. The classes correspond to sea, water, woodland, bare soil, urban, pasture, scrubs, fields, industrial, buildings, roads, and beach.

A. Vector-Map-Based Ground-Truth Map Creation Procedure

This section describes the generation of ground-truth maps from the vector maps available in the Open Data Euskadi repository.

⁷[Online]. Available: <http://opendata.euskadi.net/>



Fig. 6. Classification results. (a) Composition of the five original images of the test sites relative from left to right to La Concha bay, to the Arratz Erreka high plane, to the industrial area in the outskirts of Bilbao, to the Vitoria Gasteiz mixed urban environment and to the Urdaibai estuary protected area. (b) Corresponding composition of the five ground-truth maps based on Open Data Euskadi WMS shape files is shown in line, while (c) classification results are in line. Line (d) presents from left to right the original data, the vector-map-based ground-truth and the classification result at full resolution for the area marked with a red rectangle in the Vitoria Gasteiz ground-truth map in line (b). The classified maps are with a number of pepper-and-salt effects. In addition, some spectrally similar objects are not well identified and discriminated, such as buildings-roads-soil.

In addition to submetric resolution aerial ortho imagery of all territory, geographical information related to ground cover and usage is available in the form of vector maps. With this information, it is possible to build ground-truth models of the surface of the Basque country.

The challenge here is to complete a ground-truth model that covers all the area in the test sites to analyze, merging different categories in existing layers to obtain the most descriptive map of the area. This procedure is carried out manually: some of the

categories overlap each other and some do not appropriately cover the test sites.

The obtained ground-truth models for the test sites presented in Fig. 6(a) are shown in Fig. 6(b).

Two fundamental problems arise with the significance of the available ground truth with respect to the available imagery: a temporal and a spatial one.

A first issue is the temporal date of reference for the maps. In general, cover maps do not correspond in this respect to the

aerial imagery. A clear example of this is visible in the low tide image of the river mouth in the Urdaibai estuary site, which is represented as fully flooded in the maps.

A second issue is that the level of detail of vector data maps does not typically match that of aerial imagery with 25 cm pixel spacing, which is bound to have impacts in the performance measures. If Fig. 6(a) and (b) are compared, it is easy to detect some differences: most of the green areas in the city areas are not represented in the maps, different kinds of vegetation can be seen in the imagery that are fused in the Arratz Erreka site vector maps under the same label.

To overcome these limitations, in addition to the vector map-derived ground truth, the production of a further pixel-level, image interpretation-based ground-truth model is considered, to be able to compare obtained results with more detail.

B. Pixel-Level, Image Interpretation-Based Ground-Truth Map Creation Procedure

The definition of a multiple-class pixel-level ground-truth map based on the interpretation of submetric resolution imagery represents a significant challenge. A combination of specific training, semiautomatic tools, and careful inspection of the results are important components. A very good knowledge of and accessibility to the chosen test area are needed.

An area for testing and validation has been defined on the La Concha bay in the city of Donostia San Sebastian, where the authors are located so that field inspections can be used whenever necessary to verify the obtained results. The bay gathers different spatial contexts in a limited extension, which makes it particularly interesting as a testing location.

A first step is the selection of a representative set of semantic classes with clear meaning. In the case of our map, the eight selected classes are beach, buildings, bare soil, pasture, roads, sea, woodland, and urban area. Although these classes only represent an approximation to the twelve classes considered in the case of the vector map-derived ground truth, we still consider the set to be significant in the sense that the classes properly represent all essential visible content in the input data, and it to be orthogonal in the sense that their semantic separation is sufficiently large as to avoid significant overlaps and uncertainties in the corresponding feature space.

A well-defined procedure needs to be set up for generating an output thematic map with these characteristics from the input data. The procedure needs to exploit semiautomatic tools to generalize and extend training input provided by a human supervisor in order to speed up the overall process. The training is provided in the form of polygons covering a significant area of a given scene object. These elementary training areas need to have a sufficient geometrical extension to be able to express significant statistical descriptors from them. These statistics are computed in terms of color content as defined and simplified by means of a vector quantization with a number of levels in the order of the hundreds. To actually perform the geometrical extension of the training polygon to the observable limits of the considered scene object, the semiautomatic tools used include an edge-based segmentation routine that is launched in conjunction with every training event.

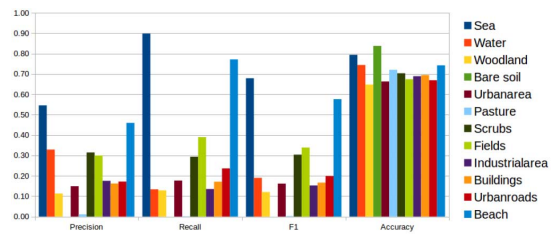


Fig. 7. Performance results for the full set of five test sites and of the 12 ground cover classes considered. Low values for classes like Beach and Water are directly related to temporal variations observable in the imagery with respect to the reference maps and to the different level of detail considered, see Fig. 4.

Once multiple tentative single-class thematic maps are defined, a procedure is needed to carefully compose them into a multiclass map. A further semiautomatic procedure is employed to highlight areas assigned to multiple classes as well as unassigned areas. The pixels in these areas are subject to an arbitration procedure in order to assign them unambiguously to a single ground-truth thematic class.

An extensive and labor-intensive supervised validation phase ensues in which the produced tentative multiclass map is inspected for dubiously labeled pixel areas.

The produced ground-truth map (Fig. 8), is currently published as open data at the URL <http://150.241.250.4:5000/earthfacets/groundtruthmap.png>. The authors hereby invite external users, specific corrections and general suggestions for improvement in the results or in the overall procedure.

C. Ground-Truth-Based Evaluation Procedure

The generated ground-truth map can be used for evaluating the performance of the system. The procedure we employ to evaluate quantitatively the obtained results can be described as follows.

- 1) Load training in the form of user selected pixel data.
- 2) The K-d trees of the tiles that compose the analysis area are calculated.
- 3) Each K-d tree is processed and a new tile is created and stored. If the training considers a multiclass problem, a new tile is created for each class.
- 4) A unique map is composed with the newly created tiles.
- 5) A confusion matrix is computed by comparing the classification output with the ground-truth map.
- 6) Performance statistics (precision, recall, F-1 measure, and accuracy, see Section V, Section IV-D) are computed from the confusion matrix.
- 7) Finally, the time required for completing the process is computed.

The designed solution allows us to experiment with different user selection options and obtain quantitative results for these selections in an easy and quick way.

D. Considered Performance Measures

To evaluate system performance we select the following statistical measures of information retrieval performance:



Fig. 8. From left to right: original image, pre-existing vector-map-based ground truth, manually produced pixel-level image interpretation-based ground truth for the La Concha test site. The subset of the defined ground cover classes includes eight elements from the original 12: building (purple), sea (blue), bare soil (light green), pasture (green), woodland (dark green), street (gray), urban roads (dark gray), and beach (brown). An increased level of detail is evident, particularly, in vegetated areas between buildings and near to Bare Soil areas.

TABLE V
PERFORMANCE MEASURES FOR THE COMPLETE TEST SITE COLLECTION OF FIVE AREAS BASED ON THE VECTOR MAP-BASED GROUND TRUTH WITH 12 GROUND COVER CLASSES

Performance statistics	Sea	Water	Woodland	Bare soil	Urban	Pasture	Scrubs	Fields	Industrial	Buildings	Roads	Beach
Precision	0.70	0.33	0.27	0.00	0.29	0.00	0.03	0.30	0.09	0.46	0.21	0.17
Recall	0.90	0.13	0.13	0.00	0.18	0.00	0.29	0.39	0.13	0.17	0.24	0.77
F1	0.79	0.19	0.17	0.00	0.22	0.00	0.05	0.33	0.10	0.25	0.22	0.28
Accuracy	0.83	0.70	0.58	0.83	0.57	0.86	0.80	0.68	0.72	0.66	0.61	0.74

Results are markedly inferior to the ones obtainable with reference to the manually curated image interpretation-based ground-truth map limited to the La Concha Site, due to differences in both the time reference chosen and the level of spatial detail considered.

precision, recall, F1, and accuracy [32]. The definition of the measures is as follows:

$$Precision = \frac{tp}{tp + fp} \quad (2)$$

$$Recall = \frac{tp}{tp + fn} \quad (3)$$

$$F1 = 2 \frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}. \quad (5)$$

In the above, tp is the number of true positive cases, tn is the number of true negative cases, fp is the number of false positive cases, and fn is the number of false negative cases.

V. PROTOTYPE EVALUATION RESULTS

A. End to End Classifier Evaluation on Complete Test Site Set

Quantitative performance measures for the classifier on the whole set of 12 ground cover classes as evaluated on the whole test site set of five areas is described in Table V and Fig. 7.

The ground cover class with the best results is sea, with scrubs, and fields showing more limited performance and with only limited results for classes like industrial and bare soil.

An inspection of the semiautomatic ground truth based on existing vector-maps shows that the results are probably significantly affected by the characteristics of the original map data, and in particular by both the limited detail available in rural areas (hence the good results for scrubs and fields) and by the significant overgeneralization for very diverse classes such as industrial. Further effects include a mismatch in between the reference dates for the maps and the image acquisition, which shows up clearly in coastal areas subject to rapid change such as the Urdaibai estuary.

This points to the need, addressed in upcoming sections, for an evaluation with respect to both a vector-map-based ground truth and a pixel-level one.

B. Evaluation of Different Feature Approaches

Classified thematic layers per each class are merged by (1) based on estimating and minimizing a per pixel distance to the nearest training element in either the feature or the geographic space, resulting in a pure multiclass classification or in a multiclass classification with a significant segmentation component related to the spatial dimension.

An example map obtained from the N-class classification process based only on color descriptors is presented in Fig. 5(b). Typical accuracy measures are around 85% for most of the classes. A better definition of the ground coverage classes can

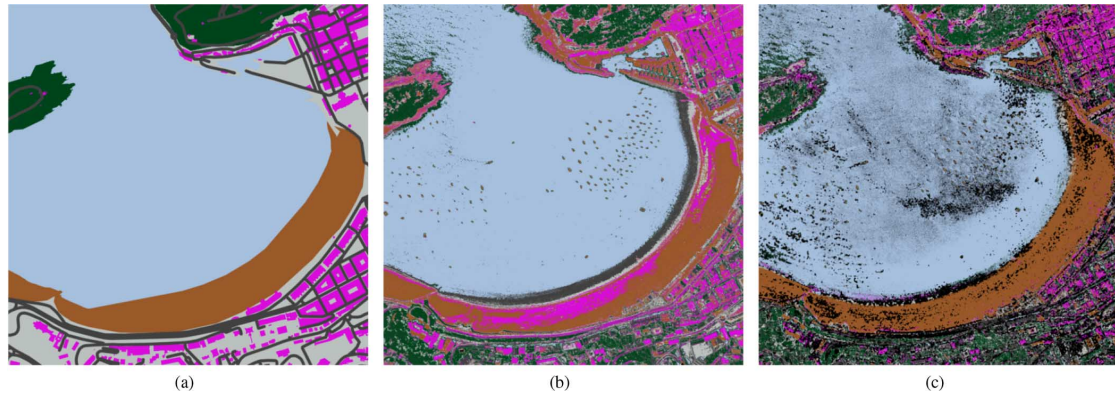


Fig. 9. (a) Ground truth based on preexisting vector map, (b) supervised classification results from color descriptors only as well as from (c) the full set of descriptors in Table II. Color encodes assigned class: building (purple), sea (blue), woodland (dark green), street (gray), urban roads (dark gray), and beach (brown). A portion of the output pixels remains unclassified (black).

be obtained by extending the training. Unlike the old town buildings which are well classified thanks their tiled roofs, the newer building area is characterized by lesser performance due to the mixed pattern of the roofs. Class street is another case for which the diversity of the patterns implies that a good characterization is more difficult to obtain.

The corresponding map obtained by considering all descriptors jointly including all the region-based ones as per Table II is in Fig. 5(b) and Fig. 10. The relative quality measures clearly show results comparable to the ones observable for the usage of purely color-based descriptors, with improvements for classes like beach that tend otherwise to be confused with the colorimetrically similar yet geometrically separate building roofs.

For the construction of a test site set, we consider five different geographical areas in the Basque country with different kinds of surfaces, as per Fig. 6. The selected areas represent a coastal bay (the La Concha beach area), a highlands area, an industrial area in the outskirts of the city of Bilbao, an urban area in the city of Vitoria Gasteiz, and a site in the protected natural area of Urdaibai, geographically located in correspondence to the red points in the map in Fig. 5.

The obtained classification result is shown in Fig. 6(c).

C. Local Classifier Evaluation Based on Both Semiautomatically Generated, Vector Map-Derived, and Manually Curated, Pixel-Level, Image Interpretation-Based Ground Truths

As described in Section IV-A, we complement the evaluation of the supervised classification system with respect to a five-site map-based ground truth with a more localized pixel-level ground truth obtained by manual interpretation of the image content.

Because of the difficulty of developing a pixel level ground truth of all areas, the analysis has been limited to one site of the five considered in previous sections. Among the five sites, we consider the La Concha bay site since it is the most complex with eight ground coverage classes: 1) beach; 2) buildings;

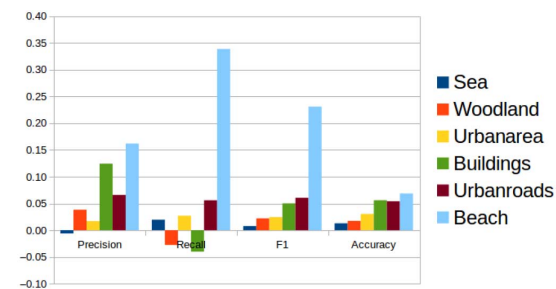


Fig. 10. Difference of quality measures by class for a classification based on color descriptors exclusively and on full set of descriptors. The results are similar, with marked improvements for homogeneous classes like beach, and a significant decrease for buildings, characterized by a very large internal variability that typically is not accounted for by the training.

3) bare soil; 4) pasture; 5) woodland; 6) roads; 7) sea; and 8) urban area.

The obtained results are shown in Table VI and Fig. 11.

Performance measures obtained with reference to the image interpretation-based ground truth tend to show improvements between 20% and 60% with respect to the ones obtained with reference to the map-based ground truth, because of both the availability of details that are not available in the maps and of the fact that the time of reference considered matches the image acquisition time, which is important in the case of rapidly changing sites such as the Urdaibai estuary. Class pasture is an example of this, where city gardens and other trees groups are not identified on shape base ground truth. Urban area and urban road classes present a decrement of precision related to the increment of false positive cases related to the different tagging in the two ground-truth maps, see top right corner of center and right images in Fig. 8.

Thematic map generation by the system requires about one minute for the whole 28 MPixel image considered on a 2.0-GHz Intel TM I7 4750HQ with 8 GB of RAM and SSD disk equipping a laptop computer.

TABLE VI
PERFORMANCE MEASURES FOR THE LA CONCHA BAY SITE WITH RESPECT TO THE MANUALLY CURATED PIXEL-LEVEL IMAGE INTERPRETATION-BASED GROUND TRUTH

Performance statistics	Sea		Woodland		Bare soil		Urban area		Pasture		Buildings		Urban roads		Beach	
	Vector	Pixel	Vector	Pixel	Vector	Pixel	Vector	Pixel	Vector	Pixel	Vector	Pixel	Vector	Pixel	Vector	Pixel
Precision	0.99	0.99	0.28	0.48	0.00	0.06	0.37	0.22	0.00	0.55	0.52	0.66	0.03	0.16	0.69	0.73
Recall	0.90	0.87	0.34	0.48	0.00	0.08	0.15	0.14	0.00	0.07	0.29	0.30	0.05	0.30	0.77	0.92
F1	0.94	0.93	0.31	0.48	0.00	0.07	0.22	0.17	0.00	0.13	0.37	0.41	0.04	0.21	0.73	0.81
Accuracy	0.91	0.90	0.89	0.91	0.96	0.95	0.84	0.87	0.99	0.94	0.91	0.91	0.89	0.92	0.91	0.94

These results are obtained by only considering the La Concha Bay Test Site, hence the number of ground cover classes is reduced from the original 12 to only 8.

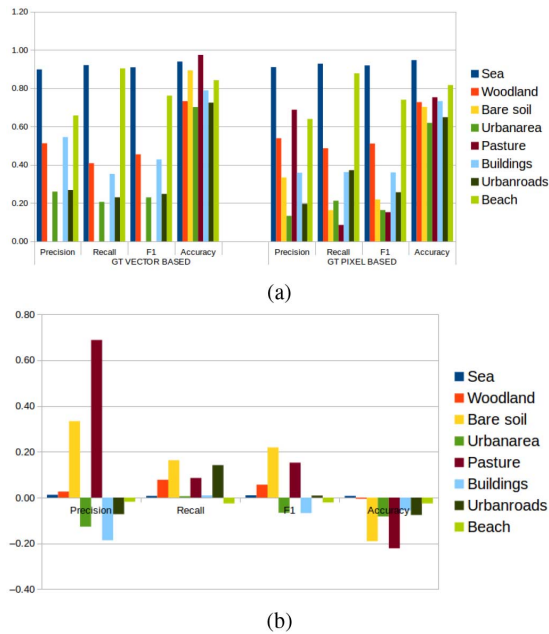


Fig. 11. Quality measures by ground cover class for a supervised classification with respect to two different ground-truth maps (existing vector map-based on the left, and manually produced pixel-level image interpretation based on the right) (a) and the difference of obtained results (b). These results are obtained by only considering the La Concha bay test site, hence the number of ground cover classes is reduced from the original 12 to only 8. Although they cannot be directly extrapolated to the full test site set, they indicate that the lack of spatial detail and the choice of a different temporal reference with respect to image acquisition can account for a 20 to 60% of difference in the obtained performance measures. Ground cover class Pasture is an example of this: gardens and small trees groups are not identified in the map-based ground truth.

VI. CONCLUSION

We have presented a prototype for thematic mapping from remote-sensing raster data.

The prototype is based on a probabilistic k-nearest neighbor supervised classification algorithm integrated in a simple Web-based architecture, and attains fast processing performance by exploiting N-dimensional data indexing structures, with the final aim of allowing users to interactively navigate and semantically map large extensions of geospatial data.

Results are promising for submetric airborne optical sensor data. The enrichment of the ground truth with the onset of

new classes and improvement on performance measures validated pixel-based ground-truth creation. Implementations on top of “big data” cluster computing framework will need to be considered to further enhance scalability.

REFERENCES

- [1] M. Quartulli and I. G. Olaizola, “A review of EO image information mining,” *ISPRS J. Photogramm. Remote Sens.*, vol. 75, pp. 11–28, Jan. 2013 [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0924271612001797>
- [2] M. Koubarakis *et al.*, “Building virtual earth observatories using ontologies, linked geospatial data and knowledge discovery algorithms,” in *On the Move to Meaningful Internet Systems: OTM 2012*, R. Meersman *et al.*, Eds. New York, NY, USA: Springer, 2012, vol. 7566, pp. 932–949.
- [3] M. Schröder, H. Rehrauer, K. Seidel, and M. Datcu, “Interactive learning and probabilistic retrieval in remote sensing image archives,” *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 5, pp. 2288–2298, Sep. 2000.
- [4] H. Costa, H. Carrão, F. Baçao, and M. Caetano, “Combining per-pixel and object-based classifications for mapping land cover over large areas,” *Int. J. Remote Sens.*, vol. 35, no. 2, pp. 738–753, 2014.
- [5] S. W. Myint, P. Gober, A. Brazel, S. Grossman-Clarke, and Q. Weng, “Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery,” *Remote Sens. Environ.*, vol. 115, no. 5, pp. 1145–1161, 2011.
- [6] U. Maulik and A. Sarkar, “Efficient parallel algorithm for pixel classification in remote sensing imagery,” *Geoinformatica*, vol. 16, no. 2, pp. 391–407, 2012.
- [7] Q. Ho, P. Lundblad, T. Aström, and M. Jern, “A web-enabled visualization toolkit for geovisual analytics,” *Inf. Vis.*, vol. 11, no. 1, pp. 22–42, 2012.
- [8] P. Keel, “Collaborative visual analytics: Inferring from the spatial organization and collaborative use of information,” in *Proc. IEEE Symp. Visual Anal. Sci. Technol.*, Oct. 2006, pp. 137–144.
- [9] A. Ferran, S. Bernabe, P. G. Rodriguez, and A. Plaza, “A web-based system for classification of remote sensing data,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 4, pp. 1934–1948, Aug. 2013.
- [10] R. Cheng, L. Chen, J. Chen, and X. Xie, “Evaluating probability threshold k-nearest-neighbor queries over uncertain data,” in *Proc. 12th Int. Conf. Extending Database Technol. Adv. Database Technol.*, 2009, pp. 672–683.
- [11] S. Maneewongvatana and D. M. Mount, “On the efficiency of nearest neighbor searching with data clustered in lower dimensions,” in *Proc. Int. Conf. Comput. Sci. I*, vol. 2073, V. N. Alexandrov, J. Dongarra, B. A. Juliano, R. S. Renner, and C. J. K. Tan, Eds. New York, NY, USA: Springer, 2001, pp. 842–851.
- [12] M. Datcu *et al.*, “Information mining in remote sensing image archives: System concepts,” *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 12, pp. 2923–2936, Dec. 2003.
- [13] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content-based image retrieval at the end of the early years,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.
- [14] R. Datta, D. Joshi, J. Li, and J. Z. Wang, “Image retrieval: Ideas, influences, and trends of the new age,” *ACM Comput. Surv.*, vol. 40, no. 2, pp. 5:1–5:60, May 2008.
- [15] J. M. Peña-Barragán, M. K. Ngugi, R. E. Plant, and J. Six, “Object-based crop identification using multiple vegetation indices, textural features and crop phenology,” *Remote Sens. Environ.*, vol. 115, no. 6, pp. 1301–1316, 2011.

- [16] M. Blume and D. R. Ballard, "Image annotation based on learning vector quantization and localized haar wavelet transform features," *Proc. SPIE* vol. 3077, pp. 181–190, 1997.
- [17] I. G. Olaizola, M. Quartulli, J. Florez, and B. Sierra, "Trace transform based method for color image domain identification," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 679–685, Apr. 2014.
- [18] I. G. Olaizola, G. Marcos, P. Kramer, J. Florez, and B. Sierra, "Architecture for semi-automatic multimedia analysis by hypothesis reinforcement," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB'09)*, 2009, pp. 1–6.
- [19] N. Chauffert, J. Israel, and B. Le Saux, "Boosting for interactive man-made structure classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS'12)*, 2012, pp. 6856–6859.
- [20] R. Spiekermann, M. Brandt, and C. Samimi, "Woody vegetation and land cover changes in the Sahel of Mali (1967–2011)," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 34, pp. 113–121, 2015 [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0303243414001718>
- [21] E. Li, J. Femiani, S. Xu, X. Zhang, and P. Wonka, "Robust rooftop extraction from visible band images using higher order CRF," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4483–4495, Aug. 2015.
- [22] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.
- [23] W. Yang, X. Yin, and G. Xia, "Learning high-level features for satellite image classification with limited labeled samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4472–4482, Aug. 2015 [Online]. Available: <http://dx.doi.org/10.1109/TGRS.2015.2400449>
- [24] A. Cheriyyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 439–451, Jan. 2014.
- [25] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 818–832, Feb. 2013.
- [26] C. Pohl and J. Van Genderen, "Review article multisensor image fusion in remote sensing: Concepts, methods and applications," *Int. J. Remote Sens.*, vol. 19, no. 5, pp. 823–854, 1998.
- [27] M. Molinier, J. Laaksonen, and T. Hame, "Detecting man-made structures and changes in satellite imagery with a content-based information retrieval system built on self-organizing maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 861–874, Apr. 2007.
- [28] J. Inglada and E. Christophe, "The Orfeo toolbox remote sensing image processing software," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS'09)*, 2009, pp. 733–736.
- [29] X. Perrotton, M. Sturzel, and M. Roux, "Automatic object detection on aerial images using local descriptors and image synthesis," in *Computer Vision Systems*. New York, NY, USA: Springer, 2008, pp. 302–311.
- [30] J. Lozano, M. Quartulli, I. Tamayo, M. Laka, and I. G. Olaizola, "Visual analytics for built-up area understanding from metric resolution earth observation data," *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XL7/W2, pp. 151–154, 2013.
- [31] J. Lozano, N. Aginako, M. Quartulli, and I. G. Olaizola, "Semi automatic remote sensing image layer generator based on web based visual analytics," in *Proc. 5th Jubilee Int. Conf. Cartography GIS*, 2014, pp. 709–714.
- [32] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.



Javier Lozano Silva (S'15) received the Technical Engineering degree in telecommunication systems from the School of Engineering of Mondragon, Mondragon Unibertsitatea, Mondragón, Spain, from 2000 to 2004. He completed his studies and graduated from the Faculty of Telecommunication Engineering between 2004 and 2007 from the School of Engineering of Mondragon, Mondragon Unibertsitatea. In 2009, he received the Diploma degree in advanced studies from the System Engineering and Automation Department, University

of the Basque Country (UPV/EHU) University, Vizcaya, Spain. He is currently pursuing the Ph.D. degree in system engineering and automation at the University of the Basque Country, Bilbao, Spain.

Afterward, he developed the End of Degree Project in Orona EIC S.COOP about "The Enhancement of the Functionalities of Teleservicio, in 2005". Since February 2008, he has been working with Vicomtech-Ik4 Research Center (www.vicomtech.org), Donostia, Spain.



Naiara Aginako Bengoa received the Telecommunications Engineering degree from the University of the Basque Country, Vizcaya, Spain, in 2005. Currently, she is pursuing the Ph.D. degree in image analysis and content-based retrieval of images and video at the University of the Basque Country, San Sebastián, Spain.

From 2003 to 2005, she collaborated the Signal Processing and Communication Group, Electronics and Telecommunications Department with the University of the Basque Country, EHU/UPV, San Sebastián, Spain. She develops and manages research projects in Vicomtech-IK4, Donostia, Spain, since 2005, in the Digital Media Department. She is also a teacher with the Politechnique School of Donostia, Donostia, Spain, for more than 3 years. Her work as a Researcher counts on a number of publications and on two patents Biography text here.



Marco Quartulli (SM'03) received the Laurea degree in physics from the University of Bari, Bari, Italy, in 1997, and the Ph.D. degree in electrical engineering and computer science from the University of Siegen, Siegen, Germany, in 2005.

From 1997 to 2010, he worked on remote-sensing ground segment engineering, image analysis, archives, and mining with the Advanced Computer Systems, Rome, Italy. From 2000 to 2003, he was with the Image Analysis Group, Remote Sensing Technology Institute, German Aerospace Center (DLR), Cologne, Germany, working on metric resolution synthetic aperture radar image understanding in urban environments and content-based image retrieval. Since 2010, he has joined the Multimedia Services Department, Vicomtech-IK4, Donostia, Spain, where he is working on large-scale analytics for the multimedia and the geospatial analysis domains.



Igor G. Olaizola (SM'14) received the Degree in electronic and control engineering from the University of Navarra, Navarra, Spain, in 2001, and the Ph.D. degree in informatics from the Faculty of Informatics of San Sebastián, University of the Basque Country, Vizcaya, Spain, in 2014.

He developed his master thesis at Fraunhofer Institut für Integrierte Schaltungen (IIS), Erlangen, Germany, in 2001, where he worked for a year as Research Assistant on several projects related to MPEG standard audio decoding. Currently, he is the Head of the Digital Interactive TV and Multimedia Services Department. In 2006, he was a Technology Consultant for 1 year with Vilau Company, Zamudio, Spain.

Dr. Olaizola has been a member of Vicomtech Technological Centre since 2002.



Ekaitz Zulueta received the B.S. degree in electronic engineering from Mondragon University, Arrasate, Spain, in 1997, and the M.S. degree in electrical engineering from Swiss Institute of Technology Lausanne, Lausanne, Switzerland, in 2000, and the Ph.D. degree in control engineering from the University of the Basque Country, Vizcaya, Spain, in 2005.

From 2000 to 2002, he has been employed as a Research Engineer with Ideko, Elgoibar, Spain and Fagor Automation, Arrasate-Mondragon, Spain. Since 2002, he has been employed as a Lecturer with the University of the Basque Country (University College of Engineering of Vitoria-Gasteiz), Vizcaya, Spain. His research interests cover a range of computational intelligence areas including image processing and wind turbines control.

2.1.2 Visual Processing of Geographic and Environmental Information in the Basque Country: Two Basque Case Studies

- **Izenburua:** Visual Processing of Geographic and Environmental Information in the Basque Country: Two Basque Case Studies
- **Egileak:** Alvaro Segura, Aitor Moreno, Igor G. Olaizola, Naiara Aginako, Mikel Labayen, Jorge Posada, Jose Antonio Aranda, Rubén García De Andoin
- **Liburua:** GeoSpatial Visual Analytics
- **Argitaletxea:** Springer
- **Orrialdeak:** 199-207
- **Urtea:** 2009
- **DOI:** http://dx.doi.org/10.1007/978-90-481-2899-0_16

VISUAL PROCESSING OF GEOGRAPHIC AND ENVIRONMENTAL INFORMATION IN THE BASQUE COUNTRY: TWO CASE STUDIES

ALVARO SEGURA, AITOR MORENO, IGOR GARCÍA,
NAIARA AGINAKO, MIKEL LABAYEN, JORGE
POSADA¹

VICOMTech, Mikeletegi 57, San Sebastian 20009, Spain

JOSE ANTONIO ARANDA, RUBÉN GARCÍA DE
ANDOIN

*Meteorology and Climatology Department of the Basque
Country, Vitoria 01010, Spain*

Abstract. The Basque Meteorology Agency is conducting an initiative to improve the collection, management and analysis of weather information from a large array of sensing devices. This paper presents works carried out in this context proposing the application of 3D geographical visualization and image processing for the monitoring of meteorological phenomena. The tools described allow users to analyze visually the state of the atmosphere and its interaction with the topography, and process live outdoor images to automatically infer weather conditions. This kind of systems can be applied in the surveillance of other environmental events and enable better decision making for several purposes, including important issues related with environmental security.

Keywords: visual analytics, GIS, geographic information, computer graphics, weather, meteorology, environmental security.

¹ Jorge Posada, VICOMTech, Paseo Mikeletegi 57, E-20009 San Sebastian. Email: jposada@vicomtech.org

2 VISUAL GIS/METEOROLOGY PROCESSING IN THE BASQUE COUNTRY

1. Introduction

Geographic information processing together with computer graphics visualization and analysis has a considerable potential in environmental monitoring and decision making. The Basque Meteorology and Climatology Agency along with VICOMTech Research is conducting a strategic project that aims at establishing tools to centralize and enable analysis of the large amounts of data collected from weather sensors spread around the Basque Country. These sensors are heterogeneous devices such as those in 96 automated weather stations (temperature, pressure, humidity, wind, solar radiation, etc.), a Doppler radar, a wind profiler and several oceanic probes.

This paper presents work carried out in the context of this initiative that brings computer graphics-based tools to assist in analyzing the global situation. Properly using the data from sensors requires calibration procedures and filtering of noise to ensure reliability.

The first part of the work involves the creation of an integrated 3D geographic information system for visually analyzing weather data, mainly weather radar scans, together with other georeferenced information. The work starts with an analysis of output data from the weather radar located on Mount Kapildui and the task of improving the quality of the readings.

The second part proposes video cameras as additional weather sensors. Computer vision techniques can process images coming from cameras in the automated stations and provide information on the state of the local atmosphere. A coordinated operation of all stations with such a system installed could give a global depiction of the state of the sky along the territory and its evolution, potentially allowing the forecast of special environmental situations.

1.1. GEOGRAPHIC VISUALIZATION AND IMAGE ANALYSIS

The Autonomous Community of The Basque Country is a territory in northern Spain bordering with southern France. It spans an area of 7234 km² and is crossed by a few mountain ranges. Established 1990, the Basque Meteorology Agency, *Eus almet*, has deployed a large network of automated weather stations, including a long-range radar, and provides past, present and forecast meteorological information.

The physical data collected by sensors in the network needs to be stored and properly managed and retrieved to provide useful information (e.g. for decision making). This process involves the use of traditional tools as well as innovative computer visualization and image processing as key technologies. Visual analysis tools help users understand the state and

evolution of the environment by providing integrated graphical representations and visual metaphors.

In the case of weather and other environmental information we consider geographic (i.e. topographic) data especially relevant. These phenomena occur in specific locations and are influenced by the topography. We thus want to present incoming sensor data coupled with a detailed 3D representation of the territory in order to give it a context and allow a visual analysis of the interaction between ground and atmosphere.

The above mentioned system transforms numerical sensor readings into visual representations to enable humans to interpret them. This initiative also proposes a system working in a very different way: taking live images of the environment, as a human observer would, and automatically process and interpret them to infer the state of the environment.

Both approaches (producing visual metaphors for humans to interpret and letting computers interpret visual information) are different aspects of the application of computer graphics processing in environment related information analysis. Current focus is in meteorology, but a similar approach can be applied to other environmental monitoring such as forest fires, pollution or floods, all with implications in environmental security.

2. Weather radar data processing and visualization

The Basque Weather service operates a dual Doppler Weather Radar, located on top of Mount Kapildui, 1000 meters high and 100 km away from the coast. It is a Meteor 1500C model from Selex-Gematronik. The radar computes the reflectivity, radial velocity and spectral width fields every 10 minutes through two volumetric and two elevation scans.

Radar scans are typically represented as 2D images in the form of either PPI (plan position indicator) or CAPPI (constant altitude PPI) products. Here we want to display the complete radar volumes, not only individual slices from it, correctly aligned and scaled over digital terrain model of the territory. The result is a form of a geographic information system (Peuquet 2004).

2.1. VOLUMETRIC DATA ANALYSIS

The volumetric data sets acquired by the radar are composed of 14 scans at increasing elevations (from -1° to 35°). Given the topography of the Basque Country, the lower scans are affected by the surrounding mountains and other topographical elements, adding almost constant noise to the data, which should be ignored. This constant noise is known as ground clutter.

4 VISUAL GIS/METEOROLOGY PROCESSING IN THE BASQUE COUNTRY

Ground clutter is noticeable in the lowest elevations, since the radar beam frequently hits the topographical elements. On the other hand, the lowest levels give more useful information to meteorologists, so a compromise has to be found. Normally, the lowest elevation free of clutter is used as the main information source. In Mount Kapildui clean scans can be obtained at elevations greater than 1° . It would be better to have lower scans (at -0.5° , 0° and 0.5°), but this part contains noticeable ground clutter.

Since ground clutter is in theory constant in time its effects in the lowest scans can be reduced by subtracting a fixed mask to the retrieved data. Basically, this clutter mask consists of the reflectivity acquired in a clear day. Under those conditions, all perceived reflectivity should be caused by surrounding topography.

We have observed that ground reflectivity is not exactly constant but has slight random variations from one scan to another. This is probably due to slightly changing atmospheric conditions and small movements of the radar support structure.

2.2. CLUTTER MASK CREATION AND SUBTRACTION

Given the variability of radar echoes caused by topography a single scan of a clear sky is not enough to create a reliable clutter filter. In order to avoid this problem, a clutter mask was created through a combination of a small set of radar scans taken at different times with a clear sky.

The resulting mask effectively removes clutter from the scans used to produce it, by definition, but may not filter correctly all ground echoes in other scans due to those random variations. In order to increase its effectiveness, the mask is processed by a *dilate* filter. While this increases the risk of producing a filter which is too aggressive our preliminary tests seem to give acceptable results.

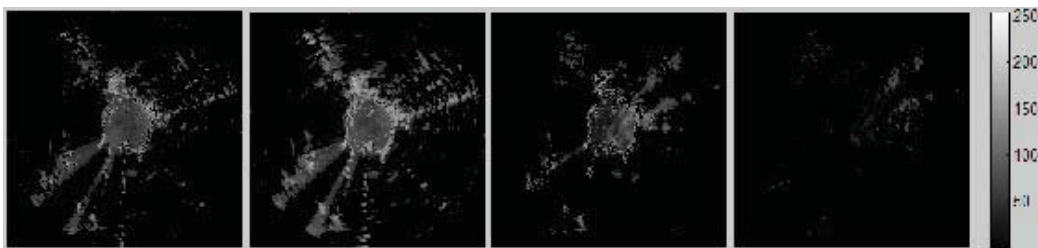


Figure 1: A low level scan subtraction. From left to right: a) original clutter mask, b) dilated clutter mask, c) a random volumetric scan and d) the subtraction of the scan and the clutter mask, removing the ground clutter.

Figure 1 shows a combination of several slices, its dilated form and an example of application (subtraction) from a new incoming reflectivity slice.

2.3. 3D VISUALIZATION ON GEOGRAPHIC MODEL

Our model of the Basque Country is based on a detailed digital elevation model (Jenson 1988) and a set of properly adjusted high resolution orthophotographs, provided by the Basque Government. In order to allow rendering at interactive rates the original elevation data in GeoTIFF format and the textures were processed to produce a set of hierarchically arranged tiles of varying resolution. The resulting data set, almost 1 gigabyte in size, enables progressive level of detail by retrieving the required terrain tiles on demand.

Volume scan files include metadata specifying the geographical location of the radar (longitude, latitude and elevation) and the sample separation. This information is used to position and scale the reflectivity field on the map. The map uses UTM coordinates and since the Basque Country is located nearly in the middle of zone 30T, very small scale distortion is expected.

The union of radar and topographic data clearly highlights the presence of ground clutter around the highest mountain ranges (see figure 2). The application currently also allows applying a precalculated clutter mask to remove such noise and produce cleaner precipitation representations.

Two visualization styles have been tried. In the simplest one, reflectivity is mapped to the opacity and greyscale intensity of all slices. In the second one, a standard reflectivity colour map is used, and values lower than 10 dBZ are completely transparent, which seems to be more intuitive to meteorologists.

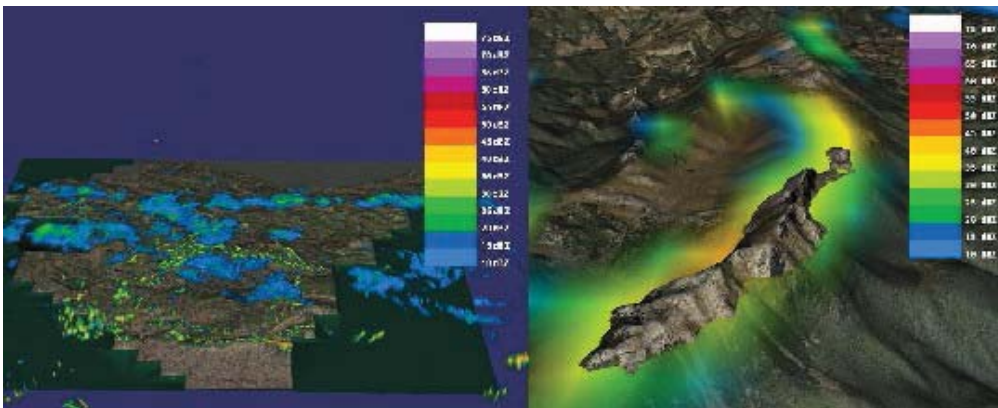


Figure 2: Unfiltered Kapildui radar volumetric information visualization using a reflectivity color map. In the left image, the rain areas can be seen in blue as well as ground clutter. In the right image, a close up of the ground clutter is shown, matching the mountain causing it.

3. Automatic analysis of sky images

The main goal of the Skeye project consists of the automated visual analysis of the images acquired by cameras located on ground stations. The Skeye architecture allows the integration of any analysis module and in this context we will focus on cloudiness estimation and fog detection that can provide information about the visibility condition in this area. Figure 3 shows the different modules developed to compose the Skeye system.

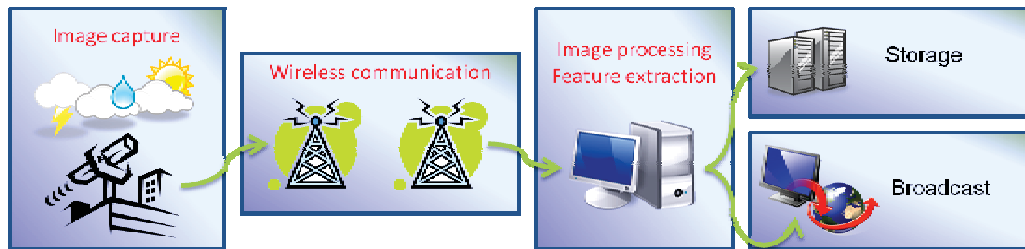


Figure 3: Skeye System Architecture

3.1. IMAGE CAPTURE

This module works at the terrestrial weather station. It takes pictures of the sky covering all elevation angles from -10° to 90° and 360° in azimuth.

Pictures can be taken in the visual electromagnetic spectrum or in infrared band. Infrared cameras have some advantages since they can work at night and the images provided contain thermal information, but on the other hand, texture-based analysis algorithms get less accurate data as input.

The amount of necessary images to cover the whole sky is inversely proportional to the field of view of the camera (Kelby, 2006). The quality and properties of the images depend on the camera parameter's settings which will be adjusted according to the environmental light conditions. The proper calibration of all these parameters will have a strong influence in the system's final precision and recall.

Data transmission is not a trivial aspect in our case. Terrestrial automated weather stations are usually placed at remote locations and the scalability of the system requires wireless solutions to keep the costs in a reasonable level. Therefore, this project has been coordinated with a WIMAX network deployment that will ensure the delivery of the information from all these remote stations. Mobile telephony networks such as GSM, GPRS, UMTS or HSDPA are also being considered depending on signal coverage.

3.2. IMAGE PROCESSING

This module centralizes all the information coming from the different terrestrial weather stations. It creates a 2D panoramic view of the whole sky dome keeping areas' relations using geometric transformations and Gall-Peters projection mode (Peters, 1983). This function allows representing the local weather conditions in a unique image and centralizing the visual information from the geographically separated places for the meteorologist.

Furthermore, it analyzes the images and extracts features using digital image processing techniques in order to segment the image and carry out the cloudiness calculation and fog detection which are processed by independent software modules.

For cloudiness, the image is segmented and labeled in 4 classes: *Earth*, *sun*, *sky* and *cloud*. Color and texture (entropy) features are used in this process (see figure 4).

The fog detector is based on the topographic outline analysis. The local terrain shape is analyzed and assumed to be fixed. Shape variations provide hints to detect fog which disturbs the terrain visual observation.

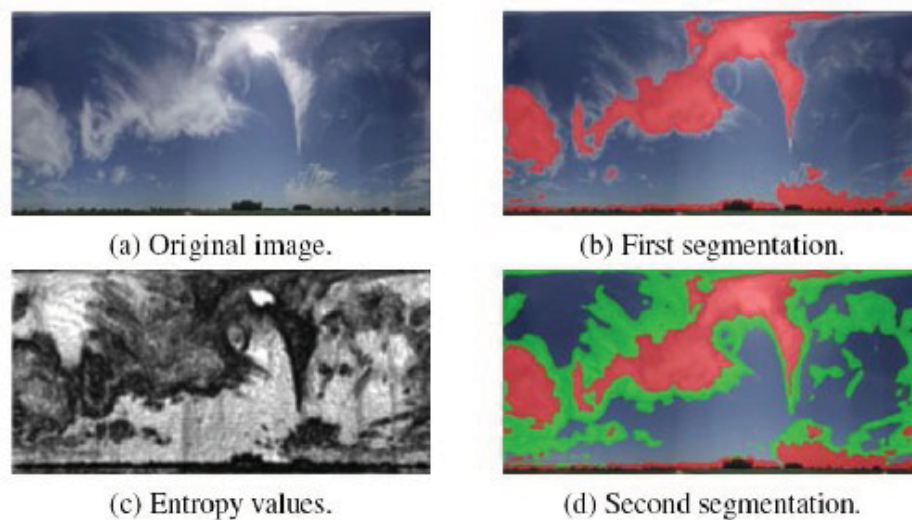


Figure 4: Cloud Segmentation Process

3.3. DATA STORAGE AND DELIVERY

The processed information is stored in common servers where different meteorological stations upload their data. This information can be accessed from anywhere and used to combine many data sources. It provides the way to find correlations among geographically separated weather phenomena and their effects or to track features in very wide areas.

4. Environmental monitoring

The potential of these two presented use cases operated by Euskalmet go much further than simple local weather analysis applications. The idea behind is a global approach to data analysis where different kinds of data (visual information, radar readings, digital terrain models, etc.) coming from different places can be combined in order to get a better understanding the state and evolution of environmental phenomena. These include weather alerts involving potential floods or fast temperature changes, chemical leaks to the atmosphere, forest fire and smoke, etc.

Current existing network infrastructures where communications costs are dramatically reduced by wireless technology deployments and the availability of a wide range of sensors and cameras provide a huge amount of data that after adequate preprocessing phases can be analyzed for different purposes. Data mining techniques can also help discover unknown correlations among geographically separated features and effects improving the knowledge of researchers and professionals.

Moreover, the network of stations can be considered as single entity able to carry out surveying activities of the areas covered by the network nodes. Some interpolation techniques could even find out effects produced in non monitored areas located among nodes.

5. Conclusions

A novel weather analysis system has been presented in this paper. In combination with classical weather instrumentation (thermometers, barometers, anemometers, etc.) the two use cases offer methods to improve forecast, general knowledge and environmental surveillance.

The Doppler weather radar visualization system with the explained clutter filtering techniques and more integrated sensors will provide the basis for a new data source to help prevent natural disasters like floods and big storms, and allows defining behavioral patterns.

The Skeye project defines a centralized image analysis framework where different cameras can be connected. All the visual information is processed by pre-calibrated analysis modules and cloudiness degree and fog presence can be automatically estimated.

6. Acknowledgements

The works here described have been funded by the Basque Government's ETORTEK Project (ISD4) and INTEK (SKEYE) research programs.

7. References

Kelby, S., 2006, *The Digital Photography Book*, Peachpit Press.

Peters, A., 1983, *Die neue kartographie/the new cartography*, Friendship Press.

Jenson, S. K., Domingue, J.O., 1988, *Extracting Topographic Structure from Digital Elevation Data for Geographic Information System Analysis*, *Photogrammetric Engineering and Remote Sensing*, Vol 54.

Peuquet, D. J., Marble, D.F., 2004, *Introductory Readings in Geographic Information Systems*, CRC Press.

2.1.3 Weather analysis system based on sky images taken from the Earth

- **Izenburua:** Weather analysis system based on sky images taken from the Earth
- **Egileak:** Mikel Labayen, Naiara Aginako, Igor G. Olaizola.
- **Proceedings:** Proceedings of the 5th International Conference Visual Information Engineering 2008 (Xian, China)
- **Orrialdeak:** 146-151
- **Urtea:** 2008
- **DOI:** <http://dx.doi.org/10.1049/cp:20080299>

Weather Analysis System Based on Sky Images Taken from the Earth

M. Labayen¹, N. Aginako¹, I. G. Olaizola¹

¹ VICOMTech, Mikeletegi Pasealekua 57, 2009 Donostia, Spain. {mlabayen, naginako, iolaizola}@vicomtech.org

Keywords: Image Processing, Meteorology, Fog Detection, Cloud Segmentation and Pattern recognition.

Abstract

Meteorological analysis has been typically associated to objective measures (temperature, humidity, air pressure, etc.) and observer's visual subjective analysis of the sky as cloudiness. Cloud amount influences strongly the radiation equilibrium, while cloudiness and fog can be seen as indicators of the type and intensity of locally active thermal processes. Land-based cloud analysis shows different local details which are unavailable in operational satellite imagery [14]. In addition, according to the available state of the art, cloudiness and fog analysis from the earth has not yet been automatized. In this study, an image based method to obtain automatic cloudiness quantification and fog detection is explained.

1 Introduction

Different meteorological devices and systems provide an automatic information about the same phenomenon: weather and climate. However, some measurements like cloudiness are estimated still by trained observers from a meteorological station on the earth.

This work aims to obtain information applying image analysis algorithms to images of the sky taken from a camera located in a terrestrial meteorological station. This information will mainly provide additional data regarding the local weather conditions based on cloudiness or fog detection.

In chapters two and three, this paper makes a short analysis of the state of the art and the description of the global system and its specifications consecutively. Afterwards, in chapter four, automatic feature extraction and the image processing functionalities are explained. Finally, in chapters five and six, the document shows the results obtained from the system test and it ends with the conclusions and future work.

2 State of the Art

Current state of the art in meteorological analysis made from the images is divided in two types according to from where

the images have been taken: based on satellite images and terrestrial images.

The analysis of images taken from satellites is an useful tool for meteorologist to weather prediction. It is the most developed automatic image analysis system for meteorological prediction purposes. The main researches in this field are made on cloud classification based on spatial textural and spectral measurements [9][2].

On the other hand, there are some experimental approaches regarding the analysis of sky images taken from the earth. They focus on cloud classification, using textural and spectral features (as edge sharpness, fibrousness and edge information) extraction techniques [14, 13, 5] or analyzing solar radiation measurements [7]. However, they do not offer any explicit solution for cloudiness quantification and fog detection which is the objective of this study.

3 System description

3.1 Overview

The goal of the system is to extract cloudiness and fog detection information from the whole sky image data. The system has to be able to capture images of the whole sky, as well as to extract useful information from the scene captured with the camera.

The general description of the system is shown in figure 1.



Figure 1: Block diagram.

3.2 Modules

The industrial project called *Skeye*, in which this research has been carried out, establishes the technical specifications that

the system has to fulfil. In this section the functionality of each system module as well as their specifications are summarized:

Image capture This module works at the terrestrial meteorology station. It takes pictures of the sky covering all the angles (100° in a vertical plane, from -10° to 90° and 360° panoramic view in a horizontal plane). The amount of necessary images to photograph the whole sky is inversely proportional to the equivalent focal length [8] of the lens.

The quality and properties of the images depend on how the camera parameters are set. The shutter speed and diaphragm aperture are configurable in order to provide following modules with images that improve the accuracy of the segmentation process. In addition, the white balance will be constant in order to achieve more homogeneity for the same colors in different images.

Wireless communication Terrestrial meteorological stations are usually placed at remote locations. Therefore, this project has been coordinated with a WIMAX network deployment that will allow the retrieval of the information provided by all these remote stations.

Image processing This module, explained in the next chapter, centralizes all the information coming from the different terrestrial meteorological stations. It is able to create a panoramic view of the whole sky keeping the physic areas relations using geometric transformations [10] and Gall-Peters projection mode [12]. This function allows to represent the local weather conditions in a unique image and centralize the visual information from the geographically separated places for the meteorologist.

Furthermore, it analyzes the images and extracts features using digital image processing techniques in order to segment the image and automatize the cloudiness calculation and fog detection.

4 Implementation of the Image Processing module

The most outstanding functionalities of the developed system are Equiareal panoramic representation, Cloudiness quantification and Fog detection.

4.1 Equiareal panoramic representation

The system has to be able to represent the environment around the terrestrial meteorology station on a two dimensional rectangular image. The state of the art presents some research

[3, 4] on image stitching based on characteristic point extraction. Unfortunately, the texture properties of sky images are too homogenous to allow good results from these methods.

For this reason, a new “panorama maker” which is independent of the content of the image has been created. This new system is based on the predefined position of the taken pictures. The direction where each image was taken from is the basic information used to align the different images in the final panoramic representation.

Geometric transformations are needed to project the three dimensional surface into a two dimensional plane keeping the physical areas relations between the elements presents in the environment (clouds, sun, sky and earth). Basic geometric transformations allow to align the different images into a panoramic representation and Gall-Peters projection mode projects the panoramic picture into a equiareal representation as follows:

$$\begin{cases} x = (\lambda - 180^\circ) \cos(\beta) \\ y = \sin(\varphi) \sec(\beta) \end{cases} \quad (1)$$

Where,

λ is the position respect a horizontal center of the projection.

φ is the position respect a vertical center of the projection.

$\beta = 44, 138^\circ$ for Gall-Peters projection mode.

Finally, multi-band Blending techniques [1, 11] are used in order to get a perfect fusion and a smooth transition between adjacent images.

As an output, the system provides a panoramic image which reflects a wider view than a single image taken by the narrow field of view lens. In addition, it has a higher resolution than images taken by wide field of view lens, as the fish-eye ones.

4.2 Cloudiness quantification

Cloudiness refers to the fraction of the sky covered by clouds. Traditionally, cloudiness is estimated by trained observers from a meteorological station on the earth and expressed in eighths. The aim of this section is to introduce a new computation method of cloudiness in numerical weather prediction models.

The mathematical definition of cloudiness is the cloud to sky ratio in the equiareal panoramic image. To be able to quantify the pixels of each region the application will segment the image in four classes: *earth, sky, clouds and sun*.

Firstly, the earth part of the image is segmented. According to experimental observations, if the image has been captured underexposing the scene, the histogram of the B channel

in a RGB representation presents easily separable two pixel densities. They represent the group of pixels belonging to the earth and the rest of the image.

Since the background where the earth silhouette is allocated is variable because of cloudiness, the binarization process needs a dynamic threshold [6]. The analysis calculation is done image by image establishing the threshold as the equation 2 shows.

$$Threshold = H^{-1}\{\min(H(x))\} \quad x_i \leq x \leq x_j \quad (2)$$

Where,

H is the histogram function.

x_i and x_j are the experimental values which delimit the possible value of the threshold.

The second aim is to separate the non-covered sky for the sun and clouds. Pixels belonging to the non-covered sky are characterized by a chromatic (colors other than the neutral colors white, black, and the pure grays) component. On the other hand, the clouds and the sun are represented in the image with achromatic (white, black, and the pure grays) values.

The HSV color space [6] contents a representation which refers to the intensity of a specific color, the saturation. The binarization of a saturation with a suitable threshold makes possible the separation between the grey level and chromatic pixels.

Nevertheless the color degradation around the sun and the earth makes that non-covered sky tends to be achromatic. In addition, the transparency of some clouds reveals the blue hue of the background which is the sky.

In order to resolve this problem, a conservative threshold that classified the most transparent clouds as sky class is established. (figure 2 (b)). In a second step with a second analysis based on texture feature extraction the transparent clouds are separated for the sky (figure 2 (d)).

The texture [15][6] represents the relation of each pixel with its neighbors. If the variance between the pixel analyzed and its neighbors into a defined window is high, the texture will be rough. In other cases the texture value is classified as homogeneous. The thin clouds over the sky generate a more rougher texture than the homogeneity of non-covered sky. This characteristic is used to segment the transparent clouds from the sky.

Measurements of texture computed using only histograms suffer from the limitation that carry no information regarding the relative position of pixels with respect to each other. The second analysis uses the co-occurrence matrix [15][6], which takes into account the positions of pixels with equal or nearly equal intensity values.

In order to characterize the content of co-occurrence matrix the entropy descriptor [16] is used. As the equation 3 shows, this descriptor is a measurement of randomness, reaching its highest value when all elements of co-occurrence matrix are maximally random (figure 2 (c)).

$$Entropy = \sum -P_{i,j} \ln(P_{i,j}) \quad (3)$$

Where,

P is the normalized co-occurrence matrix.

It is assumed that $0 \cdot \ln(0) = 0$.

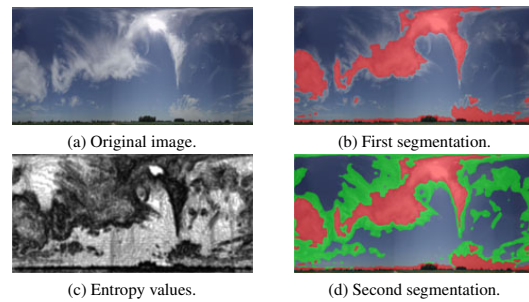


Figure 2: Non-covered sky segmentation.

Finally, the region belonging to the sun is segmented to label this region as sky. Its intensity values are represented with the maximum value of the grey level histogram. However, some clouds near to the sun can contain regions with the same property. To separate it definitively from the clouds the constant area, characteristic circle shape and orientation properties of the sun region are used.

In the case of the earth segmentation, the sky background is subject to a different illumination because of the unpredictable cloudiness. In addition, the interpolations made in different geometric transformations to create the panoramic representation miss the texture features needed to segment the transparent clouds. For these reasons, the segmentation of the earth and the non-covered sky are made image by image before applying the transformations.

On the other hand, the sun segmentation is made in the panoramic image because the system takes into account the characteristic shape and the area of the sun region, so the sun has to be fully stitched in case it has been captured in the union of various images.

4.3 Fog detection

The fog usually appears during the morning reducing visibility to less than 1000 m. It is detected because it covers parts of the earth which are usually visible without fog.

The earth class has a constant shape when it has been captured without fog. The automatic detection is based on a variances in this constant shape. The applied techniques are two: detection based on pattern recognition [6] and detection based on earth silhouette shape analysis.

For the first analysis, it is supposed that any area of the segmented earth can be considered as a constant pattern (in scale and rotation) form which will be easily detected in all of images without fog. The mountain hills are selected as pattern (figure 3(a)) because they are the first areas of the earth part of the images covered by the fog. The statistic value used to quantify the similarity of the specific area of the image under analysis with the chosen pattern is a correlation.

The pattern and the image under analysis may have been captured under different camera exposition mode because of the light presented in the scene. The luminance of each image is referenced to the same white in order to make equal as much as possible the area which will be detected.

The equation 4 shows the mathematical operation of correlation between both images ($f(x,y) \circ g(x,y)$), the pattern (g) and the image which is being analyzed (f), and the decision threshold used to detect the fog. Figures 3(b) and 3(c) support this equation graphically.

$$thr \geq f(x,y) \circ g(x,y) = \sum_{m=1}^{M-1} \sum_{n=1}^{N-1} f^*(m,n)g(x+m,y+n) \quad (4)$$

Where,
 * is the complex conjugate.
 thr is the established decision threshold.

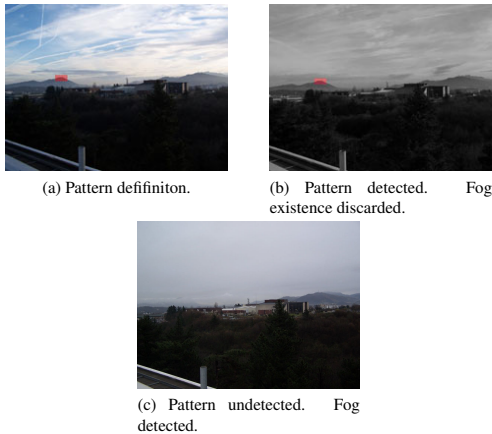


Figure 3: Pattern recognition.

The second analysis takes only into account the shape of the earth silhouette. The earth class segmentation mask is different if the fog is present in the scene. For each image under analysis the system takes the earth segmentation mask and interprets the boundary of the mask [6] as a function. It calculates the relative maximums which are the mountains hills of the image.

For a scene without fog (figure 4(a)) this characteristics point are stored (figure 4(b)) as point coordinates of the image matrix. The extracted points from each analyzed image are compared with the stored characteristic points. The fog presence is detected when there are not characteristic points in the image that is being analyzed.



(a) Original image.



(b) Characteristic points detected over the earth segmentation mask.

Figure 4: Characteristic points detection.

5 Setting up and results

The initial assessment has enabled to specify more in depth which parameters of the camera will make easier the further segmentation process. In addition, the parameters of geometric transformations and segmentation thresholds of all the used techniques have been established.

According to this first assessment, in the images which have been captured underexposing the scene and using light polarizer the segmentation is more accurate because these techniques reduce the sun color degradation effect in the sky and get darker the earth part of the image.

The system has been tested with a database provided by Euskalmet meteorological agency. The images that conform the database have been taken at different time during the daytime and under diverse meteorology conditions.

The database has 10 panoramic images (48 single images/panoramic image) which have been taken during the daytime (from 09:00 to 17:00, 2 images/2 hours) in autumn season. The results which have been contrasted with the profesional staff of Euskalmet meteorological agency, have

been acceptable for the images taken between 11:00 and 15:00. Actually, the precision in cloudiness measurement is higher than the obtained from the meteorologist visual analysis. Nevertheless, at dawn (09:00) and at nightfall (17:00) the obtained result have not been acceptable because the intensity and colorimetric properties of the sky and clouds are different because of the sun illumination, making also more difficult the visual segmentation.

An example of the successful segmentation is shown in figure 5.

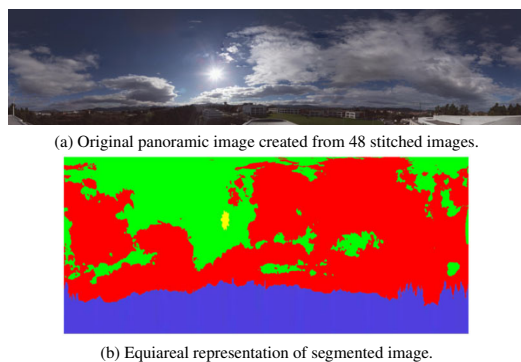


Figure 5: Segmentation result.

Since the available information comes from the visual spectrum, the analysis during the night is discarded.

Fog detection techniques depend on the correct segmentation of the earth class from the image. For this reason, in above described cases where the segmentation works properly, the fog detector has retrieved all the foggy areas, without generating any false positive (figures 3(b)3(c)4(b)).

24 images of 480 images which conform the database, contain fog. Part of these images have been taken in the morning (09:00) and the others during the rest of the daytime period. Independently of the time when the images have been taken, the results are acceptable because when the fog is present in the image the earth class segmentation is not subject to the sun undesirable effects at dawn and at nightfall.

6 Conclusions and Future work

The project concludes with a weather automatic analysis system based on images taken from the earth which carries out the specifications required for the daytime period.

Creating panoramic representations of the environment of each terrestrial meteorological station and being able to centralize them, the system enables meteorologist to cover more area

to make their visual analysis. The new “panorama maker” (adjustable to any camera and lens) has been developed entirely because according to our knowledge, the state of art did not offer any solution to the union of images with homogeneous content such as an image of a clear sky.

Current state of the art offers other solutions for weather analysis using image data coming from satellites. However, the innovative aspect of the project makes possible the consolidation of image processing techniques combinations to analyze and extract useful meteorological information from the images taken from the earth, providing meteorological sciences with an automatic application.

The study of different techniques, and their combination and customization for the specific purpose for meteorological analysis from the images taken from earth, have enabled to contribute to emergent research field using traditional image processing techniques.

Research on High Dynamic Range images will be constitute the future work in order to improve the efficiency of the system during the daytime, making more homogeneous the texture and colorimetric characteristics around the whole panoramic and reducing the undesirable sun degradation effect over the sky.

On the other hand, the analysis has to be expanded to IR in order to extend the functionalities described in this document and be able to get more information from the night period.

Acknowledgements

The authors gratefully acknowledge the collaboration offered by SPRI¹ (Society for Industrial Promotion and Restructuring), Dominion² (Services, Applications and Infrastructure for ITC'S), and Euskalmet³ (Basque Meteorology Agency).

References

- [1] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden. Pyramid methods in image processing. *RCA Engineer*, (29-6), Nov/Dec 1984.
- [2] M.R. Azimi-Sadjadi, M.A. Shaikh, B. Tian, K.E. Eis, and D. Reinke. Neural network-based cloud detection/classification using textural and spectral features. 2:1105 – 1107, 27-31 May 1996.
- [3] M. Brown and D.G. Lowe. Recognizing panoramas. *Proceedings of the 9th International Conference on Computer Vision*, pages 1218–1225, October 2003.

¹<http://www.spri.es>

²<http://www.dominion.es>

³<http://www.euskalmet.euskadi.net>

- [4] M. Brown, R. Szeliski, and S. Winder. Multi-image matching using multi-scale oriented patches. *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, June 2005.
- [5] K. A. Buch, C.H. Sun, and L. R. Thorne. Cloud classification using whole-sky imager data.
- [6] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Prentice Hall Upper Saddle River, New Jersey, second edition.
- [7] J. A. González and D. Pagès. A method for sky-condition classification from ground-based solar radiation measurements. June 20 2001.
- [8] S. Kelby. *The Digital Photography Book*. Peachpit Press, 2006.
- [9] N. Khazenie and K.A. Richardson. Classification of cloud types based on spatial textural measures using noaa-avhrr data. 3:1701 – 1705, 3-6 Jun 1991.
- [10] M. Labayen. Análisis meteorológico basado en imágenes. Master's thesis, Escuela Superior de Ingenieros Industriales y de Telecomunicación, Universidad Pública de Navarra, March 2008.
- [11] J.M. Ogden, E.H. Adelson, J. R. Bergen, and P. J. Burt. Pyramid-based computer graphics. *RCA Engineer*, (30-5), Sept./Oct 1985.
- [12] A. Peters. Die neue kartographie/the new cartography. *Friendship Press*, 1983.
- [13] M. Peura and A. Visa. Land-based cloud classification.
- [14] M. Peura, A. Visa, and I. Kostamo. A new approach to land-based cloud classification. *Proceedings of the 13th International Conference on Pattern Recognition*, 4:143 – 147, 25-29 Aug 1996.
- [15] R. Sapina. Computing textural features based on co-occurrence matrix for infrared images. pages 373 – 376, 2001.
- [16] Claude E. Shannon. *A Mathematical Theory of Communication*, volume 27. July, October 1948.

2.1.4 The COST292 experimental framework for RUSHES task in TRECVID 2008

- **Izenburua:** The COST292 experimental framework for RUSHES task in TRECVID 2008
- **Egileak:** S. U. Naci, Uros Damnjanovic, Boris Mansencal, Jenny Benois-Pineau, Christian Kaes, Marzia Corvaglia, Eliana Rossi, Naiara Aginako
- **Proceedings:** TVS '08 Proceedings of the 2nd ACM TRECVideo Summarization Workshop 2008 (Vancouver, Canada)
- **Argitaletxea:** ACM
- **Orrialdeak:** 40-44
- **Urtea:** 2008
- **DOI:** <http://dx.doi.org/10.1145/1463563.1463569>

The COST292 Experimental Framework for Rushes Summarization Task in TRECVID 2008

S.U. Naci¹, Uros Damjanovic², Boris Mansencal³, Jenny Benois-Pineau³, Christian Kaes³,
Marzia Corvaglia⁴, Eliana Rossi^{3,4}, Naiara Aginako⁵

¹Delft University of Technology, Delft, The Netherlands

²Queen Mary University London, UK

³LaBRI, University of Bordeaux 1/Bordeaux 2/CNRS/ENSEIRB, France

⁴University of Brescia, Italy

⁵VicomTech, San Sebastian, Spain

¹s.u.naci@tudelft.nl

⁴marzia.corvaglia@ing.unibs.it

ABSTRACT

In this paper, the method used for Rushes Summarization task by the COST 292 consortium is reported. The approach proposed this year differs significantly from the one proposed in the previous years because of the introduction of new processing steps, like repetition detection in scenes. The method starts with junk frames removal and follows with clustering and scene detection; then for each scene, repetitions are detected in order to extract once the real scene; the following step consists in face detections (faces are considered semantically relevant) and in pan, tilt and zoom detections (other camera motions are usually related to technical operations in the backstage); finally the summary is extracted.

Categories and Subject Descriptors

H.5.1 [Information Interfaces and Presentation]: Multimedia information systems:video

General Terms

Algorithms, Experimentation

Keywords

repetition detection, spectral clustering, normalized cuts, mid-level features

1. INTRODUCTION

In this paper we present the activities lead with the purpose of participating in the TRECVID 2008 Rushes Summarization task. The COST 292 consortium has participated to this initiative since 2006 with satisfying and progressively

improved results, as well as with a rigorous approach. Concerning the Rushes Summarization task, this year two significant improvements were performed:

- In depth study of development data which allowed the understanding of data structure and thus the consequent understanding of the best strategy for summarization;
- Introduction of repetition detection in scenes in order to remove the redundant parts and to extract only one scene in a set even if that scene was shot several times.

These two innovative aspects generated a completely revised framework based on five main milestones. Given a video, after shot boundary segmentation, first step consists of filtering that removes the junk frames from the video; junk frames are those frames with few colors, frames which are saturated and color bars. Second, clustering is performed using the spectral video clustering algorithm, specifically the *Normalized Cuts* on frames and then the scenes are extracted considering temporally continuous segments of the same clusters which are close enough. Third, within each scene, repetitions in time are detected using the spectral graph theory. Fourth, the mid-level features are extracted: face detection has been implemented in order to extract segments with additional semantic components; camera motion has been computed with the aim of ignoring segments where camera motion do not belong to the final edited video program or movie. Fifth, the summary of 2% of the initial video is extracted on the basis of the extracted features and the constraints given by NIST.

The organization of this paper is as follows: the framework proposed by COST 292 for Rushes Summarization task is presented in Section 2; results and conclusions are respectively reported in Section 3 and in Section 4.

2. COST 292 FRAMEWORK

The framework for Rushes Summarization task proposed by COST 292 is shown in Figure 1. Each box in that diagram is explained in this section.

2.1 Junk frames extraction

According to *instructions for judging video summaries*, to assess how much “junk” a summary contains, a judge has

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

TVS'08, October 31, 2008, Vancouver, British Columbia, Canada.

Copyright 2008 ACM 978-1-60558-309-9/08/10 ...\$5.00.

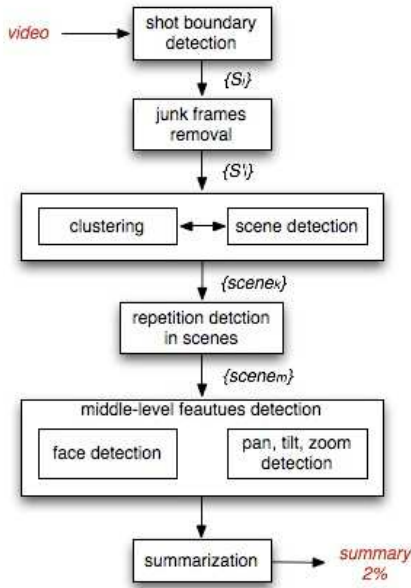


Figure 1: RUSHES framework.

to tell if there are “many color bars, clap boards, all black or all white frames”. Two kinds of color bars are presented on figure 2 a and b. We think that junk frames should also include frames with few colors, such as in figure 2 c, and saturated frames, such as in figure 2 d. All these kinds of frames would not be included in the final edited movie and thus should not be included in the summary.



Figure 2: Junk frames: a) sharp color bars; b) diffuse color bars; c) grey/black frame; d) saturated frame

To detect frames with few colors, thus in particular sharp color bars and uniform color frames, we use a thresholded histogram on each channel of a frame in RGB format. We then check that we have a reduced number of colors. Furthermore, to detect diffuse color bars (such as in Figure 2 b), we apply the same algorithm to picture downscaled to 8x8. These images can be considered as having undergone low-pass filtering.

For performance reasons, we apply this detection at I-frame temporal resolution. We filter the result with a median filter (of width 5). We then interpolate to P-frames resolution. As frames with few colors, and in particular color bars, often last several seconds, this combined filtering and interpolation method seems to work pretty well.

Our method may falsely detect parts of scripted scenes with very few colors, such as very dark scenes. But we observed that events in such scenes are not very understandable and so do not need to be in the summary.

For the detection and extraction of saturated frames in rushes videos Accumulated Histogram Difference (AHD) [8] technique is used. Each frame is converted into HSV color space in order to process each channel independently. A frame is classified as saturated if it has a low value in the S-channel and a high value in the V-channel.

Firstly, a 256 bins histogram is created for both S- and V-channels and the AHD (see Equation 1) is computed in V-channel for the last 20 bins. For S- and V-channels histogram values are summed for the first and last 35 bins, respectively. These three results are normalized and thresholded in order to classify the frame.

$$AHD_n(x) = \sum_{l=x_{min}}^{x_{max}} \Delta H_n(l) = \sum_{l=x_{min}}^{x_{max}} H_{n+1}(l) - \sum_{l=x_{min}}^{x_{max}} H_n(l) \quad (1)$$

$x_{min} \leq x \leq x_{max}$; where H_n is the histogram of frame n and x is the number of bin.

For the frames that come after a frame that is classified as saturated, only histogram values are used. In this case, the number of bins of the histogram is reduced to 25 and the threshold used is more restrictive. Therefore, only saturated frames are considered. As a consequence, meaningless saturated frames are completely extracted from rushes videos.

Our summaries have been judged as not containing too much junk frames: indeed, it seems we are 7th/44 on this criteria.

We still lack a tool to detect clap boards. This could help us to improve our results on junk frames removal. But it could also help us to better separate unscripted and scripted parts of video and thus restrain event detection to scripted parts.

2.2 Clustering and Scene detection

Clustering of frames is used as a first step of the scene detection process. Scenes are treated as segments that hold same semantic information, or which are part of the same semantic story not necessarily visually similar. Segmenting the video into scenes is more natural way of representing videos compared to the simple clustering based summarization. In the scene detection task, spectral clustering approach by normalised cuts [9] is used to cluster the frames. In normalised cuts each frame of a video is treated as a node of the graph, and algebraic graph partitioning techniques are used to find clusters in the dataset. For each frame i we extract MPEG7 Colour layout descriptor [3] and store it in a feature vector f_i of the frame i . Similarity between two data points describes the relation between two frames. If Euclidean distance $\|f_i - f_j\|_2$ between two feature vectors is high, the frames similarity will be close to zero. On the other hand if the distance is small, the similarity will be close to one. The Similarity w_{ij} between two frames i and j with feature vectors f_i and f_j is calculated using a gaussian function:

$$w_{ij} = e^{-\frac{\|f_i - f_j\|_2}{\sigma^2}} \quad (2)$$

Parameter σ serves as a scaling parameter which determine how fast will similarity measure decrease with increas-

ing distance between feature vectors. Similarity matrix $W_{n \times n} = [w_{ij}]$, where n is the number of frames, is created by calculating pairwise similarities between all frames. The matrix W created in such a way hold all information necessary to perform spectral clustering. Clusters are found using eigenvectors and eigenvalues of the similarity matrix. Clustering using eigenvectors have its origin in the problem of graph partitioning, where the goal is to find a cut in the graph that satisfies some predefined criteria. In the literature there exists a number of criteria that can be satisfied using eigenvectors of the similarity matrix [5, 6]. It can be shown that most of this criteria are similar or equivalent to each other. Let V be set of all frames of the original video, and let A and B be two clusters satisfying following conditions: $A \cap B = 0$ and $A \cup B = V$. Cut between two disjoint subsets A and B , $cut(A, B)$ of the set V is defined as $cut(A, B) = \sum_{i \in A, j \in B} w_{ij}$ and association $assoc(A, V)$ of the

subset A is defined as: $assoc(A, V) = \sum_{i \in A, j \in V} w_{ij}$. The clustering criterion $Ncut(A, B)$ that is optimised using eigenvectors of the similarity matrix W is defined as follows:

$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(A, B)}{assoc(B, V)} \quad (3)$$

The clustering problem formulated in terms of cut and association can be seen as identifying groups that have a strong connection between members of the same cluster and weak connections between members of different clusters. Clustering that satisfies these conditions gives minimal value of $Ncut(A, B)$ over a set of all possible clustering results. Searching for the specific clustering that minimise $Ncut(A, B)$ is shown to be NP-hard problem in discrete domain [9]. Algebraic theory of graph spectra properties shows that minimising $Ncut(A, B)$ can be done in continuous domain using eigenvectors and eigenvalues of the similarity matrix W . Letting clustering indicators take continuous values instead of discrete, minimal value of $Ncut(A, B)$ can be found as the second smallest eigenvalue of the similarity matrix W . Every entry $x^{(2)}(i)$ of the second eigenvector $x^{(2)}$ corresponds to one point i in the dataset, and its sign is used as a clustering indicator:

$$\text{if } x^{(2)}(i) > 0 \ i \in A \quad (4)$$

$$\text{if } x^{(2)}(i) < 0 \ i \in B \quad (5)$$

Starting from the set of all frames V , in first step $k = 1$, two clusters are found $V(1)^+$ and $V(1)^-$, $V(1)^+$ correspond to the positive eigenvector entries and $V(1)^-$ to the negative ones. In each of the clustering steps k , $Ncut(k)$ value is calculated using formula (3). If $Ncut(k)$ value is bigger than some predefined threshold $NCUT$, elements of $V(k)^+$ and $V(k)^-$ belong to a single cluster, and are left out from further clustering. Choice of $NCUT$ is done experimentally on a manually annotated training dataset. High $Ncut$ value indicates that the similarity between frames of different clusters is high, so it is most likely that they should stay in the same cluster. On the other hand, lower $Ncut$ values indicate that two clusters are separated more. Clustering process can be generalised as follows, for every step k set of frames that is clustered will be denoted by $V(k)$, cluster corresponding to positive eigenvector entries $V(k)^+$ and cluster corresponding to negative eigenvector entries $V(k)^-$. Clusters $V(k)^+$ and

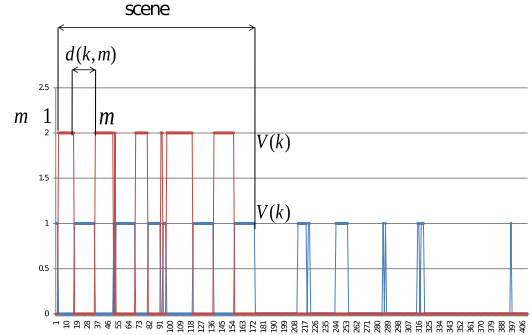


Figure 3: Scene detection example. Cluster indicators values are plotted over the set of frames. If two consecutive continuous segments m and $m - 1$ of the same cluster are close enough they belong to the same scene. Distance between consecutive segments is analysed over time, until all scenes of both positive and negative clusters are found.

$V(k)^-$ will be further clustered in the steps 2^k and $2^k + 1$ respectively if $Ncut(k) < NCUT$. Every clustering step k for which $Ncut(k) < NCUT$ gives useful information for the problem of scene detection. In every step k each frame $i \in V$ can have one of three possible labels $l(i)$:

$$l(i) = 0 \text{ if } i \notin V(k) \quad (6)$$

$$l(i) = 1 \text{ if } i \in V(k)^+ \quad (7)$$

$$l(i) = 2 \text{ if } i \in V(k)^- \quad (8)$$

If we assume that we are in the clustering step k , with parent cluster $V(k)$, and child clusters $V(k)^+$ and $V(k)^-$. When clustering video segments, parent cluster $V(k)$ is not necessarily continuous in time, resulting in $V(k)^+$ and $V(k)^-$ being scattered over the time axis, see 3. Every cluster is then composed of a number of continuous segments. We will denote $m - th$ continuous segment of the cluster $V(k)^+$ $V(k, m)^+$ and $n - th$ continuous segment of the cluster $V(k)^-$ $V(k, n)^-$. Let $d(k, m)^+$ be the distance in time between $m - th$ and $(m - 1) - th$ segment of the positive cluster, and $d(k, n)^-$ be the distance between $n - th$ and $(n - 1) - th$ segment of the negative cluster. Scene boundary detection starts from the first frame of $V(k)$ on the time axis. We define scene as a time segment which contain segments with $d(k, m)^+ < Dseg$ where $Dseg$ is temporal threshold. It means that scene will be formed of temporally continuous segments with distance between every consecutive segment being smaller than $Dseg$. $Dseg$ is not set to fix value, but is dynamically determined on the run. Two consecutive segments $V(k, m)$ and $V(k, m - 1)$ will be put in the same scene if distance between them $d(k, m)$ is smaller than a weighted sum of lengths of the two segments:

$$d(k, m) < T * (length(V(k, m)) + length(V(k, m - 1))) \quad (9)$$

here the weight T is an experimentally determined constant.

2.3 Repetition detection

Repetition detection is based on the spectral graph theory, saying that eigenvector entries that belong to the same

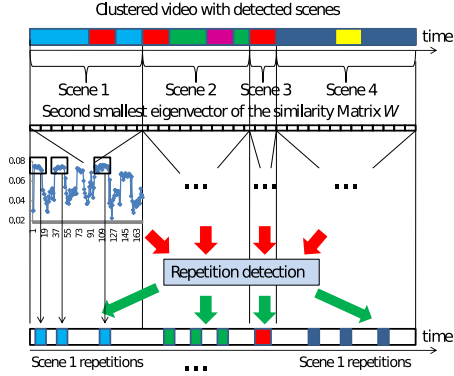


Figure 4: Overview of the repetition detection process. Different colours represent different clusters, with detected scenes. Entries of the second smallest eigenvector of the similarity matrix W , corresponding to the specific scene are used in the repetition detection process. Detecting similar patterns of the eigenvector entries is done for every scene until all repetitions are found.

cluster, in ideal case will be pair wise constant [5]. This practically means that all elements of the same cluster, in ideal case, will have same eigenvector entries which are different from entries belonging to other clusters in the dataset. It is also worth noting that stronger the similarity between two clusters is, stronger the similarity between eigenvector entries corresponding to these clusters will be. On the other hand difference between entries of dissimilar clusters will be high. We use these facts to analyse scenes by analysing the structure of its eigenvector entries. Let S be the scene with $n(S)$ frames, in which we are looking for the repetitions, with $i_{1(S)}$ being the first frame of the scene, and $i_{n(S)}$ last frame. We use second smallest eigenvector of the first clustering step, which is done on the whole video. The second smallest eigenvector of the first step is $x_i^{(2)}(1)$, $i \in 1..n$ where n is a number of frames in the video. For each scene S analysis is done on a subset of $x_i^{(2)}(1)$ corresponding to the scene S :

$$i_{1(S)} \geq t \geq i_{n(S)} \quad (10)$$

Repetitions are found by searching for segments in the $x_i^{(2)}(1)$ that have similar value of eigenvector entries, that are separated in time, and have similar temporal structure. Condition that repetitive segments have to have similar eigenvector entries comes from the fact that these segments have similar if not the same visual layout. Separation is logical consequence of the repetition definition, as repetitive taking of the same scene mixed with recordings of technical preparations is found to be main feature of the BBC Rushes videos. Since interesting segments are taken according to a script which is usually fixed, similar duration of the repetitive segments showed to be useful condition for detecting these scenes. It means that the repetitive segments will have similar distribution of eigenvector entries in time. Assuming we are in the scene S , we analyse second smallest eigenvector

entries corresponding to the frames belonging to the scene S , see (4). Let $x_{max}^{(2)}(S)$ be the maximal eigenvector value within the scene S and $x_{min}^{(2)}(S)$ be the minimal value. We define span of the scene S as a segment of eigenvector values between minimal and maximal value:

$$span(S) = x_{max}^{(2)}(S) - x_{min}^{(2)}(S) \quad (11)$$

We divide the span of the eigenvector entries $span(S)$ into r partitions P_i , $i = 1..r$, of length $span(S)/r$. Every partition P_i will have some number of points $N(P_i)$, whose eigenvector entries belong to P_i . By defining partitions P_i we captured information about eigenvector entries distribution over the $span(S)$. Looking into distribution of $N(P_i)$ over the set of partitions P_i we can detect such partitions P_i^m that have maximal values of $N(P_i)$ on the local level. These partitions correspond to peaks in the histogram of $N(P_i)$ over P_i and satisfy following conditions:

$$P_i^m > P_{i-1} \text{ and } P_i^m > P_{i+1} \quad (12)$$

Now, we treat each local peak as a centre of a possible repetition e in the space of eigenvector values. Possible repetition e is defined as a set of frames whose eigenvector entries fall into area around the local peaks P_e . P_i^m denote detected local peak in the histogram of $N(P_i)$, and $N(P_i^m)$ is a number of frames whose eigenvector entries belong to the partition P_i^m . Area $span(P_i^m)$ around the local peak P_i^m which defines the possible repetition e is defined as a set of partitions with number of frames being bigger then half the number of the frames $N(P_i^m)$ in the partition P_i^m . By the nature of repetitions, it is assumed that they are all of similar duration. Every possible repetition e_i of the same event E which have significantly different duration $dur(e_i)$ compared to other possible repetition of the same event is discarded. First we calculate mean duration of the repetition segments $dur(R_i \in E)_{mean}$ of the event E , and then we test if the current segment satisfies following conditions:

$$R_{max} * dur(R_i \in E)_{mean} \geq dur(R_i) \quad (13)$$

$$dur(R_i) \geq R_{min} * dur(R_i \in E)_{mean} \quad (14)$$

R_{max} and R_{min} are thresholds used to keep duration values have small standard deviations. This is done for all scenes and all events within the scene until all scenes are analysed and all repetitions are found.

2.4 Detection of Mid-level features

In order to create a summary, due to the nature of the rushes content, the detection of events constitutes an interesting tool. Our approach is based on the usage of some meaningful mid-level features that, according to our experience, are relevant for event detection. Specifically we have implemented: face detection and camera motion description which is a significant mid-level feature because directors usually use this characteristic to highlight some relevant events in the film and explicitly camera events make a part of semantic Ground Truth annotation. The developed algorithm tags frames for the different camera motion types.

We believe a human is one of the most recognizable object in a video, especially for a human summarizer. Moreover, most of the events in ground truth produced for evaluation contain a reference to a human posture or action. Our approach is a combination of two detectors: one of Viola and Jones, extended by Lienhart, from OpenCV [1], using Haar-like features, thus working on textural features, the other

one uses skin color appearance model trained on the faces detected by OpenCV. The first implementation of this approach was described in [2]. Compared to pure OpenCV, this method allows to increase the recall, without degrading precision.

For significant camera motion detection, we use the algorithm described in Kraemer et al [4]. First, we estimate the global camera motion, extracting only motion vectors from P-frames of MPEG compressed stream. Then we use a likelihood significance test of the camera parameters to classify specific camera motions. The algorithm of [4] allows for classification of camera motion as pure physical motions, such as “pan/travelling”, “tilt”, “zoom”, “rotation” or complex motions.

In rushes content, during scene setup, there are often short, noisy camera motions. However, such unwanted motions are often complex and thus can be discarded by algorithm [4].

2.5 Merging and summary production

The merging system is developed to utilize the units mentioned above to produce summaries of the rushes videos which are *plausible to watch, informative, clear from redundancy and with the maximum coverage of the events*. The system functions in four steps to produce the final summary whose length is limited to maximum 2% of the input video:

1. Detection of redundant parts in the video.
2. Detection of separate scenes and repetitions in the scenes.
3. Aligning the repetitions and calculation of importance within the repetitions.
4. Creating the summary selecting relevant parts from each repetition.

Proposed system puts emphasis on the “watchability” issue. In that sense the system attempts to produce outputs that follow the story line of the content, that contain the important events and actions in the video and also, in the automatic editing process, it takes the measures to prevent any difficulty or annoyance for the viewer. Firstly, the system never displays segments shorter than 2 seconds. In addition to this, we do not use techniques like fast forwarding and frame-in-frame which limit the access of viewer to audio modality and boost artificially the content of events.

3. EXPERIMENTAL RESULTS

As can be seen in the detailed results in [7], the proposed method has performed quite well in many criteria. Especially taking into account the criteria *repeated segments* (3rd best result among all), *pleasant tempo and rhythm* (2nd best result among all) and *junk frame removal* (7th best result among all), the system can be referred to as capable of producing pleasant and informative summaries of the videos which are cleared from the redundancy. On the other hand, in the results we observe that the *inclusion of the events* (36th best result among all) is below the average and we investigated the reasons behind this. Firstly it is important to consider this important measure together with the other performance measures. Otherwise we observe that the best performing system in this ranking is the CMU’s base

system which mainly plays the videos in a fast-forward manner. This system, which includes almost all the events (as expected) shows poor performance in other criteria, i.e. it produces a “summary” which shows a piece of every event existing in the video in an unpleasant and difficult to understand manner. In general when we look at all the scores in different criteria and from different groups, we observe an inverse relation between the *inclusion of the events* and other three criteria mentioned above.

4. CONCLUSIONS

In this paper we introduced a system for creating summaries from unedited videos. The system is designed to create an output that is free from redundancy and repetition, and which highlights the most important events in the video, and does all of this in a generic manner. For the system the “watchability” was another issue that we put emphasis on: the created summaries should be composed of continuous segments which are long enough to let the users understand the content in audio and visual modalities. This suggests that the possible techniques that might be employed for using the 2% time limit more efficiently (i.e. fast forwarding, frame-in-frame or multiple very short segments) are avoided in the summaries. In such a system where the number of video segments that can be added to the summary is limited, the success in choosing the right segments becomes critical. Our efforts in this direction has resulted in a better inclusion rate compared to the last years system, although the summary length has been halved. We plan to continue working on this issue by using the mid-level features in a more efficient way and proposing improvements on the scene clustering algorithms.

5. REFERENCES

- [1] Opencv. <http://opencvlibrary.sourceforge.net>, 2007.
- [2] A. Don, L. Carminati, and J. Benois-Pineau. Detection of visual dialog scenes in video content based on structural and semantic features. In *Proc. CBMI’05*, Létonie, 2005.
- [3] E. Kasutani and A. Yamada. The mpeg-7 color layout descriptor: a compact image feature description of high-speed image/video segment retrieval. In *ICIP 2001*, Greece, 2001.
- [4] P. Kraemer, J. Benois-Pineau, and M. Gràcia Pla. Indexing camera motion integrating knowledge of quality of the encoded video. In *Proc. SAMT’06*, 2006.
- [5] M. Meila and J. Shi. Learning segmentation by random walks. In *NIPS*, 2000.
- [6] M. Meila and J. Shi. A random walks view of spectral segmentation, 2001.
- [7] P. Over, A. F. Smeaton, and G. Awad. The TRECVID 2008 BBC rushes summarization evaluation. In *TVS ’08: Proceedings of the International Workshop on TRECVID Video Summarization*, pages 1–20, New York, NY, USA, 2008. ACM.
- [8] X. Qian, G. Liu, and R. Su. Effective fades and flashlight detection based on accumulating histogram difference. *IEEE Transactions On Circuits And Systems For Video Technology*, 16(10), 2001.
- [9] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 2000.

2.2 Irudien ulerpenerako ikasketa automatikoko metodoen erabilera

Ikasketa automatikoko sailkatzaileak erabiliz ikusmenaren barruan dauden arazo askoren irtenbidea aurkitu daiteke. Sailkatzaile hauek metodo numerikoek ebatzi ezin ditzaketen problemak ebazteko gaitasuna daukate. Horregatik, irudien prozesaketa eta ulerpenaren ikerketa munduan, sailkatzaileen erabilpena oso jorratuta dagoen gaia da. Bien arteko elkarrekintza izugarri hobetzen ditu lortutako garapenen emaitzak. Are gehiago, *ikasketa* faktoreak domeinu dependentzia murrizten laguntzeko estrategiak eskaintzen ditu.

Ikerketa hau ez da sailkatzaile horien analisisira bideratu, hau da, sailkatzaileak ez dira izan ikerketaren helburua baizik eta irtenbidea aurkitzeko giltza. Batez ere, gain-begiratutako ikasketa metodoen barruan dauden sailkatzaileak erabili dira domeinu itxi bateko aplikazioetan erabili direlako. Beraz, sailkatzaileen entrenamendurako aurretik etiketatutako datu multzoak erabili dira.

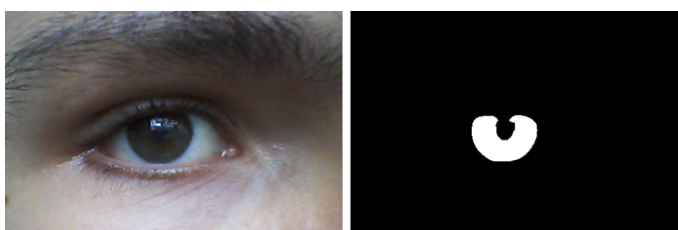
Ikerketa lerro honen ekarpen nagusiak bi multzo hauetan banatzen dira:

1. **Irudien sailkapenerako desberdintasun faktore berri** baten definizioa. Faktore hau bi irudien artean dagoen desberdintasun maila kalkulatzeko balio du eta beraz, irudiak klase desberdinetan sailkatzeko erabilgarria da. Desberdintasun faktore hau sailkatzaileek emaitza bezala ematen duten probabilitate banaketan distantzia bezala kalkulatzen da. Distantzia horri atalase maila bat esleitzen zaio eta honen arabera irudien sailkapena egiten da (ikusi 2.2.1 argitalpena).
2. **Irudien ezagutzarako sailkatzaileetan oinarritutako metodologiaren** diseinu eta garapena. Metodologia hau irudien behe-mailako ezaugarrietan eta sailkatzaileetan oinarritzen da. Metodologia hau garatzeko momentuan bi helburu izan dira nagusi: batetik, irudiaren behe-mailako ezaugarrien aukeraketa prozesuaren egokitasuna aztertzea eta bestetik, ikasketa automatikoak irudien ezagutzan duen potentziala aztertzea. Horretaz gain, prozesuaren etapa bakoitzean lortzen den informazioaren fusioa egiteko modu posibleak eta honek emaitzetan duen eragina ikertu egin da (ikusi 2.2.2 eta 2.2.3 argitalpenak).

Bi alor desberdinetako irudiak erabili dira aipatu diren ikerketa hauek aurrera eramateko. Batetik, errekonozimendu biometrikoaren barruan kokatzen den irisaren ezagutzarako irudiak eta bestetik landareen irudiak. **Irudien sailkapenerako desberdintasun faktorea**, [Agi+17b] **Iris matching by means of machine learning**

paradigms: a new approach to dissimilarity computation argitalpenean aurkeztu dena, gailu mugikorrek ateratako bi iris irudien arteko desberdintasun maila kalkulatzeko erabili da. Horrela, bi iris irudi pertsona berdinari dagozkion edo ez zehaztu daiteke eta irudien sailkapena egin. ICPR 2016 [ICP16] konferentzian aurkeztu den Mobile Iris CHallenge Evaluation II (MICHE II) [MIC] lehiaketarako antolatzaileek prestatutako eta etiketatutako iris irudiak erabili dira sistema entrenatu eta testatzeko. Irudi prozesamenduari dagokionez, irudi transformazio sinpleak erabili dira sailkatzailetara sartu diren atributuak ateratzeko. Nahiz eta teknika oso sinpleak erabili, lortutako emaitzek oso doitasun maila altua izan dute eta beraz faktore hau irisen irudien sailkapenerako oso egokia dela frogatzen da. Kasu honetan, garatutako metodoaren muina ez da irudi prozesamenduan aurkitzen baizik eta sailkatzaileen erabilpenean eta desberdintasun faktore berrian.

Hurrengo bi argitalpenetan berriz, irudi prozesamenduaren garrantzia handiagoa da. Bi aplikazioen kasuan, hau da, iris irudien eta landare irudien sailkapenen kasuan, arazora moldatutako behe-mailako deskriptoreak ateratzen dira (ikusi 2.2 irudia). [Agi+17c] **Periocular and iris local descriptors for identity verification in mobile applications** argitalpenean, irisaren eta irisaren inguruko zatiaren ezaugarri lokalak ateratzen dira iris irudia zein pertsonari dagokion zehaztu ahal izateko. Pertsonen ezagutzarako sistemetan iris inguruko informazioa gehitzeak, lortutako emaitzak hobetzen dituela baieztatzen du hemen aurkeztutako ikerketak.



(a) Irisaren detekzioa



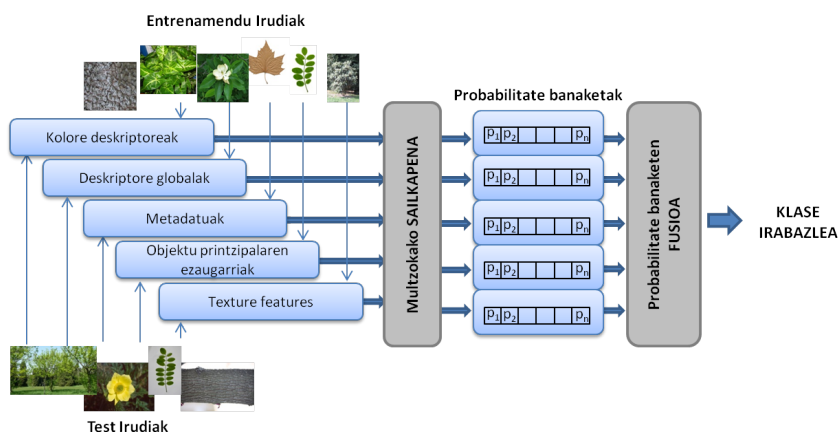
(b) Irisaren banaketa ezaugarri lokalak ateratzeko

Irudia 2.2.: Iris irudien behe-mailako deskriptoreak ateratzeko prozesua

[Agi+14] **Identification of plant species on large botanical image datasets** argitalpenean, landareen irudiak sailkatzeko metodologia bat proposatzen da. Horretarako LifeCLEF 2014 [GJB] lehiaketak parte-hartzaileen eskura landareen 47815 irudi uzten ditu. Irudi hauek 6 taldetan banatzen dira kontutan izanik landarearen zein parteri dagozkion: hostoa, eskaneatutako hostoa, lorea, enborra, fruitua eta

landarea bere osotasunean. Irudi hauen sailkapena egiteko metodologia irudien ezaugarriak multzoka aztertu eta sailkapenaren irteeran fusiorako algoritmoak aplikatzean datza. Multzo bakoitzeko atributuak sailkatzaileera sartu eta honek irteera bezala ematen duen probabilitate banaketan fusioa eginez, irudi bakoitzaren klasea zein den zehazten da.

Orokorrean, irudi batetik ateratako ezaugarri guztiak sailkatzaileetara sartu baino lehen batu egiten dira eta ondorioz sailkatzaileek irteera bakar bat ematen dute. Kasu honetan, ezaugarri guztiak batu beharrean, haien jatorriaren arabera multzokatu dira (ikusi 2.3 irudia). Multzoketa hori jarraituz sartu dira sailkatzaileetara, multzo bakoitzerako entrenamenduan emaitza onenak eman dituen sailkatzailea erabiliz testerako. Emaitzen fusioa sailkatzaileen irteera den probabilitate banaketa horien gainean egin da. Bi hurbilketen emaitzak alderatu dira eta aurkeztutako metodologia honek hartutako estrategia egokia dela frogatu da. Gainera, sistemak askoz azkarrago lan egiten du sailkatzaileetara sartzen den atributu kopurua murriztu egiten delako.



Irudia 2.3.: Landareen sailkapenerako garatutako fusio metodologia

2.2.1 Iris Matching by means of Machine Learning Paradigms: a new Approach to Dissimilarity Computation

- **Izenburua:** Iris Matching by means of Machine Learning Paradigms: a new Approach to Dissimilarity Computation
- **Egileak:** Naiara Aginako, J.M. Martinez-Otzerta, Igor Rodriguez, Elena Lazkano, Basilio Sierra
- **Aldizkaria:** Pattern Recognition Letters (PRL)
- **Argitaletxea:** Elsevier
- **Inpaktu-faktorea (urtea):** 1,586 (2015)
- **Kuartila:** Q2
- **Zenbakia (Orrialdeak):** -
- **Urtea:** 2017
- **DOI:** <http://dx.doi.org/10.1016/j.patrec.2017.01.019>



Iris Matching by means of Machine Learning paradigms: a new Approach to Dissimilarity Computation

Naiara Aginako^{a,*,*}, Goretti Echegaray^a, J.M Martínez-Otzeta^b, Igor Rodríguez^b, Elena Lazkano^b, Basilio Sierra^b

^aApplied Mathematics Department (UPV-EHU), Robotics and Autonomous Systems Research Group, Donostia 20018, Spain

^bComputer Sciences and Artificial Intelligence Department (UPV-EHU), Robotics and Autonomous Systems Research Group, Donostia 20018, Spain
<http://www.sc.ehu.es/ccwrobot>

ABSTRACT

This paper presents a novel approach for iris dissimilarity computation based on Computer Vision and Machine Learning. First, iris images are processed using well-known image processing algorithms. Pixels of the output image are considered the input of the previously trained classifiers, obtaining the *a posteriori* probability for each of the considered class values. The main novelty of the presented work remains in the computation of the dissimilarity value of two iris images as the distance between the aforementioned *a posteriori* probabilities. Experimental results, based on the testing dataset given by the MICHE II Challenge organizers, indicate the appropriateness of the deployed method for the iris recognition task. Best results show a precision score above 90% even for iris images of new individuals.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Biometric services are increasingly addressing the need for reliable authentication methods. They are required in everyday tasks such as micro-transactions, insurance entitlements and payments. Furthermore, the availability of these services in mobile platforms, which are ubiquitous and omnipresent, exalts the benefits of these technological solutions.

More precisely, iris recognition methods are essential for these applications due to the enormous number of discriminative features of the iris. As stated in (Wildes, 1997), iris recognition can provide the basis for truly non-invasive biometric solutions as it is highly distinctive and stable regardless of the age of the user.

Even though, discriminative features are usually affected by the acquisition process of iris images and consequently adopted solutions are not always as precise as required. Therefore, user recognition processes require new technological approaches to overcome these limitations.

Aligned with the exposed needs, this paper presents a novel approach that combines image processing and classification to

increase the accuracy of the recognition of iris images. The main contribution relies on the computation of the dissimilarity value that determines if a pair of iris images belongs to the same person or not. To sum up, the main concepts of the proposed approach are the following:

- Machine Learning and Computer Vision are combined for iris image classification: first, images are processed by means of diverse image processing algorithms in order to highlight the most discriminative features of the irises. Pixels of the processed image are directly considered the values of the feature vector and thus, the input of the classifiers.
- Computation of the dissimilarity value of two iris images as the distance between the two *a posteriori* probability vectors that are the output of the classification process.

MICHE dataset (De Marsico et al., 2015) is used to evaluate the performance of the proposed method. It contains iris images captured using three different mobile devices, two of which are mobile phones. Besides, this dataset has been proposed by the Mobile Iris Challenge Evaluation (MICHE-II), now part of the 23rd International Conference on Pattern Recognition (ICPR). This challenge is a new edition of the MICHE-I contest and its

*Corresponding author:

e-mail: naiara.aginako@ehu.eus (Naiara Aginako)

main objective is to collect novel and suitable techniques for iris recognition in images captured by mobile devices.

The rest of the paper is organized as follows: Section 2 presents the related work in the area of mobile biometric recognition and more precisely iris recognition using mobile devices. Section 3 describes the proposed approach and Section 4 summarizes the experimental setup. Section 5 gives the experimental results in order to show the suitability of our method for the presented challenge. Finally, in Section 6 conclusions of the work and the future work are presented.

2. Related Work

Biometric recognition is the automated identification of individuals based on their behavioral and biological characteristics. As stated in (Prabhakar et al., 2003) a biometric system is essentially a pattern-recognition system that recognizes an individual based on a feature vector derived from specific biological or behavioral characteristics.

There are several biometric recognition methods such as the ones presented in (Delac and Grgic, 2004); some of them are already implemented in real systems, while others are still under research.

Reliable automatic person identification has long been an attractive goal. For this purpose, as stated in (Daugman, 2002), (De Marsico et al., 2012) (Daugman, 2003), iris recognition is one of the major biometric recognition methods due to the reliability of iris patterns used in visual recognition of persons. The iris has the great advantage that its pattern variability among different persons is enormous and even more, it is stable over time.

Iris recognition systems consist mainly of four different steps (Sheela and Vijaya, 2010): image acquisition, image pre-processing, feature extraction and pattern classification. The phase-based method presented by (Daugman, 1993) was one of the first commercial solutions. Later, some other authors such as (Martin-Roche et al., 2001) and (Masek, 2003) developed similar ideas.

The implementation of new modalities based on thermal imaging are considered possible future solutions for iris recognition systems (Ramaiah and Kumar, 2016). More concretely, the fusion of the information obtained from images acquired using different types of sensors is being considered as one of the most challenging research directions for pattern recognition systems. For example, in (Susperregi et al., 2013a) and (Susperregi et al., 2013b), images from different sensors are combined to increase the accuracy of person's detection.

Another open issue in biometric recognition is the need of evolving the image acquisition step. Biometric identification devices must be designed for more intuitive, faster and lighter operations, and therefore, research is moving forward to the use on images captured from mobile devices. A complete study of these solutions is presented in (Kim et al., 2016).

3. Proposed Approach

As previously mentioned, the main objective of the presented approach is to determine the dissimilarity between two iris im-

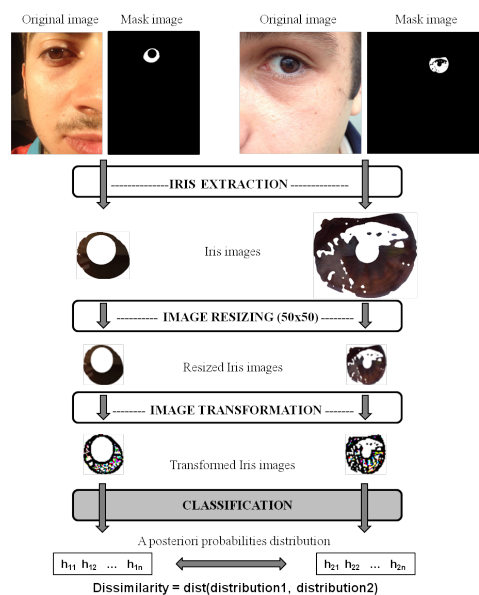


Fig. 1. Steps of the comparison of two iris images

ages in order to consider if they belong to the same person or not. To do that, this work combines the extraction of image features through the application of diverse Computer Vision algorithms with the use of Machine Learning techniques. There are two main phases: training phase where classifiers adapted to the concerning problem are built testing phase where two iris images are compared computing their dissimilarity value using the output of the classifiers.

During the training phase, classifiers (see subsection 3.2 for more detailed information) are built using MICHE II dataset training images. First, images are processed using different image transformations and the pixels of the processed image are directly considered the input vector of the classifiers. Each image has associated a class that identifies the individual it belongs to. Taking into account this class, classifiers learn about the distinctive features of individual's iris images. Training process is totally blind for the user.

During the testing phase, as well as in the previous training phase, iris images are processed by pairs using different image processing algorithms to extract characteristics that are relevant for the identification of iris patterns. Afterwards, each pixel vector is classified to obtain the *a posteriori* probability vector used for the computation of the dissimilarity value. If this distance value stands under a predefined threshold, two images are considered of the same person.

In more detail, the main steps of the testing process can be summarized as follows (see Figure 1):

- Starting from a pair of given images and their masks, extract the images corresponding to the iris.
- In order to obtain the same number of attributes to be inserted in the classifiers, resize iris images to a fixed size

(for instance, 50x50 pixel size).

- Apply the image transformation used in the learning phase of the classifiers to both images.
- Classify both images using the induced classifier, obtaining not only the winning class, but also the *a posteriori* probability distribution of all considered classes.
- Compute the dissimilarity between both iris images as the Manhattan distance of the two probability distributions.
- Determine if two iris images belong to the same person using a threshold value.

In the following sections a short revision of the image processing process, applied classifiers and dissimilarity computation is presented.

Table 1. Used image transformations

Transform	Command	Effect
Transf. 1	Convolve	Apply a convolution kernel to the image
Transf. 2	Despeckle	Reduce speckle within an image
Transf. 3	Edge	Detect edges in the image
Transf. 4	Enhance	Digital filter to enhance the image
Transf. 5	Equalize	Perform histogram equalization
Transf. 6	Gamma	Perform a gamma correction
Transf. 7	Gaussian Blr.	Reduce image noise and detail levels
Transf. 8	Lat	Local adaptive thresholding
Transf. 9	Linear Str.	Linear with saturation histogram stretch
Transf. 10	Median	Apply a median filter to the image
Transf. 11	Modulate	Vary the brightness, saturation, and hue
Transf. 12	Negate	Replace each pixel with its complementary
Transf. 13	Radial-blur	Radial blur the image
Transf. 14	Raise	Lighten/darken edges to create 3-D effect
Transf. 15	Selective-blur	Selectively blur pixels within threshold
Transf. 16	Shade	Shade the image using a distant light
Transf. 17	Sharpen	Sharpen the image
Transf. 18	Trim	Trim image edges
Transf. 19	Unsharp	Unsharpen the image

3.1. Image Processing

A bank of filters with large variability is applied to highlight different characteristics of iris images (see some examples in Figure 2). To achieve this, we have selected some of the most common transformations in order to show the benefits of the

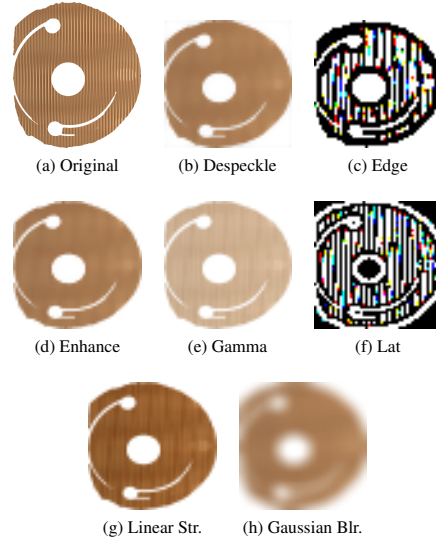


Fig. 2. Some of the transformations

proposed approach making use of simple algorithms. Table 1 presents the transformations used as well as a brief description of each of them. It is worth to point out that any other Image Processing transformation could be used apart from the selected ones.

3.2. Machine Learning classifiers

During the training phase, the building of a classifier implies a learning process from a hand-labeled dataset of images. As classifiers we use four well known ML supervised classification algorithms (Mitchell, 1997): IB1, Naive-Bayes, C4.5 and Random Forest. They have completely different approaches to learning and a long tradition in different classification tasks.

In the following paragraphs the main characteristics of each of the classifiers are summarized:

IB1. The IB1 described in (Aha et al., 1991) is a case-based, Nearest-Neighbor classifier. To classify a new test sample, all training instances are stored and the nearest training instance with respect to the test instance is found: its class is retrieved to predict this as the class of the test instance.

Naive-Bayes. The Naive-Bayes (NB) rule presented by (Cestnik, 1990) uses the Bayes theorem to predict the class for each case, assuming that the predictive features are independent given the category. To classify a new sample characterized by d features $\mathbf{X} = (X_1, X_2, \dots, X_d)$, the NB classifier applies the following rule:

$$c_{N-B} = \arg \max_{c_j \in C} p(c_j) \prod_{i=1}^d p(x_i | c_j)$$

where c_{N-B} denotes the class label predicted by the Naive-Bayes classifier and the possible classes of the problem are grouped in $C = \{c_1, \dots, c_l\}$.

C4.5. As defined in (Quinlan, 1993), C4.5 represents a classification model by a decision tree. It is run with the default values of its parameters. The tree is constructed in a top-down way, dividing the training set and beginning with the selection of the best variable in the root of the tree. The process stops at each node of the tree when all cases in that point belong to the same category or the best split of the node does not surpass a fixed chi-square significant threshold. Then, the tree is simplified by a pruning mechanism to avoid overspecialization.

Random Forest (RandomF). This classifier constructs a combination of many unpruned decision trees (Breiman, 2001). The output class is the mode of the classes of the individual trees.

3.3. Dissimilarity computation

In general, classification approaches consider the resulting winning class as the solution of the classification process. Therefore, for the concerning task, if two iris images differ in their resulting winning class they will be considered as not belonging to the same person. But, taking into account that the differences between iris images captured with mobile devices are very subtle, there is a need of a more robust solution.

Therefore, in the presented work dissimilarity value of two iris images is computed as the distance between the *a posteriori* probability vectors. This vector represents the probability of belonging to a certain class. Consequently, the length of this vector is the number of considered classes in the training phase. The distance between these two probability distributions, which can be also seen as histograms, is the measure of the dissimilarity given by our approach. In this case, Manhattan distance has been used. The resulting value is finally compared with a threshold value to determine the belonging or not to the same individual. Hence, two iris images with different output winning classes can still be considered of the same person.

4. Experimental Setup

As previously mentioned, the proposed approach has been designed and implemented for the recognition of iris images captured from mobile devices. Therefore, in order to prove its suitability for resolving this task, MICHE-II proposes two separated datasets:

- **Training dataset:** This contains a total of 3148 iris images of 75 individuals, that is a total of 75 classes. It has been used for the learning process of the classifiers and the identification of the best image transformation and classifier pair.
- **Test dataset:** This dataset contains two different folders: Probe and Gallery, each of them has a total of 60 iris images. Some of these images belong to individuals that are not present in the training dataset. It has been used to evaluate the performance of the presented approach.

First of all, all iris images are extracted from original images processing them with their corresponding masks that have been computed automatically by the executable available for

MICHE II participants. Then, after resizing all the images to a fixed size of 50x50 pixels, the 19 transformations described in Table 1 are applied. Our aim is to measure the classification accuracy of the four classifiers previously described (IB1, Naive-Bayes, C4.5 and Random Forest) over the 19 primitive features. During the learning process of the classifiers, predictor variables are the pixels of the processed image and the input class is the individual's ID code. This code identifies to whom belongs each iris image.

Weka machine learning utility (Hall et al., 2009) has been used to perform the experiments. More in detail, an ARFF file (the file format used by Weka) has been built from the images under each transformation. This ARFF file defines 50x50=2500 features, which are the pixels of the processed image, and the class for each image.

5. Experimental Results

In order to identify the best image processing and classifier couple, a total of 80 experiments have been achieved (Original images + 19 transformations evaluated using four different classifiers) using the training dataset. Performances are measured by the classifier accuracy after a 10-fold cross-validation. The results of these experiments are shown in Table 2. As presented in this table, the best results are obtained by the combination of IB1 classifier and the Edge primitive feature.

Table 2. 10 fold cross-validation accuracy percentage obtained for each classifier using each of the proposed transformations.

Images	RandomF	NB	C4.5	IB1
Original	6.74	25.54	22.74	55.03
Transf. 1	42.05	25.54	22.55	55.03
Transf. 2	44.34	15.59	23.63	54.96
Transf. 3	13.61	25.70	10.02	59.06
Transf. 4	44.27	11.23	21.00	54.87
Transf. 5	33.78	15.27	19.08	57.86
Transf. 6	42.33	12.37	22.11	57.44
Transf. 7	43.13	11.80	23.51	57.51
Transf. 8	6.58	25.54	7.86	50.92
Transf. 9	42.08	11.80	20.61	55.53
Transf. 10	42.97	11.26	21.00	55.53
Transf. 11	37.56	25.54	19.75	55.41
Transf. 12	6.74	25.54	22.74	55.03
Transf. 13	44.91	21.31	23.38	56.55
Transf. 14	42.08	25.54	22.39	55.03
Transf. 15	42.05	25.54	22.55	55.03
Transf. 16	21.37	18.51	12.88	41.00
Transf. 17	35.18	25.54	17.08	52.48
Transf. 18	42.05	25.54	22.55	55.03
Transf. 19	35.72	11.58	17.49	52.10

For the evaluation process, Gallery and Probe datasets are used. For that end, each Gallery set image is compared with each Probe set one. In this way, a total of 3600 comparisons are made during the validation process. It is worth mentioning that 38 among the 60 images of both folders correspond to

Table 3. Results of the evaluation process

Model	Classifier	Set	Errors	Comparisons	Accuracy
Model1	IB1	All v All	67	3600	98.14
Model1	IB1	Known v Known	1	1144	99.93
Model1	IB1	Unknown v All + All v Unknown	66	2156	96.69
Model1	IB1	Unknown v Unknown	48	484	90.08
Model2	IB1	All v All	80	3600	97.78
Model2	IB1	Known v Known	3	1444	99.79
Model2	IB1	Unknown v All + All v Unknown	77	2156	96.43
Model2	IB1	Unknown v Unknown	54	484	88.84
Model3	RandomForest	All v All	551	3600	84.69
Model3	RandomForest	Unknown v All	53	1444	96.33
Model3	RandomForest	Unknown v All + All v Unknown	498	2156	76.90
Model3	RandomForest	Unknown v Unknown	48	484	90.08

Known people, individuals that are already present in the training dataset. The remaining 22 images belong to unidentified users, that are named *Unknown* in our experiments. In order to better analyze the obtained results we show in Table 3 four different aspects of the Gallery-Probe experiment:

- **All v All:** All 60 images of Probe folder against Gallery folder, in total 3600 comparisons.
- **Known v Known:** Comparisons between iris images that belong to *Known* individuals, people whose images are in the training dataset. In total $38 \times 38 = 1444$ comparisons.
- **Unknown v All + All v Unknown:** Comparisons between all the images of Probe folder against *Unknown* individuals of the Gallery and vice versa.
- **Unknown v Unknown:** Comparisons between all the *Unknown* individuals' images of both folders. In total $22 \times 22 = 484$ comparisons.

The evaluation has been performed for the three models which were considered the best ones in the ICPR contest (Aginako et al., 2016); the models were obtained by using IB1 (Model1, Model2) and Random Forest (Model3) as base classifier. It is worth mentioning that classifier families different from KNN obtain a significantly worse result.

As it can be seen in Table 3, obtained accuracy results are over 90% for all the considered subsets in the best model. Moreover, the results using only *Known* individuals' images obtain a precision of 99.93% and only a single error among the 1444 comparisons performed. This indicates the adequateness of the proposed approach for the recognition of the iris images of individuals already present in the training dataset. Furthermore, the results using only *Unknown* individuals are also remarkable because the precision is between 90.08% of well classified cases in the best case and a 88.84% in the worst one. Consequently, this approach can be considered as a promising solution when dealing with individuals that are not previously known for the system.

6. Conclusions

This work presents a novel iris recognition method applicable to mobile captured images. It combines techniques from Computer Vision and Machine Learning. Several experiments using a combination of different image transformations and Machine Learning algorithms have been accomplished to select the best solution. As a conclusion, mention that best results are obtained using Edge transformation followed by IB1 classification method.

The validation process implies a comparison between a pair of iris images to identify their belonging to the same individual. To that end, three best image processing and classifiers combinations have been tested. Obtained results highlight the adequateness of this method both for iris images of individuals already present in the dataset and new individuals' iris images. Remark that the obtained results with the latter case are very promising. Moreover, they indicate that the approach could be adapted to deal with the so-called One Class Classification problem (OCC) (Irigoien et al., 2014) in which the number of classes can be adapted during the execution of the process.

As a novelty, the dissimilarity computation between two images has been computed as an *a posteriori* histogram distance of the classes' distributions returned by classifiers. More concretely, the Manhattan distance has been utilized for the computation of this difference. The calculation of this value increases the precision of the classification process.

As future work, regarding iris recognition solutions, more classifier combinations –see Mendialdua et al. (2015)– and different histogram distances are going to be tested. Furthermore, new datasets with more individuals or a variety of mobile devices would be also of interest to further validation.

Even more, the presented dissimilarity computation method is going to be tested as a new methodology for image classification. The idea is to make a comparison between different distances results using different image datasets.

Acknowledgments

This work has been partially supported by the Basque Government (IT900-16) and the Spanish Ministry of Economy and Competitiveness MINECO (TIN2015-64395-R).

References

- Aginako, N., Martínez-Otzeta, J., Rodríguez, I., Lazkano, E., Sierra, B., Castrillón-Santana, M., Lorenzo-Navarro, J., 2016. Machine learning approach to dissimilarity computation: Iris matching, in: International Conference on Pattern Recognition.
- Aha, D.W., Kibler, D., Albert, M.K., 1991. Instance-based learning algorithms. *Machine learning* 6, 37–66.
- Breiman, L., 2001. Random forests. *Machine learning* 45, 5–32.
- Cestnik, B., 1990. Estimating probabilities: a crucial task in machine learning., in: *ECAI*, pp. 147–149.
- Daugman, J., 2002. How iris recognition works. *IEEE Transactions on Circuits and Systems for Video Technology* 14, 21–30.
- Daugman, J., 2003. The importance of being random: statistical principles of iris recognition. *Pattern Recognition* 36, 279 – 291. URL: <http://www.sciencedirect.com/science/article/pii/S0031320302000304>, doi:[http://dx.doi.org/10.1016/S0031-3203\(02\)00030-4](http://dx.doi.org/10.1016/S0031-3203(02)00030-4).
- Daugman, J.G., 1993. High confidence visual recognition of persons by a test of statistical independence. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15, 1148–1161. doi:10.1109/34.244676.
- De Marsico, M., Nappi, M., Riccio, D., 2012. Noisy iris recognition integrated scheme. *Pattern Recognition Letters* 33, 1006–1011. URL: <http://dx.doi.org/10.1016/j.patrec.2011.09.010>, doi:10.1016/j.patrec.2011.09.010.
- De Marsico, M., Nappi, M., Riccio, D., Wechsler, H., 2015. Mobile iris challenge evaluation (MICHE)-I, biometric iris dataset and protocols. *Pattern Recognition Letters* 57, 17–23.
- Delac, K., Grgic, M., 2004. A survey of biometric recognition methods, in: *Electronics in Marine, 2004. Proceedings Elmar 2004. 46th International Symposium*, pp. 184–193.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: An update. *SIGKDD Explorations* 11.
- Irigoin, I., Sierra, B., Arenas, C., 2014. Towards application of one-class classification methods to medical data. *The Scientific World Journal* 2014.
- Kim, D., Jung, Y., Toh, K.A., Son, B., Kim, J., 2016. An empirical study on iris recognition in a mobile phone. *Expert Systems with Applications* 54, 328 – 339. URL: <http://www.sciencedirect.com/science/article/pii/S0957417416300148>, doi:<http://dx.doi.org/10.1016/j.eswa.2016.01.050>.
- Martin-Roche, D.D., Sanchez-Avila, C., Sanchez-Reillo, R., 2001. Iris recognition for biometric identification using dyadic wavelet transform zero-crossing, in: *Security Technology, 2001 IEEE 35th International Carnahan Conference on*, pp. 272–277. doi:10.1109/.2001.962844.
- Masek, L., 2003. Recognition of Human Iris Patterns for Biometric Identification. Technical Report.
- Mendialdua, I., Arruti, A., Jauregi, E., Lazkano, E., Sierra, B., 2015. Classifier subset selection to construct multi-classifiers by means of estimation of distribution algorithms. *Neurocomputing* 157, 46–60.
- Mitchell, T.M., 1997. *Machine Learning*. 2 ed., McGraw-Hill, New York.
- Prabhakar, S., Pankanti, S., Jain, A.K., 2003. Biometric recognition: Security and privacy concerns. *IEEE Security and Privacy* 1, 33–42.
- Quinlan, J.R., 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufman Pub., Inc., Los Altos, California.
- Ramaiah, N.P., Kumar, A., 2016. Advancing Cross-Spectral Iris Recognition Research Using Bi-Spectral Imaging. Springer India, New Delhi. pp. 1–10.
- Sheela, S., Vijaya, P., 2010. Iris recognition methods-survey. *International Journal of Computer Applications* 3, 19–25.
- Susperregi, L., Arruti, A., Jauregi, E., Sierra, B., Martínez-Otzeta, J., Lazkano, E., Ansuategui, A., 2013a. Fusing multiple image transformations and a thermal sensor with kinect to improve person detection ability. *Engineering Applications of Artificial Intelligence* 26, 1980 – 1991. URL: <http://www.sciencedirect.com/science/article/pii/S0952197613000791>, doi:<http://dx.doi.org/10.1016/j.engappai.2013.04.013>.
- Susperregi, L., Sierra, B., Castrillón, M., Lorenzo, J., Martínez-Otzeta, J.M., Lazkano, E., 2013b. On the use of a low-cost thermal sensor to improve kinect people detection in a mobile robot. *Sensors* 13, 14687. URL: <http://www.mdpi.com/1424-8220/13/11/14687>, doi:10.3390/s131114687.
- Wildes, R.P., 1997. Iris recognition: an emerging biometric technology. *Proceedings of the IEEE* 85, 1348–1363. doi:10.1109/5.628669.

2.2.2 Periocular and iris local descriptors for identity verification in mobile applications

- **Izenburua:** Periocular and iris local descriptors for identity verification in mobile applications
- **Egileak:** Naiara Aginako, J.M. Martinez-Otzerta, Basilio Sierra, Modesto Castrillón-Santana, Javier Lorenzo-Navarro
- **Aldizkaria:** Pattern Recognition Letters (PRL)
- **Argitaletxea:** Elsevier
- **Inpaktu-faktorea (urtea):** 1,586 (2015)
- **Kuartila:** Q2
- **Zenbakia (Orrialdeak):** -
- **Urtea:** 2017
- **DOI:** <http://dx.doi.org/10.1016/j.patrec.2017.01.021>



Periocular and iris local descriptors for identity verification in mobile applications

Naiara Aginako^a, Modesto Castrillón-Santana^{b,**}, Javier Lorenzo-Navarro^b, José María Martínez-Otzeta^a, Basilio Sierra^a

^aUniversidad del País Vasco (UPV-EHU), Spain

^bSIANI, Universidad de Las Palmas de Gran Canaria (ULPGC), Spain

ABSTRACT

The 2016 International Conference on Pattern Recognition hosted the MICHE-II Contest, with the aim at biometric identification on mobile devices. This paper describes the ideas behind one of the contest submissions, in particular the one ranked 6th (4th in cross-device), including different novelties in relation to the original proposal. Our approach is based on the extraction of local descriptors previous to classification. In this sense, starting from a common iris segmentation information, different normalization procedures are considered, analyzing the use of both iris and periocular patterns. A collection of local descriptors is computed on those patterns, evaluating their performance by means of different classification paradigms in a 10-fold cross validation experiment. Our results suggest the great utility of the periocular area for this problem and dataset. Finally, periocular based classifiers are evaluated on the test set, evidencing an improvement in relation to our original submission, with a promising close future improvement if a fusion approach is adopted.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Iris and fingerprint recognition are the biometric technologies with largest impact in terms of deployed applications. The iris trait in particular, provides a contact-less biometric solution with high accuracy.

The proven iris recognition reliability added to the increasing computational power of mobile phones have attracted the interest of researchers to integrate this technology in smartphones. First attempts go back to the early century, see Cho et al. (2006). More recently the number of proposals is increasing as several mobile applications require security, see for example De Marsico et al. (2014); Barra et al. (2015).

However, the mobile scenario represents one of the remaining challenges related to iris recognition. A systematic evaluation was firstly tackled on the Mobile Iris CHallenge Evaluation (MICHE-I), see De Marsico et al. (2015). The authors main motivation was to evidence the difficulties present in applications where sensors are located in mobile devices, configuring a different playground for traditional biometric systems.

The development of best solvers in this new challenging uncontrolled context would spread the use of automatic identification technologies to mobile devices. This fact would also increase the use of applications where the identification of the user is critical as financial or payment applications.

Similarly to the noisy iris imagery, where two different competitions, NICE-I and NICE-II (Noisy Iris Challenge Evaluation), were devoted to segmentation and classification, respectively, MICHE-I evaluated segmentation approaches; while the 2016 contest edition, MICHE-II, hosted by the 23rd International Conference on Pattern Recognition (ICPR), assumes as starting point the best segmentation approach submitted to MICHE-I. This fact will focus the aim of the challenge to compare features and classification proposals in the mobile iris recognition context independent of the segmentation process. The MICHE II Contest ICPR paper reported results achieved by the seven best performing submissions, see Castrillón-Santana et al. (2016a).

In this paper, we further evaluate the ideas involved in our submission to the ICPR MICHE II Contest, a joint work by the UPV-ULPGC team, see Aginako et al. (2016). The original submission studies the use of a wide collection of local descriptors, to select those that are better suited for this particular problem, to finally fuse them. The main novelties in this extended work are the integration of the periocular analysis in

**Corresponding author: Tel.: +34-928-458-743; fax: +34-928-458-711;
e-mail: modesto.castrillon@ulpgc.es (Modesto Castrillón-Santana,

the process, and the exploration of alternatives within the normalization step.

The integration of periocular information is based on previous good results combining periocular and iris recognition as evidenced in the literature, see Nigam et al. (2015), and even in the MICHE-II contest, see Castrillón-Santana et al. (2016a). These evidences are added to our recent achievements in gender classification combining face and periocular features Castrillón-Santana et al. (2016d).

1.1. Iris datasets

Iris recognition evaluation is a known problem in the computer vision literature, see Burge and Bowyer (2013); De Marsico et al. (2016). To this aim, different datasets have been made available to the community to serve as benchmark for researchers.

The CASIA Iris Image Database (CASIA-Iris) is a known example that after being deployed for the first time in 2002, its fourth version is currently available. CASIA-IrisV4¹ contains 54,601 iris near infrared illumination (NIR) images of roughly 2,000 genuine and 1,000 virtual subjects respectively.

We may also mention a more recent but smaller database, the UTIRIS, see Hosseini et al. (2010), which contains samples of 79 individuals captured in Visible Wavelength (VW) and NIR imaging.

Covering more experimental setups, the Computer Vision Research Lab at the University of Notre Dame² has made available iris datasets for different purposes such as gender or age classification and cross-sensors scenarios.

Considering another point of view, UBIRIS (Noisy Visible Wavelength Iris Image Databases)³, see Proença and Alexandre (2005), was conceived to reduce the controlled conditions present with user cooperation, where distance and pose are particularly relevant. This database was used for NICE-I and NICE-II Contests.

The aforementioned MICHE-I and MICHE-II contests define a benchmark for current mobile user needs, that require robust iris recognition in uncontrolled scenarios in terms of pose, sensor, resolution, illumination, and so on.

The 2016 International Conference on Image Processing (ICIP) presented results for mobile ocular biometrics on the VI-SOB dataset, see Rattani et al. (2016). This collection contains a larger number of individuals, over 500, acquired with smartphones of three different brands and three illumination conditions. The image set contains acquisitions captured in two sessions separated at least two weeks in time.

2. Proposal description

This section describes the steps involving our proposal, summarizing the different alternatives analyzed below. Considering that, as suggested by the MICHE-II protocol, a common iris segmentation approach must be adopted, we made use of the

unsupervised iris detection solution developed by Haindl and Krupicka (2015).

We include here the original proposal submitted to MICHE-II by the ICPR16 deadline, added to new variants that modify the normalization step, and enclose the periocular area. Also, an additional dataset was used to evaluate the performance of the proposed method.

Later, a large collection of local descriptors is considered, analyzing different classification approaches. In a final step, those providing the best accuracy for this particular problem are selected, evaluating also some immediate fusion alternatives.

2.1. Normalization

As mentioned above, iris detection is provided by the solution of Haindl and Krupicka (2015). Fig. 1 illustrates the resulting segmentation mask obtained for an iris capture sample.

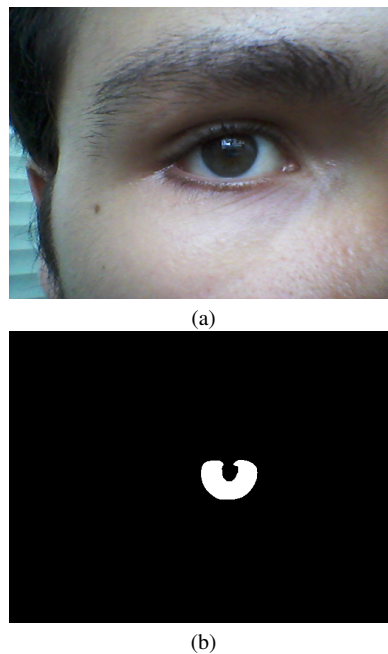


Fig. 1. (a) MICHE-II input image sample and (b) its corresponding segmented iris mask.

The normalization step in our original proposal, adopted the criteria of rescaling the resulting rectangular masked iris pattern to 50×50 pixels, without paying attention to keep the aspect ratio. As seen in Fig. 2a, the presence of occlusions, due to the eyelid in the sample image, may introduce a deformation in the resulting normalized pattern, as it is visible in the mentioned normalized iris.

Added to the adoption of a normalization step that keeps the aspect ratio, the analysis of the average dimension of the iris patterns in the MICHE-II dataset, reveals that it is remarkably larger than the adopted normalized dimensions for at least two of the three sensors. For that reason, a normalized iris pattern of 100×100 pixels, as shown in Fig. 2b, is also evaluated below.

¹<http://biometrics.idealtest.org/dbDetailForUser.do?id=4>

²<https://sites.google.com/a/nd.edu/public-cvrl/data-sets>

³<http://iris.di.ubi.pt/>

The same resolution has been considered for the periocular region, as shown in Fig. 2c. In both cases, the original images may be certainly larger, but the use of very large images would increase significantly the processing cost for feature extraction and do not introduce additional discriminant information.

So far, the normalization step has skipped the common dimensionless polar coordinate system before extracting features, see Daugman (2004). To explore that approximation, we have also adopted a polar representation of 100×400 pixels as shown in Fig. 2d.

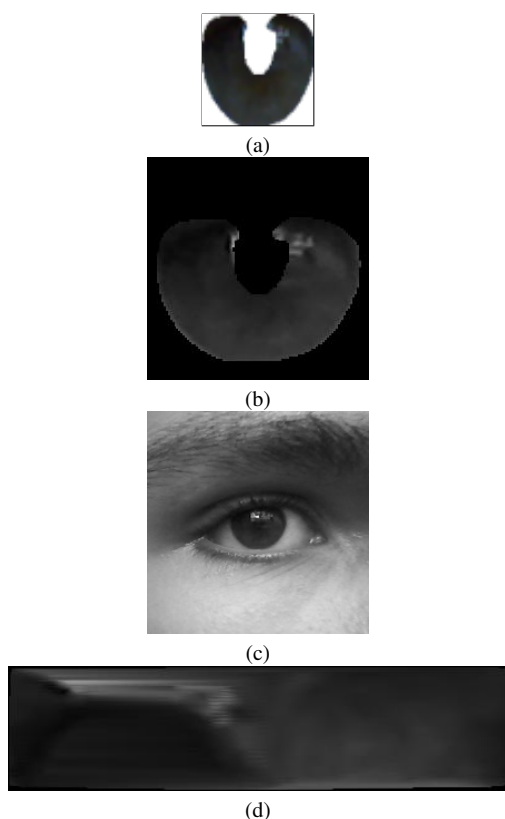


Fig. 2. Normalized iris following the (a) initial MICHE-II ICPR16 Contest submission (50×50 pixels), $I_{50 \times 50}$, the corresponding (b) iris and (c) periocular normalized images keeping the aspect ration (100×100 pixels), i.e. $I_{100 \times 100}$ and $P_{100 \times 100}$. Finally, the iris normalized in polar coordinates, $I_{p_{100 \times 400}}$, is presented in (d) with a resolution of 100×400 pixels.

2.2. Features

As mentioned above, we have adopted an approach based on local descriptors. The use of local descriptors for iris analysis is not new, as evidenced in most approaches submitted to MICHE-II, see Castrillón-Santana et al. (2016a), and previous literature on the topic, see Li et al. (2015), covering also spoofing detection in biometrics, see for example Gragnaniello et al. (2015).

Instead of deciding in advance the particular local descriptor to apply, and based on our previous experience related to

facial analysis, we have selected a collection of local descriptors present in the literature, to evaluate their individual performance in this particular benchmark.

Local descriptors describe a pixel in relation to its neighborhood, re-coding the pixel value. The resulting codes are used to describe the image, commonly making use of an histogram, h_i , where each histogram bin indicates the number of occurrences of each descriptor code. This idea follows a Bag of Words scheme, see Csurka et al. (2004).

Certainly, a histogram serves to reduce the feature vector dimension, but at the same time loses spatial information. That is an undesirable effect. In the context of facial analysis, that loss has been attenuated since the work by Ahonen et al. (2006), where any image is divided into a grid of non overlapping cells, see Fig. 3. The resulting feature vector is composed by the concatenation of the respective cell histograms. In this sense, once the system designer has defined the number of horizontal and vertical cells, respectively cx and cy , the resulting concatenated $cx \times cy$ histograms would conform the image feature vector, $x^d = \{h_1, h_2, \dots, h_{cx \times cy}\}$, where h_i is the corresponding histogram of cell i , and d is a particular descriptor. Thus, given the number of bins per histogram, n_{bins} , the feature vector would contain $cx \times cy \times n_{bins}$ features. There is no doubt that this new feature vector would be larger, but tuning properly the number of cells would offer higher accuracy while keeping a low dimension space with affordable time processing.

Here we adopt the same approach for iris or periocular description, evaluating multiple local descriptors and grid setups. The motivation is that different descriptors would provide different point of views about the image appearance, that could serve to improve performance combining the complementary information in a later fusion stage. Below we summarize the set of descriptors analyzed:

- Histogram of Oriented Gradients (HOG), see Dalal and Triggs (2005). This well known descriptor counts the number of gradient orientations in each image cell. Thus, it is taking edge information into consideration.
- Local Binary Patterns (LBP) and uniform Local Binary Patterns (LBP^{u2}), see Ahonen et al. (2006). Both descriptors have been extensively used in image processing. They both encode a pixel analyzing whether its gray value is greater or not each of its neighbors, composing a binary code. Considering the eight neighbors, 256 different codes are possible. LBP^{u2} focuses on the subset of codes that are more present in texture images.
- Local Gradient Patterns (LGP), see Jun and Kim (2012). Instead of making use of gray values as LBP, LGP makes use of gradient values to encode each pixel.
- Local Ternary Patterns (LTP), see Tan and Triggs (2010). LTP considers three possible relations instead of the two possible analyzed by LBP. The resulting LTP ternary code may be separated into high and low parts, that could be evaluated separately, i.e. LTP_{low} and LTP_{high} .
- Local Salient Patterns (LSP), see Chai et al. (2013). This codification considers the largest differences computed

Table 1. Number of bins per cell histogram.

Descriptor	Number of bins
HOG	9
LBP ^{u2} , NILBP	59
LBP, LGP, LPQ, WLD, LTP _{high} LTP _{low}	256
LOSIB	8
LSP ₀ , LSP ₁ , LSP ₂	57
LSP ₀₁	114
LSP ₀₁₂	171

within each pixel neighborhood. The motivation is to reduce noise influence when gray pixel values are quite similar. We have analyzed the following variants: LSP₀, LSP₁, LSP₂, LSP₀₁ and LSP₀₁₂.

- Weber Local Descriptor (WLD), see Chen et al. (2010). Human perception of a pattern depends both on the change of a stimulus and also on its original intensity. This observation is modeled by the Weber's Law, and integrated in this descriptor.
- Local Phase Quantization (LPQ), see Ojansivu and Heikkilä (2008). With the motivation of being insensitive to centrally symmetric blur, this descriptor computes the short-term Fourier transform (STFT) within the neighborhood.
- Intensity based Local Binary Patterns (NILBP), see Liu et al. (2012). Quite similar to LBP, but considering the neighborhood mean instead of the center pixel gray value.
- Local Oriented Statistics Information Booster (LOSIB), see García-Olalla et al. (2014). Designed as a texture enhancer, it is based on LBP, computing the local statistical information in the whole cell.

However, a consideration must be given to the feature vector dimension, as a larger feature vector would be likely time consuming. For each descriptor, the number of bins per cell histogram is shown in Table 1. They range from 8 to 256 bins per cell.

2.2.1. Classification

The MICHE-II dataset contains images captured with mobile sensors in visible light. Contrary to other well known datasets, this collection avoids the use of NIR images, as those sensors are still not widely available in mobile devices. Extensive details related to the dataset are provided by the ICPR 2016 MICHE-II comparative paper, see Castrillón-Santana et al. (2016a).

MICHE-II proposes the problem of given two iris captures, determine whether they both belong to the same individual or not. For this end, the MICHE-II evaluation protocol has defined two separated datasets:

- The **main dataset**. Provided in advance to ICPR Contest participants to tune their respective submissions. This dataset is basically the same used for MICHE-I containing

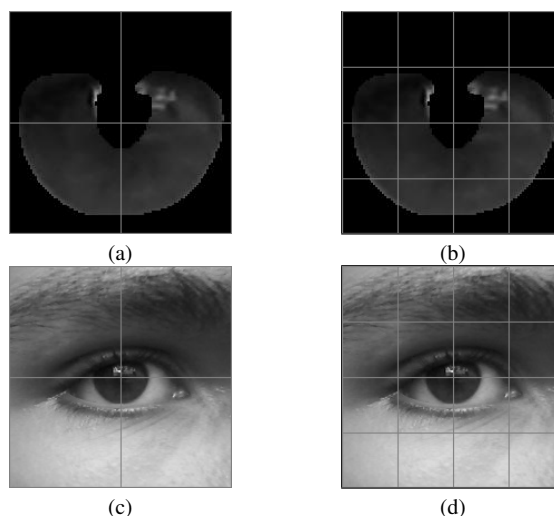


Fig. 3. Normalized masked iris and periocular patterns at (a, c) 2×2 and (b, d) 4×4 grid setups, respectively.

samples of 75 individuals, who have been captured with different mobile sensors, with a total of 3148 samples.

- The **test dataset**. This second dataset was used to rank ICPR 2016 MICHE-II submissions. This collection contains 120 samples, containing also samples of individuals not present in the previous dataset. The total number of different identities reaches 86.

3. Evaluation results

This section is divided into different subsections according to the study done on each dataset. The mentioned **main dataset** serves to tune and select both features and classifiers, that are later used to evaluate their performance on the **test dataset**. An additional evaluation is summarized for VISOB based on the best rates achieved for MICHE-II.

The main idea is firstly to explore multiple descriptor-classifier combinations, evaluating those relevant later with the test dataset. Whenever it is possible, we would also take into account the possibility of fusion based approaches, that may integrate different descriptors and patterns. This attempt is based on our previous experience, particularly on gender classification where the combination of descriptors has evidenced a reduction both in error and ambiguous cases, see Castrillón-Santana et al. (2016b,c).

3.1. MICHE-II main dataset evaluation

In a first step, we analyze the performance considering different classification frameworks applied to the MICHE-II **main dataset**. To do so, we have adopted Weka, see Hall et al. (2009), to explore different classifiers. This is done designing a classification experiment considering 75 classes, defining a 10-fold cross validation.

Table 2. Top-5 descriptors for $I_{50 \times 50}$, and their respective recognition rates in MICHE-II.

Descriptor	Grid	RR (%)	Classifier
WLD	3×3	66.09	K-NN
LBP ^{u2}	4×4	64.82	K-NN
NILBP	4×4	64.38	K-NN
LPQ	2×2	63.68	K-NN
LSP ₀₁₂	4×4	63.68	K-NN

No image pairs combinations are considered in this particular experiment. Therefore, the reported results described below in this subsection, correspond to classification among the mentioned 75 classes.

As mentioned above, different patterns, iris and periocular, and three iris normalization approaches have been analyzed separately:

- Normalized masked iris at 50×50 pixels, $I_{50 \times 50}$, see Fig.2a.
- Normalized masked iris at 100×100 pixels, $I_{100 \times 100}$, see Fig.2b.
- Normalized masked iris in polar coordinates at 100×400 pixels, $I_{p100 \times 400}$, see Fig.2d.
- Normalized periocular area at 100×100 pixels, $P_{100 \times 100}$, see Fig.2c.

Once the normalization step is done, a descriptor d is chosen, computing for each normalized pattern, np , its corresponding feature vector, x_{np}^d . From this point, a collection of classifier paradigms have been studied: K-NN, Bagging, Random Forest, Naive Bayes and C4.5.

For the mentioned respective patterns $I_{50 \times 50}$, $I_{100 \times 100}$, $I_{p100 \times 400}$ and $P_{100 \times 100}$ single recognition rates (RR) are respectively summarized in Tables 2, 3, 4 and 5. In each table, we do not present the results for the whole collection of descriptors and grid setups, but the top descriptors for each pattern. Each table also encloses information related to the classifier and grid setup used. However, the reader must observe that multiple configuration setups for a single descriptor may beat other descriptors. However, we are interested in showing the best descriptors for each particular pattern to the interested reader.

For $I_{50 \times 50}$, see Table 2, WLD reported the best RR hardly larger than 66%. Other classification paradigms instead of K-NN reported significantly worse results. The increase in the normalization resolution, $I_{100 \times 100}$, provides more information which translates into an interesting improvement in the recognition rate, see Table 3. Specifically, LPQ descriptor is greatly affected by the larger resolution, as it provides a remarkable worse performance at lower resolution. Other descriptors do not exhibit a similar behavior. The reported RR is achieved using both K-NN or Bagging for classification.

For $I_{p100 \times 400}$, only the best two RRs are presented, see Table 4. These RRs are the worst among the different patterns and

Table 3. Top-5 descriptors for $I_{100 \times 100}$, and their respective recognition rates in MICHE-II.

Descriptor	Grid	RR (%)	Classifier
LPQ	2×2	74.17	K-NN/Bagging
WLD	4×4	69.03	K-NN
LBP ^{u2}	4×4	67.15	K-NN
NILBP	4×4	63.31	K-NN
HOG	4×4	62.04	K-NN

Table 4. Top-2 recognition rates for $I_{p100 \times 400}$ in MICHE-II.

Descriptor	Grid	RR (%)	Classifier
WLD	4×4	65.18	KNN
WLD	4×4	64.26	Bagging

normalization procedures. A reason to these low RRs +4using the polar representation may be due to the need of requiring more dense grids, that would also affect the feature vector length.

Observing the results for $P_{100 \times 100}$, see Table 5, there is an evident improvement in RR, reaching a value over 91% using LPQ, for the defined 10-fold cross validation in the main dataset. These results evidence the impact of the periocular area for the task and dataset, even if the iris resolution is significantly lower in the processed pattern. Similarly to previous results, LPQ and LBP^{u2} seem to be also relevant for this pattern, but compared to the iris, there are other descriptors providing significant features.

In general K-NN and Bagging reported the best results, being slightly better for K-NN in most cases. Clearly periocular based results are significantly better, observing a difference in the descriptor nature required for different patterns and normalization methods.

3.2. MICHE-II test dataset evaluation

The test set evaluation contains 120 samples divided into two subsets corresponding to probe and gallery. The whole collection of possible combinations defines an experiment with 3600 image pairs.

For evaluation, a pair of iris images is analyzed, deciding whether they both belong to the same identity or not. Given

Table 5. Top-5 descriptors for $P_{100 \times 100}$, and their respective recognition rate in MICHE-II.

Descriptor	Grid	RR (%)	Classifier
LPQ	2×2	91.07	K-NN
LBP ^{u2}	4×4	88.50	K-NN
LSP ₀₁₂	4×4	88.05	K-NN
HOG	4×4	87.77	K-NN
LTP _{low}	3×3	87.73	K-NN

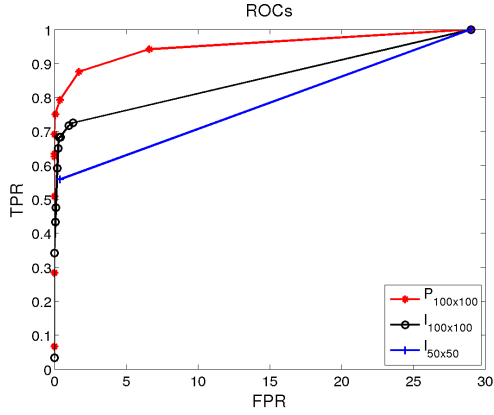


Fig. 4. Best ROC curves obtained for patterns $P_{100 \times 100}$, $I_{100 \times 100}$ and $I_{50 \times 50}$.

an image pair of normalized images, the process continues as follows:

1. Each normalized pattern, np_a and np_b , is classified separately based on their respective feature vectors $x_{np_a}^d$ and $x_{np_b}^d$.
2. Given the number of individuals or classes, c , the resulting posterior distribution histograms, $ph_{np_a}^d \equiv P(c|x_{np_a}^d)$ and $ph_{np_b}^d \equiv P(c|x_{np_b}^d)$, are used to compute the similarity between both patterns using a distance measure. For that purpose, we have adopted the Histogram Difference (HD) computed as the L1-norm of the difference between the resulting posterior distribution histograms.

$$HD(ph_{np_a}^d, ph_{np_b}^d) = \|ph_{np_a}^d - ph_{np_b}^d\|_1 \quad (1)$$

We have evaluated in this dataset the top configurations described in the previous subsection. The best results obtained for the best performing patterns, i.e. $P_{100 \times 100}$, $I_{100 \times 100}$ and $I_{50 \times 50}$, are compared in Fig. 4. There is a clear evidence of the significant improvement provided by the use of the periocular area, i.e. $P_{100 \times 100}$ reported the best performance. This fact made us to include a more detailed study on the results for this pattern.

Fig. 5 presents the receiver operating characteristic (ROC) curves of the top-5 descriptor-classification combination. As mentioned above, the same descriptor with different grid and classifier may be present in these curves. Notation D_G^C corresponds to the classifier C obtained with descriptor D computed in a grid of size G . The observation suggests for the periocular pattern a better behavior for $LPQ_{2 \times 2}^{KNN}$, $LPQ_{2 \times 2}^{BAGG}$ and $HOG_{4 \times 4}^{RF}$.

An alternative may be the combination of multiple classifiers using a classifier ensemble. The idea here is to take advantage of the different information compressed by descriptors of diverse nature. A straightforward possibility is the addition of the dissimilarity measures, combining two or three of the previously mentioned best classifiers. The resulting ROC curves are depicted in Fig. 6. The fusion respective areas under the curve (AUC) are slightly better, suggesting again the benefits of integrating multiple descriptors and resolutions for a problem. However, the application of such simple fusion approach

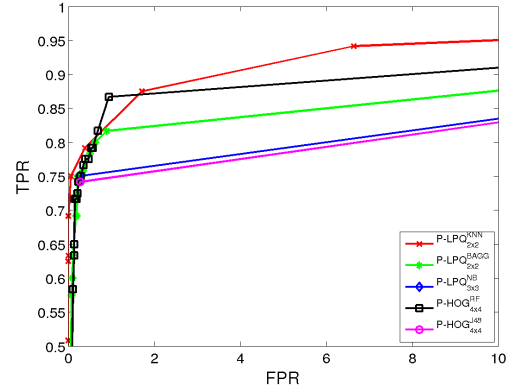


Fig. 5. ROC curves for the periocular based top-5 combinations descriptor-classifier according to the main dataset.

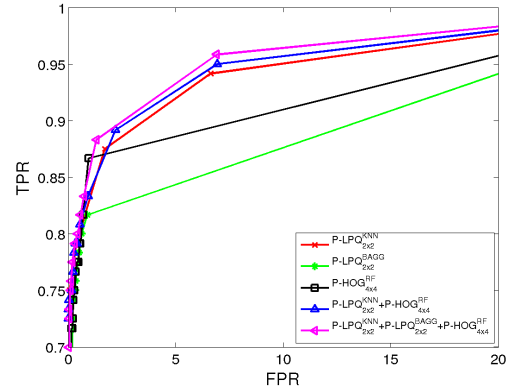


Fig. 6. Best Periocular and fusion ROCs.

for classifiers computed for different patterns, did not reported relevant results. Our intention is to evaluate other fusion possibilities in the future.

3.3. VISOB evaluation

As mentioned in the introduction section, the ICIP 2016 edition hosted an iris recognition challenge on the VISOB (Visible Light Mobile Ocular Biometric) dataset, see Rattani et al. (2016). Similarly to MICHE-II, this challenge focuses on creating a large dataset of iris samples recorded with mobile sensors. The dataset includes images of more than 500 individuals captured with three sensors under three illumination conditions. The samples correspond to two visits, Visit 1 and Visit 2. Each visit encloses two sessions separated roughly 15 minutes. For the ICIP challenge Visit 1 was provided in advance to participants, evaluating their approaches for iris recognition using Visit-2. For this evaluation set, session 1 and 2 comprise roughly 31,000 and 27,000 samples, respectively.

To provide a comprehensive comparison, we have carried out a classification experiment on Visit 2 starting from the best results achieved for MICHE-II. In this sense, we have evaluated only top-5 descriptors for the best two patterns, i.e. $I_{100 \times 100}$ and $P_{100 \times 100}$, which were normalized following the same iris segmentation technique. We would like to point out that the resulting periocular area covered a slightly narrower area in this dataset.

The results obtained for the 10-folds cross-validation on the training set, i.e. Visit 2 first session are summarized in Tables 6 and 7 corresponding to $I_{100 \times 100}$, and $P_{100 \times 100}$, respectively.

Using the iris pattern, $I_{100 \times 100}$, once more the LPQ descriptor achieves higher RR. However, the top values, about 60%, are significantly lower than those obtained by the top-5 for MICHE-II. A possible explanation might be the large number of identities and samples involved in this evaluation. But, as shown below, this effect is not so clearly evidenced using the periocular pattern. Therefore, the iris segmentation approach could also play a role for VISOB.

As mentioned above, the RR loss was not evidenced for the periocular pattern, even if a narrower area was used to extract features. In this sense, the results achieved for $P_{100 \times 100}$ agree with those obtained for MICHE-II. Again the leading descriptor, LPQ, reported and RR over 90%, slightly larger than those obtained for MICHE-II. The rest of top-5 descriptors performed slightly worse, but HOG and LBP^{m2} provided quite similar RRs.

Finally, for both patterns, in a training set and test set experimental setup, the reported RRs are worse, that seems to be explained by the larger test set which encloses more appearance variations. Differences between RR for a single sensor and the theoretically harder cross-device classification are reduced.

The achieved results are not easily comparable to those reported in the ICIP 2016 challenge. Certainly the participants made use of Visit 1 to tune their approaches, reporting the final paper results related to Visit 2, splitting results according to sensor and lighting conditions. Considering that test scenario, three participants reported EER under 10%. Our leading cross-device results reported a classification error around 13.2%, after selecting descriptors and grid setups based on MICHE-II.

3.4. Discussion

The results summarized in this paper have tackled some of the pending tasks not covered in the authors ICPR 2016 paper. Indeed, we have redesigned the normalization process, exploring polar coordinates, see Daugman (2004), pattern resolution, and the adoption of a scaling procedure that keeps the pattern aspect ratio. The increase in pattern resolution, and keeping the pattern aspect ratio have evidenced an improvement in RR.

Last but not least, previous subsections have evidenced the benefit of the use of the periocular pattern for this particular contest, for both main and test datasets. A collection of iris segmentation problems due to occlusions, highlights and artifact, see Castrillón-Santana et al. (2016a), may serve to explain the significantly lower RR achieved for the iris compared to the more stable periocular area.

A deeper analysis of the error distribution in MICHE-II for the best ensemble classifier that fuses $P-LPQ_{2 \times 2}^{KNN}$, $P-LPQ_{2 \times 2}^{BAGG}$

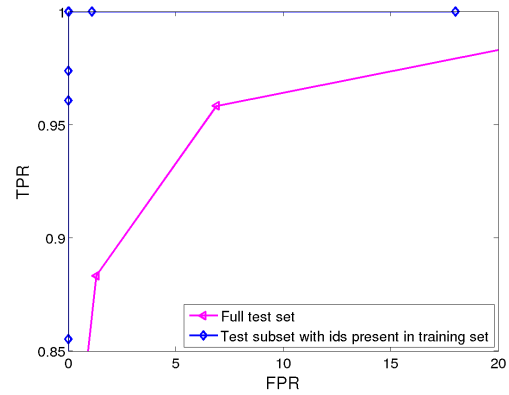


Fig. 7. Best P and only samples in train

and $P-HOG_{4 \times 4}^{RF}$, suggests that the approach is particularly robust for identities present in the training dataset. As seen in Fig. 7 the AUC is rather high if only image pairs with both identities present in the main, i.e. training, dataset are considered, meaning that most errors are present for unseen identities.

To this respect, our approach projects to the domain of known identities, we have not addressed the ideal number of required identities, nor the need to compute any principal component or similar. Unfortunately, existing datasets comprise a reduced number of identities.

Meanwhile, there is still however a wide domain of possibilities to explore, including classification approaches more elaborated such as the Stacked Generalization, see Mendialdua et al. (2015), or the evaluation of a Classifier Subset Selection.

Also, different histogram distances must be explored, to analyze whether they could be more appropriated to compute the dissimilarity, i.e. Kullback-Leibler, Chi Square, Mahalanobis or Jeffrey divergence some possibilities.

As mentioned above, a close future must be the integration in the ensemble of multiple patterns, and the exploration of complementary information. Certainly, there is no clear evidence that the ensemble of top single classifiers would perform better than others. Indeed, top classifiers might share features, being more interesting to combine those that provide complementary information.

4. Conclusion

This paper describes a wider evaluation of an approach submitted to the ICPR 2016 MICHE-II Contest. The approach is based on Machine Learning paradigms and Computer Vision techniques. The proposal studies firstly an interesting range of combinations among patterns, local descriptors and classification approaches, for the labeled as **main dataset**. Later, top setups are evaluated in the **test dataset** suggesting an improvement compared to our previous ICPR submission, particularly when using the periocular area. A similar behavior is observed for VISOB.

Table 6. Recognition rates achieved using KNN for $I_{100 \times 100}$ in VISOB Visit 2 session 1.

Descriptor	Grid	10-fold RR				Train vs test RR			
		all	iphone	oppo	samsung	all	iphone	oppo	samsung
LPQ	2x2	60.1	59.6	68.0	63.1	45.0	42.8	52.5	49.3
WLD	4x4	58.4	60.4	61.5	59.4	42.2	42.7	44.5	43.7
LBP ^{u2}	4x4	59.4	58.5	63.6	63.5	45.8	43.2	49.5	50.0
NILBP	4x4	51.6	52.9	55.0	54.7	39.5	38.9	40.4	42.8
HOG	4x4	50.2	52.0	56.6	53.5	35.5	36.1	40.8	38.5

Table 7. Recognition rates achieved using KNN for $P_{100 \times 100}$ in VISOB Visit 2 session 1.

Descriptor	Grid	10-fold RR				Train vs test RR			
		all	iphone	oppo	samsung	all	iphone	oppo	samsung
LPQ	2x2	92.1	92.9	90.7	92.4	86.8	86.6	84.7	87.9
LBP ^{u2}	4x4	86.2	88.5	83.7	89.0	77.4	80.3	72.6	83.0
LSP ₀₁₂	4x4	81.8	83.3	79.8	84.2	72.5	73.4	69.1	74.9
HOG	4x4	85.3	86.8	84.5	86.5	75.2	76.3	68.8	80.0
LTP _{low}	3x3	77.9	77.0	76.7	78.5	67.2	65.5	64.6	68.3

Finally, a preliminary exploration on MICHE-II of ensemble classifiers suggests the benefits in this particular problem of descriptors fusion. Close future work must enclose the evaluation of multiple pattern fusion, along with classifiers and histogram distances.

Acknowledgments

This work has been partially supported by the Basque Government (IT900-16) and the Spanish Ministry of Economy and Competitiveness MINECO (TIN2015-64395-R).

Authors would like to thank Dr. Ajita Rattani and Dr. Reza Derakhshani for providing the VISOB dataset.

References

- Aginako, N., Martínez-Otzerta, J.M., Sierra, B., Castrillón-Santana, M., Lorenzo-Navarro, J., 2016. Local descriptors fusion for mobile iris verification, in: International Conference on Pattern Recognition.
- Ahonen, T., Hadid, A., Pietikäinen, M., 2006. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 2037–2044.
- Barra, S., Casanova, A., Narducci, F., Ricciardi, S., 2015. Ubiquitous iris recognition by means of mobile devices 57, 66–73. doi:10.1016/j.patrec.2014.10.011. mobile Iris CHallenge Evaluation part I (MICHE I).
- Burge, M.J., Bowyer, K., 2013. *Handbook of Iris Recognition*. Springer Science & Business Media.
- Castrillón-Santana, M., De Marsico, M., Nappi, M., Narducci, F., Proença, H., 2016a. Mobile iris challenge evaluation ii: results from the ICPR competition, in: International Conference on Pattern Recognition.
- Castrillón-Santana, M., De Marsico, M., Nappi, M., Riccio, D., 2016b. MEG: Texture operators for Multi-Expert Gender classification. *Computer Vision and Image Understanding* (in press). doi:10.1016/j.cviu.2016.09.004.
- Castrillón-Santana, M., Lorenzo-Navarro, J., Ramón-Balmaseda, E., 2016c. Multi-scale score level fusion of local descriptors for gender classification in the wild. *Multimedia Tools and Applications* (in press). doi:http://dx.doi.org/10.1007/s11042-016-3653-2.
- Castrillón-Santana, M., Lorenzo-Navarro, J., Ramón-Balmaseda, E., 2016d. On using periocular biometric for gender classification in the wild. *Pattern Recognition Letters* 82, Part 2, 181–189. doi:10.1016/j.patrec.2015.09.014.
- Chai, Z., Sun, Z., Tan, T., Mendez-Vazquez, H., 2013. Local salient patterns - a novel local descriptor for face recognition, in: International Conference on Biometrics (ICB).
- Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X., Gao, W., 2010. WLD: A robust local image descriptor. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32, 1705–1720. doi:10.1109/TPAMI.2009.155.
- Cho, D.h., Park, K.R., Rhee, D.W., Kim, Y., Yang, J., 2006. Pupil and iris localization for iris recognition in mobile phones, in: Seventh ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing.
- Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C., 2004. Visual categorization with bags of keypoints, in: Workshop on Statistical Learning in Computer Vision, ECCV, pp. 1–22.
- Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: Schmid, C., Soatto, S., Tomasi, C. (Eds.), International Conference on Computer Vision & Pattern Recognition (CVPR), pp. 886–893.
- Daugman, J., 2004. How iris recognition works 14, 21–40.
- De Marsico, M., Galdi, C., Nappi, M., Riccio, D., 2014. FIRME: Face and iris recognition for mobile engagement 32, 1161–1172. doi:10.1016/j.imavis.2013.12.014.
- De Marsico, M., Nappi, M., Riccio, D., Wechsler, H., 2015. Mobile iris challenge evaluation (MICHE)-I, biometric iris dataset and protocols. *Pattern Recognition Letters* 57, 17–23.
- De Marsico, M., Petrosino, A., Ricciardi, S., 2016. Iris recognition through machine learning techniques: A survey 82, Part 2, 106–115.
- García-Olalla, O., Alegre, E., Fernández-Roble, L., González-Castro, V., 2014. Local oriented statistics information booster (LOSIB) for texture classification, in: International Conference on Pattern Recognition (ICPR).
- Graganiello, D., Poggi, G., Sansone, C., Verdoliva, L., 2015. An investigation of local descriptors for biometric spoofing detection 10, 849–863.
- Haindl, M., Krupicka, M., 2015. Unsupervised detection of non-iris occlusions. *Pattern Recognition Letters* 57, 60–65.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: An update. *SIGKDD Explorations* 11.
- Hosseini, M.S., Araabi, B.N., Soltanian-Zadeh, H., 2010. Pigment melanin: Pattern for iris recognition. *IEEE Transactions on Instrumentation and Measurement* 59, 792–804.
- Jun, B., Kim, D., 2012. Robust face detection using local gradient patterns and evidence accumulation. *Pattern Recognition* 45, 3304–3316.
- Li, C., Zhou, W., Yuan, S., 2015. Iris recognition based on a novel variation of local binary pattern. *The Visual Computer* 31, 1419–1429.
- Liu, L., Fieguth, P., Zhao, L., Long, Y., Kuang, G., 2012. Extended local binary patterns for texture classification. *Image and Vision Computing* 30, 86–99.
- Mendialdua, I., Arruti, A., Jauregi, E., Lazkano, E., Sierra, B., 2015. Classifier subset selection to construct multi-classifiers by means of estimation of distribution algorithms. *Neurocomputing* 157, 46–60. doi:10.1016/j.neucom.2015.01.036.
- Nigam, I., Vatsa, M., Singh, R., 2015. Ocular biometrics: A survey of modalities

- ties and fusion approaches. *Information Fusion* 26, 1–35.
- Ojansivu, V., Heikkilä, J., 2008. Blur insensitive texture classification using local phase quantization, in: Elmoataz, A., Lezoray, O., Nouboud, F., Mamass, D. (Eds.), *Image and Signal Processing, LNCS 5099*. Springer, pp. 236–243.
- Proença, H., Alexandre, L.A., 2005. Ubiris: a noisy iris image database, in: *13th International Conference on Image Analysis and Processing*, Springer, pp. 970–977.
- Rattani, A., Derakhshani, R., Saripalle, S.K., Gottemukkula, V., 2016. ICIP 2016 competition on mobile ocular biometric recognition, in: *IEEE International Conference on Image Processing (ICIP)*, pp. 320–324.
- Tan, X., Triggs, B., 2010. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Image Processing, IEEE Transactions on* 19, 1635–1650.

2.2.3 Identification of plant species on large botanical image datasets

- **Izenburua:** Identification of plant species on large botanical image datasets
- **Egileak:** Naiara Aginako, Javier Lozano, Marco Quartulli, Basilio Sierra, Igor G. Olaizola.
- **Proceedings:** 1st International Workshop on Environmental Multimedia Retrieval co-located with ACM International Conference on Multimedia Retrieval (ICMR 2014).
- **Argitaletxea:** CEUR
- **Zenbakia (Orrialdeak):** 1222 (38–44)
- **Urtea:** 2014
- **URL:** <http://ceur-ws.org/Vol-1222/paper6.pdf>

Identification of plant species on large botanical image datasets

Naiara Aginako
Vicotech-IK4
Paseo Mikeletegi 57
20009 Donostia-San Sebastián
+34943309230
naginako@vicotech.org

Javier Lozano
Vicotech-IK4
Paseo Mikeletegi 57
20009 Donostia-San Sebastián
+34943309230
jlozano@vicotech.org

Marco Quartulli
Vicotech-IK4
Paseo Mikeletegi 57
20009 Donostia-San Sebastián
+34943309230
mquartulli@vicotech.org

Basilio Sierra
Computer Sciences and
Artificial Intelligence Department
University of the Basque Country
+34943015102
b.sierra@ehu.es

Igor G. Olaizola
Vicotech-IK4
Paseo Mikeletegi 57
20009 Donostia-San Sebastián
+34943309230
iolaizola@vicotech.org

ABSTRACT

The continuously growing amount of multimedia content has enabled the application of image content retrieval solutions to different domains. Botanical scientists are working on the classification of plant species in order to infer the relevant knowledge that permits them going forward in their environmental researches. The manual annotation of the existing and newly creation plants datasets is an outsized task that is becoming more and more tedious with the daily incorporation of new images. In this paper we present an automatic system for the identification of plants based on not only the content of images but also on the metadata associated to them. The classification has been defined as a classification plus fusion solution, where the images representing different parts of a plant have been considered independently. The promising results bring to light the chances of the application computer vision solutions to botanical domain.

1. INTRODUCTION

The digital age has brought the development of new technologies that allow making deeper studies about our reality and therefore, winning a more exhaustive knowledge. In addition, the ever increasing use of digital cameras and sensors in several fields, has led to an exponential growth in the amount of multimedia content being generated every day in the world. Nowadays, the whole society is involved in the generation of any kind of content; it's already a fact that digital technologies are introduced in all aspects of our daily lives.

Although the multimedia analysis techniques in their beginning were focused on application sectors directly related with the technology, their penetration in divergent sectors such as medicine, meteorology, environment it's a reality that is bringing huge progress.

Regarding environmental multimedia content, there is an increasing need of techniques for analyzing, interpreting and labelling of the content in order to enrich the actual knowledge. This automatically extracted knowledge leads to the adoption of new strategies that can improve the actual insight of the environment to move forward in the deployment of new directives to help in its protection and care.

Initiatives such as Tela Botanica and projects such as PI@ntNet foster the development of this kind of technologies. Even more, open competitions as ImageCLEF[1], and more precisely plant identification task [2], where technological researchers focused on multimedia content analysis take part, promote the approach of these two worlds. Newborn mobile applications such as Plantifier, LeafSnap or NatureGate are also examples of the natural synergy tendencies.

The image-based identification of different species of plants that both botanical scientists and expert users have collected has become a key study among plant biology science. On the one hand, one of the peculiarities of plant image analysis is that such images may belong to different plant parts such as leaf, stem or flower. On the other hand, content is also time dependent, thereby increasing the difficulty of the identification task. The latter can be mitigated by using not only image content but also the linked metadata. Thus, the analysis process is enriched and more accurate results can be obtained. This metadata is not only the data that users can add manually but information that nowadays digital cameras impress automatically.

One of the biggest handicaps of multimedia content analysis is to determine the working domain so that afterwards, more domain specific implementations are applied. In the case of ImageCLEF dataset, there is a division of 6 subcategories that identify these domains. Each image has an associated XML which specifies what subcategory belongs to, permitting the abstraction from the domain categorization issue.

In our plant identification approach we used ImageCLEF dataset. This competition was first turned up in 2003. Since then, it has become a benchmark platform for the evaluation of image annotation and retrieval algorithms in several domains such as medical imagery, robot vision imagery or botanical collections. This year, a new lab dedicated to life media LifeCLEF which includes plant identification task has been released. In the past

Copyright © by the paper's authors. Copying permitted only for private and academic purposes.

In: S. Vrochidis, K. Karatzas, A. Karpinnen, A. Joly (eds.): Proceedings of the International Workshop on Environmental Multimedia Retrieval (EMR2014), Glasgow, UK, April 1, 2014, published at <http://ceur-ws.org>

edition, 2013, there were 33 submitted runs. Training data resulted in 20985 images while testing data resulted in 5092.

The rest of the paper is organized as follows: section 2 describes category dependent image analysis (section 2), divided into two subsections that go in depth in the metadata analysis (section 2.1) and in the image content analysis (section 2.2). Section 3 is focused on the classification algorithms for the plant identification purpose while in section 4 fusion and merging methodologies are described. We conclude with a summarization of the obtained results (section 4), pointing out the challenges ahead for the use of content based retrieval technologies in botanical domain.

2. CATEGORY DEPENDENT IMAGE ANALYSIS

As mentioned in the prior sections, the available dataset for ImageCLEF2013 Plant Identification Task is segmented into 2 main categories, *NaturalBackground* and *SheetAsBackground*, that are also divided into several sub-categories: Scan and Scan-like for *SheetAsBackground* category, that are considered equally in our system, and *Leaf*, *Flower*, *Fruit*, *Stem* and *Entire* for *NaturalBackground* category. Both training and testing images have an associated XML describing their metadata that permits the system to separate the images into groups for the later processing and classification.

This subcategory based groups are the key units of the overall plant identification process until the merging done taking into account the Individual Plant Identification, a metadata parameter that determines images that belong to the same plant. For each of the subcategories or groups is necessary to extract all the relevant knowledge. First, inferring this knowledge from the metadata such as localization and date and second, describing the content of images as in detail as possible and using discriminative factors. Not all the implementations have been considered for all the groups, taken decisions permit obtaining better results.

In the next subsections, more detailed explanations are presented regarding the metadata analysis and the deployed image content description algorithms.

2.1 Image metadata analysis: georeference and seasonal nature

Considering the metadata information attached to each of the images, we determine the inclusion of two metadata parameters: GPS data and the date in order to extract knowledge that can improve the plant identification process. These parameters are included not only for the training dataset but also for the testing dataset.

The schema of categories and subcategories of the image dataset delimits the use of these metadata parameters to the *Natural Background* category. Images included within *SheetAsBackground* category don't belong to natural environments; consequently, their latitude, longitude and date parameters don't represent the plant ecosystem. Including these data in the classification process can insert too much noise in the system preventing good results.

2.1.1 Georeferenced data

Since ancient times, studies to determine the influence of topography on species identification have been done. One of the most important factors is the altitude at which each species grows. Therefore, altitude has been considered one of the key indicators for the classification process. Altitude values have been extracted using the actual digital elevation model (DEM) for Europe as the

vast majority of the images belong to France. The inputs to these models are the latitude and longitude data (GPS data).

In this case, the classification process has been focused in the analysis of the altitude parameter, not taking into account longitude and latitude variables as we judge that it could increase the noise level as all the images pertain to a specific country.

2.1.2 Seasonal nature classification

The plants are species that change throughout the seasons. Although not all plants undergo this change that doesn't affect to different parts of the plants in the same way, this seasonal concept has been considered an important factor that can be determinant in recognizing the plant. As a consequence, date metadata parameter has been added to the classification attribute list.

2.2 Image content analysis

After analyzing the metadata associated to each image, only its content can contribute some meaningful information to improve the classification process. We implemented approaches based on algorithms for the extraction of global characteristics such as DITEC[3]; textural characteristics such as Haralick [4], Zernike [5][6] and Local Binary Patterns (LPB)[7]; and parameters to characterize the principal object of the image calculating its solidity, eccentricity, dominant colour and area-perimeter relationship.

DITEC algorithm is based on the statistical modelling of the Trace Transform for global image description. Its main strength lays on the capacity for the description of an image as it extracts the most robust features for the interpretation of the content. This algorithm provides highly discriminative global descriptors at very low dimensionality.

Textural characteristics are very meaningful when the elements representing the image have texture. This kind of features give information about the spatial arrangement of colour or intensities and are very present in natural scenes. When using these attributes we intend to find repeated spatial patterns in images to make a distinction between them.

The identification of an image principal object brings the capacity to describe the overall image more accurately. Doing a good segmentation of the object is crucial since the parameters that are being extracted are totally related with this object withdrawing all the other elements that form the image. Parameters associated to the description of principal objects have been applied only in the case of *SheetAsBackground* and *Fruit* categories.

Regarding *SheetAsBackground*, is the shape of the object that represents the leaf which best discriminates between different species. Different measure properties of image regions related with shape description have been applied:

- **Eccentricity:** It is the ratio of the distance between the foci of the ellipse and its major axis length.
- **Solidity:** Scalar specifying the proportion of the pixels in the convex hull that are also in the region.
- **Area-perimeter relationship:** The number of pixels that belong to the area of the object divided by the number of pixels belonging to the object perimeter.

Even though, we didn't consider it a very discriminative parameter we also added the dominant colour of the segmented object.

Concerning *Fruit* subcategory, as the segmentation process was not as accurate as in the previous case because the photos had been taken in real scenarios, only dominant colour parameter was

extracted as it's also a factor that can make the difference between different types of plants.

2.2.1 Segmentation

Although usually image segmentation has a crucial significance for content description, as mentioned before, our system only uses it for the *SheetAsBackground* category and *Fruit* subcategory. In the first case, an isolated leaf is represented in the image with uneven illumination and possible shadows. We implemented colour clustering techniques based on Local Relative Entropy Method (LRE) [9] for the subtraction of the background. As this background doesn't represent a real scenario, the results for the segmentation of this uniform area are promising and therefore, valid for the implementation of an automatic segmentation process.

In the case of *Fruit* image segmentation, the assumption about the importance of the flower object itself carries the necessity of isolating it from the forest background. As well as in the previous approximation, colour clustering techniques based on Joint Relative Entropy method (JRE) [10] are used.

Even more, we observe that *Stem* subcategory contains predominantly images with tree trunks both in vertical and horizontal that fill the majority of the image. Hence, in order to minimize the effect of the insertion of noisy backgrounds to the system, four fifths of the images are cropped in a fixed direction. To determine this orientation of the trunk along the image, local gradients are analyzed.

3. PLANT CLASSIFICATION

All the image content retrieval solutions include a classification stage where data mining algorithms are implemented. These algorithms are necessary to infer knowledge from the extracted

features. Five different algorithms have been studied with the aim of determining the best one for each of the subcategories: Bayesian Network [11], Naive Bayes [12], SMO [13], SVM [14] and Kstar [15]. For the comparison between classification algorithms, training dataset is split into two subsets, one for the training and the other one for the validation of the implementation. KNIME [16] is an appropriate framework to carry out this learning approach and for the experimentation with a range of algorithms and parameterization of them. It permits working with several feature spaces at a time, therefore it a very suitable framework to undertake the evaluation of the algorithms with the best performance.

As starting point, we considered the classification as totally independent problem for each of the subcategories. The interdependency between some of the images has not been taking into account till the merging of the results. Most suitable features (see section 5) are extracted from all the images belonging to the same subcategory and they are gathered into five groups when all present. Each of the group is also considered an independent classification approach; therefore, the overall classification process is atomized as a subcategory classifications solution based on feature associations.

For the learning of the classification algorithms the training subset of images has been used and we validate the performance of the five implemented classification algorithms using the validation subset. As a result, we got at most five classification modules per category for each feature group. These modules output is a *ClassID* probability list that represents the probability of each image to belong to a plant species.

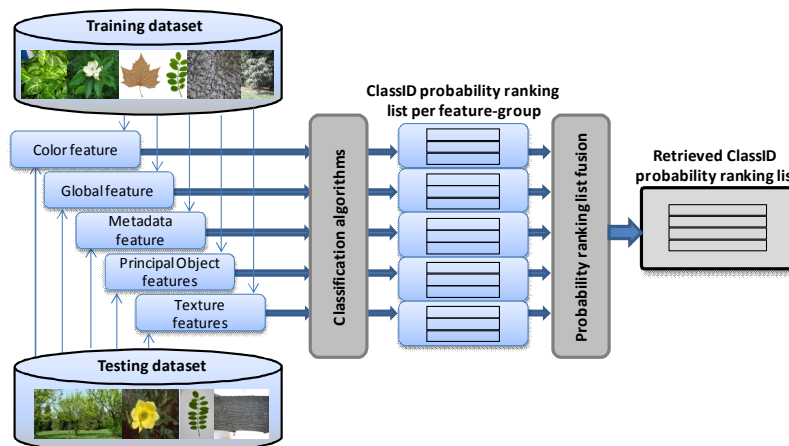


Figure 1 Computation of plant identification based on feature space level fusion

4. FUSION AND MERGING OF CLASSIFICATION RESULTS

We grouped the extracted features in five different groups to analyze their relevance in the identification task results. In general, most of the Content Based Image Retrieval (CBIR)

systems employ a unique probability output to determine the belonging class of a new query image. Multiple feature fusion is a classical technique used in CBIR and pattern recognition to improve the efficiency and robustness of results but this fusion is usually done to feature level. As an alternative to this, we propose an approach that computes the fusion of the classification results at feature space level. Probability scores lists for each of the

feature group are fused using a Leave Out algorithm (LO) [8]. Despite the algorithm was defined for its application using similarity scores, the adoption to probability lists is direct.

For the plant identification of a new query image, its features are extracted taking into account the aforementioned five feature spaces: colour, principal object, texture, global (DITEC) and metadata (see Figure 1). Classification modules have been already trained at feature space level so each feature group vector is classified by the corresponding classifier. As the output of this classification stage, we get a *ClassID* probability ranking list that denotes the probability for that image to belong to each of the plant classes regarding a concrete feature space.

In order to get a unique output, these probability lists are fused by setting the probability of an image belonging to a class to the maximum of the probabilities in each list. The resulting probability list represents the ranking for the plant identification *ClassID*.

$$M_{ij} = \text{Prob.}(\text{img}_{\text{feature space } i} \in \text{ID}_{W_j})$$

where; $\text{ID}_{W_j} = \text{sort}\{\text{Prob}(\text{img} \in \overline{\text{ID}})\}$ $\overline{\text{ID}} = \{\text{ClassID}\}$

M_{ij} matrix is composed of cells representing a tuple that contains the *ClassID* and the probability value of pertaining to that class. Each of the columns represents the probability ranking list for each of the feature spaces.

$$\vec{R}_{jDC} = \text{sort}\{\vec{M}_j\} \quad j = 1, \dots, J$$

\vec{R}_{jDC} vector represent the retrieved *ClassID* probability ranking list (see Figure 1).

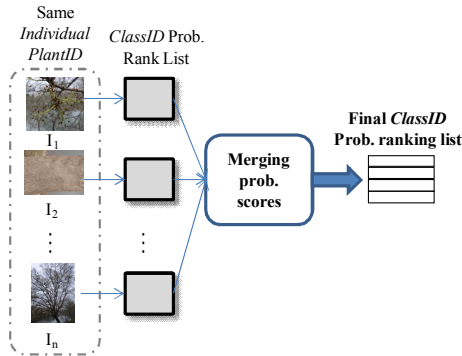


Figure 2 Merging of *ClassID* identification results for images belonging to the same plant (same *IndividualPlantID*)

But there is another fact that must be taken into consideration when estimating classification results: ImageCLEF dataset includes a metadata that must be considered during the plant identification; it is the *IndividualPlantID* which represents an exclusive number identifying images taken from the same plant. Therefore, there is a need of merging results coming from the same plant (see Figure 2). The *ClassID* probability lists belonging to the same plant are merged by means of empirically obtained weights for each of the subcategories.

$$\vec{W}_{SC} = \text{mean}(\overline{\text{Accuracy}})_{\{\text{ClassID}\}}|_{\{T_{SC}\}}$$

where $\{T_{SC}\}$ is the group of images selected for the validation of the classification modules and the definition of the weights.

First, retrieved *ClassID* probability lists (\vec{R}_{jDC}) with the same *IndividualPlantID* are gathered. Taking into consideration the subcategory that the images belong to, probabilities are multiplied by a factor that has been deduced from the performance of the system for each of the subcategories (\vec{W}_{SC}). More precisely, the weight represents the mean accuracy value of the two best classification methods for each of the subcategories. In order to infer this value training dataset has been split into two sets, one for the training and the other for the validation of the classification system. *SheetAsBackground* and *Flower* subcategories are the ones with the highest weight while *Stem* and *Entire* have rather lower values.

Table 1 Weight values for each of the subcategories

SaB	Flower	Fruit	Leaf	Stem	Entire
0,6	0,562	0,409	0,534	0,175	0,07

Second, weighted probability lists are merged by means of the highest probability score that will determine the *ClassID* of the images with the same *IndividualPlantID*.

$$\vec{V}_{kj} = W_{\text{Type}} \cdot R_{1_{kj}}|_{\{k\}}; \quad k = \forall \text{img} \cap \text{PlantID}_Z$$

$$\text{Prob}(\text{img} \in \text{ClassID}) = \max(\vec{V}_{kj})|_{\{\text{ClassID}\}}$$

5. RESULTS

In order to validate the influence of each of the extracted features in the overall process of plant identification we considered to analyze the results of the classification process for each of the subcategories. The results presented in this section are the rate of correct predictions for each of the subcategories. These prediction results have been computed using only the training dataset, splitting this dataset into two sets, 90% of the images for the training of the classification and fusion modules and the other 10% for the validation.

As summarized in **Table 2**, not all the features have been contemplated for all the subcategories, as an example aforementioned associated metadata has not been included in the classification of images pertaining to *SheetAsBackground* category. In addition, all the extracted attributes concerning the identification of the principal object of the image such as the solidity, eccentricity or area-perimeter relationship has only been rated for the *SheetAsBackground* category. By contrast, principal object dominant colour attribute is extracted from both *Flower* and *SheetAsBackground* categories.

Concerning *Leaf* and *Stem* subcategories, metadata, textural and DITEC attributes have been included as the most representative features. As there is no a clear principal object in the image and the colour is not something characteristic other attributes were not considered.

In the case of *Entire* subcategory, images contain the entire natural scene where the plants grow, so the elements of the image are very diverse. This fact introduces lots of noise in the system and the classification of this subcategory is considered the most ambitious. In this case, metadata features and DITEC have been selected for the description.

Fruit and *Flower* are the subcategories where image colouring is a leading figure. Hence, for both subcategories metadata and colour attributes are extracted. Even at first it was considered to add the dominant colour attribute to both cases, due to the weak

results of the segmentation algorithms for *Flower* images we dismiss that possibility and it was only included for *Fruit*. The

opposite of textural features, that are more descriptive in the case of *Flower* subcategory.

Table 2 Subcategory dependent image description features extraction

Subcategory	Texture features			Metadata features		Principal Object features				Global feature	Color
	LBP	Zern.	Haral.	Data	Geo.	Dom. Co.	Sol.	AP rel.	Ecc.	DITEC	HSV
SaB	x	x	x	x	x	x	✓	✓	✓	✓	✓
Flower	✓	✓	✓	✓	✓	x	x	x	x	x	✓
Fruit	x	x	x	✓	✓	✓	x	x	x	x	✓
Leaf	✓	✓	✓	✓	✓	x	x	x	x	✓	x
Stem	✓	✓	✓	✓	✓	x	x	x	x	✓	x
Entire	x	x	x	✓	✓	x	x	x	x	✓	x

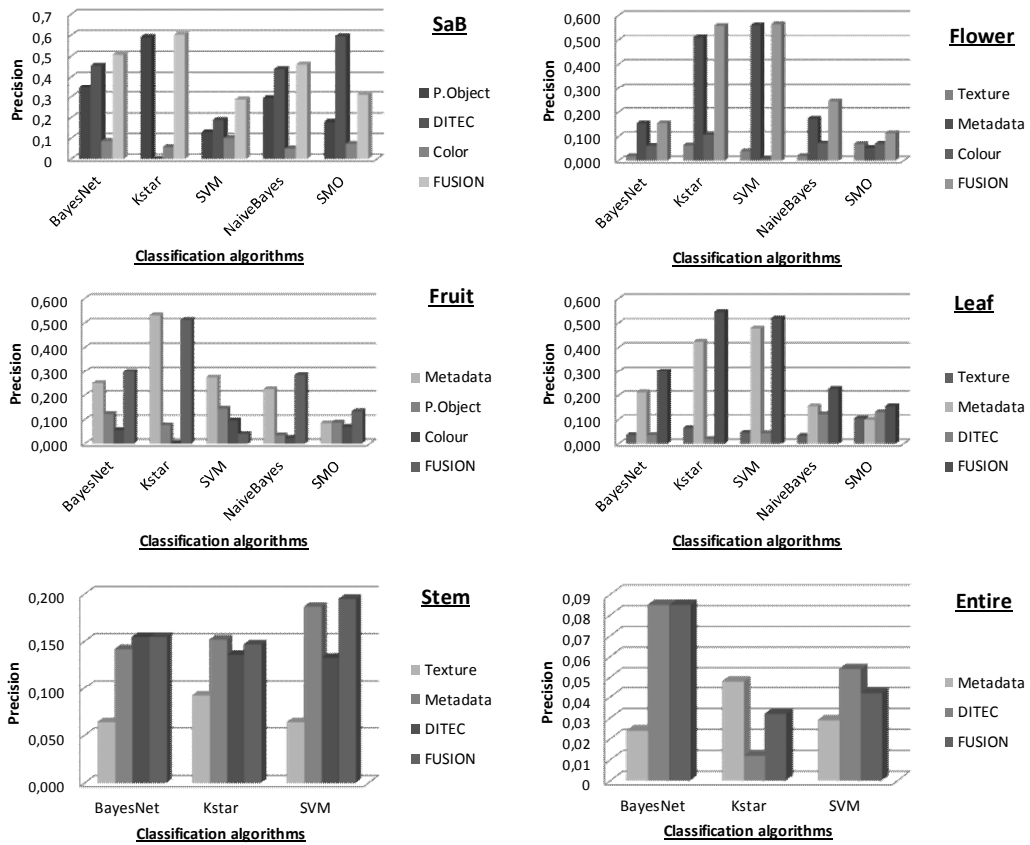


Figure 3 Precision results for each of the subcategories and using different classification algorithms. FUSION label represent the precision results after the application of probability list fusion.

In Figure 3 we resume the results obtained for the classification process visualized separately for each subcategory. For *Flower*, *Fruit* and *Leaf* categories metadata attributes are the ones with the best precision rates. The results for the *SaB*, *Flower*, *Fruit* and *Leaf* categories are quite promising while *Stem* and *Entire* classification doesn't give very good results. In the case of *Entire* category, the inclusion of very diverse elements in the images can distort the general perception of the plant itself and therefore identification task becomes quite difficult. However, if we consider the *Stem* category, we conclude that the extracted features are not feasible for the identification of this type of images.

In general, fusion algorithms increase precision results so a deeper analysis of the consequences of the utilization of these approaches is recommended for plant identification solutions.

5.1 Comparison with ImageCLEF official results

In this subsection some comparative indicators about the results obtained with the method presented in this paper and the overall results of ImageCLEF participants are presented. ImageCLEF results are divided into two different blocks: one of them including only image from *SheetAsBackground* category and the other one for the rest of the dataset images considered as *NaturalBackground* category. All the values for the final validation have been computed only for the testing dataset.

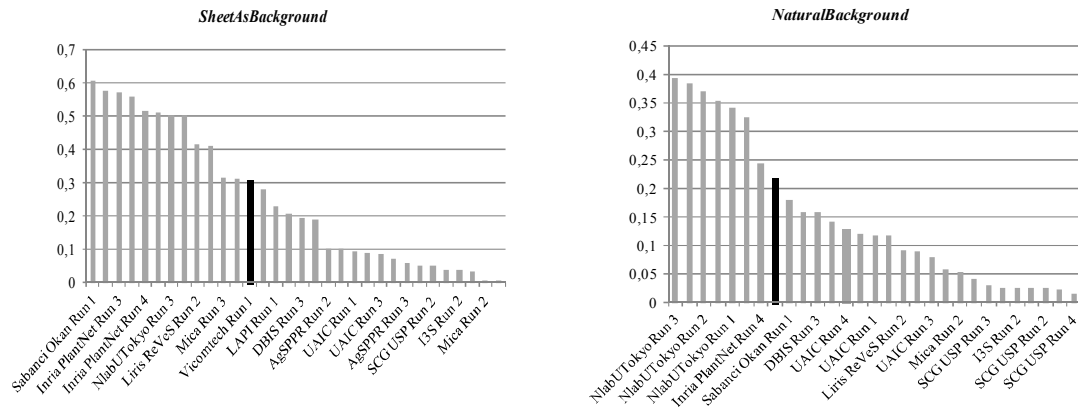


Figure 5 Scores obtained for *SheetAsBackground* and *NaturalBackground* categories

As appreciated in the figures, the results obtained with the described method are among the first half of the participants. In the case of *SheetAsBackground* category, more emphasis must be done in the segmentation process in order to have a better defined content for the analysis.

The bad results obtained for *Stem* and *Entire* subcategories have a direct influence in the scores of the *NaturalBackground* category so better approaches for the classification of these two subcategories are going to be implemented in the near future.

6. CONCLUSION

This paper presents a system for the identification of several plant species based on the analysis of metadata associated to an image and the content of the image. The inclusion of metadata

$$S = \frac{1}{U} \sum_{u=1}^U \frac{1}{P_u} \sum_{p=1}^{P_u} \frac{1}{N_{u,p}} \sum_{n=1}^{N_{u,p}} S_{u,p,n}$$

Figure 4 Primary metric used to evaluate the submitted runs in plant identification task of ImageCLEF 2013

As shown in Figure 4 the metric is a score related to the rank of the correct species in the list of retrieved species, where,

- **U** : number of users (who have at least one image in the test data)
- **P_u** : number of individual plants observed by the u-th user
- **N_{u,p}** : number of pictures taken from the p-th plant observed by the u-th user
- **S_{u,p,n}** : score between 1 and 0 equals to the inverse of the rank of the correct species (for the n-th picture taken from the p-th plant observed by the u-th user)

In the following figures, highlighted in the graphics, the results of the described method for both categories compared with the results of all the participants of ImageCLEF 2013.

parameters reveals an opportunity to refine the results of the image content analysis. Even the described system has been proved for ImageCLEF dataset, the approaches defined in this paper are applicable to collections that contain plant images, only the categorization of plant parts should be keep in mind.

Concerning technical aspect of the system, remark the need of the inclusion of new algorithms that overcome the actual results especially for *Entire* and *Stem* categories. Additionally, merging strategies should consider the insertion of unique image instance identifiers previously in the classification process.

The growing botanical collections ease the inclusion of image retrieval solutions which are considered as very promising by experimented scientists. Competitions such as ImageCLEF are key factors on the approach between image analysis research groups and botanists which permits faster scientific discovery.

Having an accurate knowledge about the identity of plant species is essential for our biodiversity conservation.

7. ACKNOWLEDGMENTS

Our thanks to ImageCLEF organizers and all members of Pl@ntNet project and Tela Botanica initiative who brought us the possibility of researching on the application of multimedia analysis techniques applied to environmental data.

8. REFERENCES

- [1] Caputo, Barbara and Müller, Henning and Thomee, Bart and Villegas, Mauricio and Paredes, Roberto and Zellhofer, David and Gobeau, Herve and Joly, Alexis and Bonnet, Pierre and Martínez-Gómez, Jesus and García-Varea, Ismael and Cazorla, Miguel. ImageCLEF 2013: *The Vision, the Data and the Open Challenges*. Springer, *Lecture Notes in Computer Science*, p. 250-268, 2013
- [2] Gobeau, Herve and Joly, Alexis and Bonnet, Pierre and Bakic, Vera and Bartheemy, Daniel and Boujemaa, Nozha and Molino, Jean-Francois. The imageCLEF 2013 Plant Identification Task, *CLEF 2013 Evaluation Labs and Workshop, Online Working Note*, 2013.
- [3] Olaizola, I.G; Quartulli, M.; Florez, J.; Sierra, B. Trace Transform Based Method for Color Image Domain Identification. 2014. *Multimedia, IEEE Transactions on*
- [4] Haralick, R.M. and Shanmugam, K. and Dinstein, Its'Hak, 1973. Textural Features for Image Classification. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-3 610-621.
- [5] Hu, Ming-Kuei. Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on*, 8.2 (1962): 179-187.
- [6] Teague, Michael Reed. Image analysis via the general theory of moments*. *JOSA* 70.8 (1980): 920-930.
- [7] DC. He and L. Wang (1990), Texture Unit, Texture Spectrum, And Texture Analysis, *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 28, pp. 509 - 512.
- [8] Jović, Mladen and Hatakeyama, Yutaka and Dong, Fangyan and Hirota, Kaoru. Image Retrieval Based on Similarity Score Fusion from Feature Similarity Ranking Lists, *Springer Berlin Heidelberg*, 4223, 2006, pp. 461-470, DOI: http://dx.doi.org/10.1007/11881599_54
- [9] Chein-I Chang, Kebo Chen, Jianwei Wang and Mark L. G. Althouse, A relative entropy-based approach to image thresholding, 1994, Pergamon Pattern Recognition. Vol. 27, No. 9. pp. 1275 1289.
- [10] Wang, J. and Eliza Yingzi Du and Chein-I Chang and Thouin, P.D. Relative entropy-based methods for image thresholding, *Circuits and Systems, 2002. ISCAS 2002. IEEE International Symposium on*, II-265-II-268 vol.2.
- [11] Pearl, J. (1985), Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning, *Proceedings of the 7th Conference of the Cognitive Science Society*, University of California, Irvine, CA. pp. 329-334.
- [12] George H. John, Pat Langley., Estimating Continuous Distributions in Bayesian Classifiers, *Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, 338-345, 1995.
- [13] J. Platt, Fast Training of Support Vector Machines using Sequential Minimal Optimization, *B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning*, 1998.
- [14] Cristianini, Nello; Shawe-Taylor, John; An Introduction to Support Vector Machines and other kernel-based learning methods, *Cambridge University Press*, 2000
- [15] John G. Cleary, Leonard E. Trigg: K*: An Instance-based Learner Using an Entropic Distance Measure. *12th International Conference on Machine Learning*, 108-114, 1995.
- [16] <http://www.knime.org/>

2.3 Bideo analitika

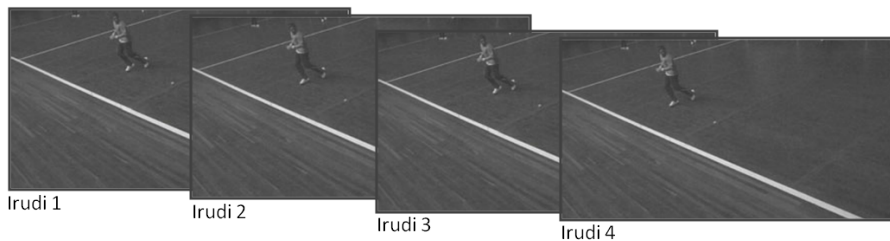
Irudien analitika hasi eta bideoen analitikara bideratu da ikerketa. Hemendik aurrera, bideoa ez da soilik irudien sekuentzia bat bezala ulertuko baizik eta denbora ezaugarria kontutan hartuko da. Hau da, ez dira irudiak independenteki aztertuko; haien arteko harremana kontutan hartuko da. Ikerketa lerro honen barruan honako hauek izan dira ekarpen nagusiak:

1. **Objektuen identifikaziorako metodoen** diseinu eta garapena. Aplikazio konkretu baten esparruan, objektuen segmentazio, jarraipen eta identifikaziorako irtenbideak proposatu dira. Horretarako irudien behe-mailako ezaugarri lokalak eta globalak aztertu eta objektua identifikatzeko arauak ezarri dira. Bideoan zehar objektuaren jarraipena egiteko, bideoaren fotogramen arteko denbora ezaugarria kontutan hartzen da (ikusi 2.3.3 argitalpena eta 2.3.1, 2.3.2 patenteak).
2. **Ekintzen sailkapenerako metodologia** berri baten diseinu eta garapena. Bideoetan agertzen diren ekintzak sailkatzeko irudi prozesamendu eta ikasketa automatikoko metodoen konbinaketan oinarritzen den metodologia bat aurkeztu egin da. Ekintza gertatzen den bideo segmentutik abiatuz, irudien ezaugarri lokalak erabiltzen dira bideo zati hauetako informazioa konprimitzeko. Lortutako ezaugarri berri hauen prozesaketa eta sailkapena eginez ekintza identifikatzen da (ikusi 2.3.4 argitalpena).

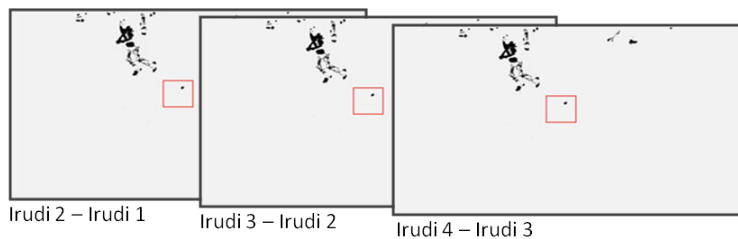
Objektuen identifikaziorako diseinatu eta garatu diren metodoak, telebistatik igortzen diren eduki mota desberdinen analisia egiteko moldatuta dira. [OAL12] **Method of detection and recognition of logos in a video stream** patenteak logoen detekziorako garatutako metodoa babesten du. Logo hauek edukian bertan txertatuta agertzen dira. Bilatzen diren logo posibleak bilatzeko, logoen ezaugarrietan oinarritzen da algoritmoa, hau da, forma erregularrak, kolore biziak eta testua duten irudi zatiak bilatzen dira. Ateratako deskriptoreak, datu base batean gordeta dauden logo posibleen deskriptoreekin konparatu eta logoa dagoen edo ez ondorioztatzen da haien arteko berdintasun faktorea aurretik finkatutako atalase balio baten ginetik baldin badago.

[Ola+08] **Method for detecting the point of impact of a ball in sports events** patentean eta [Lab+14] **Accurate ball trajectory tracking and 3D visualization for computer-assisted sports broadcast** argitalpenean, pilota partidu batean pilotak lurreko zein puntutan jotzen duen aurkitzeko garatutako metodoa aurkezten da. Puntu hori kalkulatu ahal izateko kamara bakar batekin ateratako bideoaren analisia egiten da eta bertan pilotaren forma duten objektuak detektatzen dira irudi

prozesamenduko erremintak erabiliz (ikusi 2.4 irudia). Behin detektatuta, tracking edo jarraipenerako algoritmoa garatu egin da sistemak pilota uneoro non dagoen dakin dezan. Azkenik, bote puntua detektatzeko, pilotaren ibilbidearen norabide aldaketa aztertzen da orduan baita pilotak lurra jotzen duen momentua. Pilotaren kokapena jokalekua mugatzen duen lerroarekiko konparatuz, barruan edo kanpoan jo duen zehazten da (ikusi 2.5 irudia).

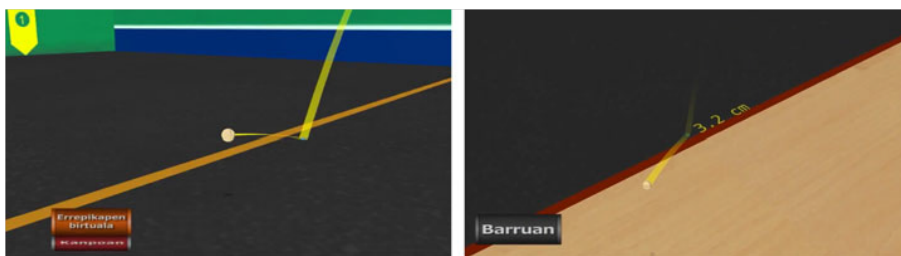


(a) Bideoaren fotogramak



(b) Pilotaren segmentazioa eta jarraipena

Irudia 2.4.: Pilotaren segmentazioa eta jarraipena lurrarekiko ukitze puntua kalkulatzeko



Irudia 2.5.: Pilotaren ibilbidearen berreraiketa

Azkenik, [Agi+17a] **Machine Learning for Video Action Recognition: a Computer Vision Approach** artikuluan bideoetan dauden ekintzen sailkapenerako metodologia berria aurkezten da. Metodologia hau bideoaren denbora ardatzean dagoen informazioaren ezaugarriak ateratzen saiatzen da. Horretarako, fotograma bakoitzetik, behe-mailako deskriptore lokalak ateratzen dira. Deskriptore multzo honek sortzen duen matrizea irudi berri bat balitz bezala ulertzen da. Modu honetan, ezaugarriak denboran zehar nola aldatu diren irudi baten bidez irudikatzen da. Irudi horri, irudi prozesamenduko oinarritzko transformazioak aplikatzen zaizkio informazioa murrizteko eta sailkapenera sartuko diren atributu kopuruak mugatu

ahal izateko. Lortutako emaitzek metodologia berri hau ekintzen sailkapenerako egokia dela adierazten dute.

Laburbilduz, bideoaren analitika barruan egindako lana oso emaitza adierazgarriak loratu ditu. Objektuen identifikazioari dagokionez diseinatu eta garatu diren irtenbideek aplikazioarekin erlazio estua dute eta ondorioz, antzeko arazoetara moldatzeko aldaketak egin beharrezkoak izango lirateke. Ekintzen sailkapenean egindako lana, bideo barruan gertatzen ari diren gauzak ikasteko oinarriak ezarri ditu. Garatutako metodologia domeinuaren independentea da eta ondorioz, orokorragoa. Ikasketa metodoen erabilerak orokortasun hori gehitzea ahalbidetzen du.

Bideo analitikaren inguruan egindako ikerketa, Vicomtech-IK4-en garatu diren I+G proiektu hauetan oinarritu da batik bat (eranskinean proiektu hauen deskribapen sakonago bat eta bakoitzean lortutako emaitzak adierazten dira):

- SIRA- Diseño y desarrollo de un sistema de reconocimiento de marcas comerciales en emisiones televisivas (Proiektu Autonomikoa) (A.5)
- CAPER- Collaborative information, Acquisition, Processing, Exploitation and Reporting for the prevention of organised crime (Proiektu Europarra) (A.6)
- P-REACT - Petty criminality diminution through sEarch and Analysis in multi-source video Capturing and archiving plATform (Proiektu Europarra) (A.7)
- BEGIRA– Diseño y desarrollo de un sistema seguimiento preciso de objetos en transmisiones deportivas (Proiektu Autonomikoa) (A.8)

2.3.1 Method of detection and recognition of logos in a video data stream

- **Izenburua:** Method of detection and recognition of logos in a video data stream
- **Egileak:** Igor G. Olaizola, Naiara Aginako, Mikel Labayen
- **Patente zenbakia:** ES2395448 (T3)
- **Igorpen data:** 2013-02-12



(11) EP 2 259 207 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
08.12.2010 Bulletin 2010/49

(51) Int Cl.:
G06K 9/32^(2006.01)

(21) Application number: 09382086.8

(22) Date of filing: 02.06.2009

(84) Designated Contracting States:
AT BE BG CH CY CZ DE DK EE ES FI FR GB GR
HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL
PT RO SE SI SK TR

- Aginako Bengoa, Naiara
20009, SAN SEBASTIAN (ES)
- Labayen Esnaola, Mike!
20009, SAN SEBASTIAN (ES)

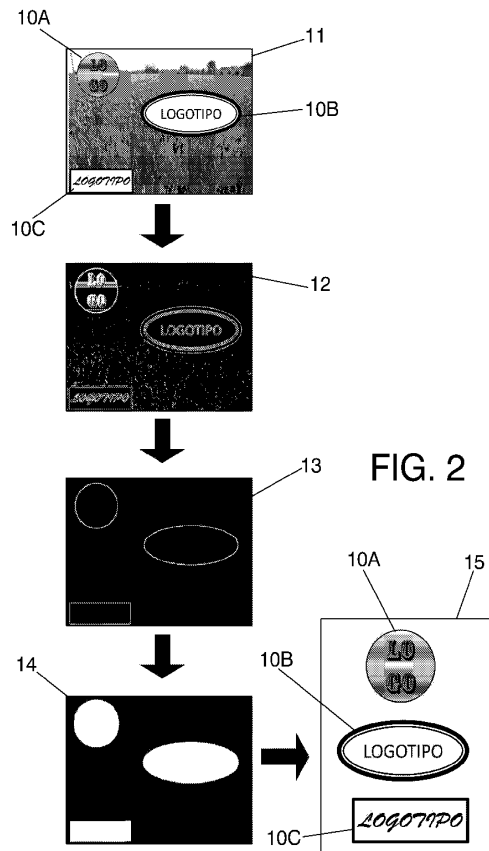
(71) Applicant: Vicomtech-Visual Interaction and
Communication Technologies Center
20009 San Sebastian (ES)

(74) Representative: Carpintero Lopez, Francisco et al
Herrero & Asociados, S.L.
Alcalá 35
28014 Madrid (ES)

(72) Inventors:
• Garcia Olaizola, Igor
20009, SAN SEBASTIAN (ES)

(54) Method of detection and recognition of logos in a video data stream

(57) The present invention relates to a method of detection and recognition of logos in a video data stream comprising the steps of sampling (1) frames of said video data stream; segmenting (3) regular shapes such as, for example, circles, ellipses and rectangles; generating (4) a vector of feature parameters of an image contained in each of said shapes; and comparing said feature parameters with a database for determining (5) whether the images correspond to logos. The frames are captured preferably using a sampling frequency which is dynamically adapted to processing times, for the purpose of allowing the system to work in real time.



EP 2 259 207 A1

DescriptionField of the Invention

[0001] The present invention applies to the field of television and advertising. More specifically, the present invention relates to a method of detection and recognition of logos in video broadcasts.

Background of the Invention

[0002] Advertising and television have always been closely linked, advertising being one of the main sources of income in television broadcasts. However, this business model is undergoing huge changes which are even destabilizing the accounts of television channels.

[0003] There are different factors which are causing this substantial change in the television advertising business. On one hand, the introduction of digital television allows developing systems capable of identifying commercials such that they can be directly eliminated from recordings. For example, US 6,100,941 A discloses a method of detection of video segments corresponding to commercials, using, among other parameters, fade to black frame detection.

[0004] On the other hand, the proliferation of television channels is enormous. New DVB-T (Digital Video Broadcasting Terrestrial) channels, new options such as Imagenio, Internet channels, Video On Demand services, etc. must be added to the offer of analog television, satellite or cable channels. This enormous fragmentation reduces the value of commercials, diverting the value of advertising towards the televised contents themselves instead of towards the commercial breaks characteristic of each broadcast of said contents. This is due to the fact that the linearity of the broadcasts is broken in many cases, whereas the same content can be offered by several channels or on several occasions.

[0005] New solutions and new business models able to continue providing a profit to broadcasters and producers must be conceived against these changes. Pay channels are a classic way to subsist economically, avoiding a strong dependence on commercial breaks. Better service is thus offered. However, it has been seen that the mass of users is not enough to support a large number of such companies, since television viewers do not tend to subscribe to more than one provider.

[0006] As a result, and given the lack of efficiency of commercial breaks in this new environment, there is a growing trend to insert advertising directly in the contents. This is known as Product Placement. It is thus assured that the product will be associated to the contents, which changes the business model. Time slots are no longer valued as highly, but rather the assessment of the content itself is what sets the price for the commercials which are inserted therein. However, this means of advertising products introduces certain technical difficulties when evaluating the duration of brand advertisements in broad-

casts. As they are integrated in the rest of the broadcast image and are not part of a pre-calculated temporal planning, it is not so easy to determine when a brand logo is inserted.

[0007] There are currently a large number of lines of research in the field of Artificial Vision addressing these issues. As a result, there are various solutions for the processing, analysis, segmentation and tracking of image patterns, which is the main objective when finding a specific logo in a scene. However, as occurs in most of the lines relating to artificial intelligence, multipurpose intelligent systems cannot be developed today. Only expert systems achieve satisfactory results, those systems in which knowledge of the problem to be addressed is reduced and can be transmitted to the system. This means that for each case it is necessary to develop, or in the best case scenario, adapt systems because it is impossible to implement a multipurpose solution. W00045291 describes, for example, a system based on the search for images containing text patterns stored in a database, information which can be used to classify the frames and detect commercials.

[0008] However, for a detection of logos integrated in a scene, a robust system which is able to automatically detect multiple logos in an uncontrolled environment, in which the position, luminosity or perspective of said logo are unknown, is necessary. It is also necessary for the method to involve a controlled computational load to avoid collapses and to allow working in real time.

Summary of the invention

[0009] The present invention solves the aforementioned problem by means of a method of detection of logos in a video data stream which speeds up the location of said logos on the basis that most logos have a regular geometric shape, within which there is inscribed the characteristic design of each brand. By searching for logos exclusively within said geometric shapes, the computational load of said search with respect to other earlier, more general location systems is reduced, allowing the present invention to work in real time.

[0010] To that end, the method first samples frames of the video data stream with a certain sampling frequency. For each frame sampled by the method, those areas within the frame having a regular geometric shape are located first. Then, each image contained in one of these areas is characterized by means of a set of parameters which are stored in a vector of feature parameters. These parameters are compared with a database containing vectors with reference values of said parameters, previously calculated for each of the logos which are to be located. This comparison, which involves the calculation of a distance between vectors, allows determining if the image contained within said regular geometric shape corresponds with any of the logos the parameters of which are included in the database.

[0011] Preferably, to facilitate the task of the geometric

figure location algorithms, and to optimize their result, the sampled frames are pre-processed by means of edge enhancement and noise reduction filters. Also preferably, for the purpose of providing a final report to the user, every time a logo is detected in a frame, information relating to which logo is detected, as well as information as to its instant of appearance in the video data stream is stored in a file.

[0012] Preferably, to enable the method to work in real time, being adapted to the processing time of the system and without causing data saturation, the method of the invention contemplates a dynamic modification of the frame sampling frequency depending on the time required by the operator to perform the different steps of the method. Regardless of said sampling frequency, the method preferably samples at least one frame of each scene of the video data stream the duration of which exceeds a pre-determined threshold for the purpose of not losing information relating to any scene when the sampling frequency is very low, avoiding at the same time extracting information from scenes the duration of which is so brief that the information they contain is unnoticeable for the user.

[0013] The geometric shapes which the method must locate are preferably circles, ellipses, and rectangles, preferably in this same order to optimize the computational load generated by said searches upon eliminating, by means of the first searches, candidates for the following searches, which are more difficult in terms of computational load.

[0014] With regard to the parameters which are used to characterize the images and the logos, they preferably comprise, without excluding other possible parameters, a subgroup or all of the following parameters:

- An indicator of the geometric shape of the segmented area.
- Mean and variance values for the colors of the pixels contained in the segmented areas. Each color is defined by a series of parameters within a color space (for example, the values of the red, green and blue channels in the RGB space or of the intensity in a black and white image). These mean and variance values are preferably not obtained as a whole for the entire surface of the logo or image, but rather said surface is previously divided into a set of regions the color of which is characterized individually.
- A measure of the entropy of the image.
- Descriptors of feature points determined by a SIFT (Scale-Invariant Feature Transform) algorithm. Each descriptor is stored in a vector, the number of elements of which is reduced by means of applying a principal component analysis (PCA). This reduction allows reducing the processing time of the step of comparing feature parameters in that the most relevant part of the extracted information by means of the SIFT algorithm is conserved with a lower number of elements.

- A relation between the surface and the perimeter of the logo or segmented area.

[0015] Greater weight is preferably given to the indicator of the geometric shape when performing the comparison with the patterns of the database.

[0016] The described method therefore allows a computationally efficient detection of logos, and it makes use of the usual features of said logos for their characterization, allowing it to work in real time on a video stream. These and other advantages will be apparent in view of the detailed description of the invention.

Brief Description of the Drawings

[0017] For the purpose of aiding to better understand the characteristics of the invention according to a preferred practical embodiment thereof and in order to complement this description, the following figures are attached as an integral part thereof, having an illustrative and non-limiting character:

Figure 1 shows a general scheme of the method according to a preferred embodiment of the present invention.

Figure 2 shows in detail the results of a specific example of application of the method on a frame of a video data stream.

Detailed Description of the Preferred Embodiments

[0018] Figure 1 shows a general scheme of the steps of the method of detection and recognition of logos of the invention according to a preferred embodiment thereof. Said steps are performed sequentially by the following modules of a system also according to a preferred embodiment thereof:

- Capture module, which implements the step of sampling 1 frames of the video data stream with a sampling frequency
- Filtering module, which implements the step of pre-processing 2 the sampled frames, such that the images are prepared for a more successful and computationally efficient processing in the following step by means of noise reduction and edge enhancement filters.
- Segmentation module, which implements the step of segmenting 3 areas with regular geometric shapes.
- Parameter extraction module, which performs the step of extracting 4 feature parameters from each image contained in said regular shapes.
- Classification module, which performs the step of determining 5 whether said images correspond with any logo the parameters of which are contained in a database of the system.

- Indexing module, which performs the step of storing 6 the result of the classification in storage means 9, for example a register.
- Time management module, which performs the step of time management 7 in which the sampling frequency is dynamically modified so that the system works in real time.

[0019] The mentioned steps are described below in detail.

[0020] The necessary number of frames per second to be analyzed so that the system can work in real time is defined in the capture module. In a preferred embodiment in which the method of the present invention is applied to the case of digital television, I-type frames are the most appropriate for being used in the extraction of logos since they are those which have the lowest number of errors. In another preferred embodiment, in addition to these frames, intermediate frames (B) can also be introduced between the I-frames in order to obtain a more precise result since a better extraction of the regions of interest is thus observed.

[0021] In order to be able to adapt the images for the step of segmenting images, the filtering module performs a pre-processing to remove most of the noise from the image and extract the edges from the image. The image is then binarized in order to be able to obtain a contour definition that is adapted to the needs of the system. Mainly due to the phenomena known as blurring, which is the noise introduced by the movement occurring between continuous frames of the video, said pre-processing of the image is complicated in the case of images extracted from a video signal and it becomes a very important part of the method of the present invention. Therefore algorithms are applied to reduce such noise due to the movement. Anisotropic filtering is first applied to smooth the edges of the image and thus reduce the noise of the images extracted from the videos.

[0022] Then each of the channels of the image (for example, the YUV color space, which includes one luminance value and two chrominance values, is used in television) is taken and the Sobel operator is applied to detect the edges in each of them. The edges of each of the channels are considered as an edge of the image and a binary image representing all the pixels belonging to the edges is formed. Thus, after pre-processing the image, an edge image is obtained in which most of the edges of the image are present. This image is still not suitable for segmenting regions of interest. It is therefore necessary to perform further processing. In this processing, morphological operations known by those skilled in the art, such as erosion and dilation, are applied to obtain a more defined edge image. A hybrid median filter can also be applied to disregard isolated pixels which do not belong to the edge of an object. Finally, the image is filled in order to be able to establish the main areas of the image and to thus define the contours of said areas, which will be the contours from which the logos of the

images will be extracted.

[0023] The element identified as possible logo is extracted or segmented in the segmentation module. To that end, it is considered that the logos, which are generally easily and simply identified by people, are designed as a regular shape with text or other motifs in it. Therefore, the segmentation algorithms are precisely regular shape extraction algorithms. Circles, ellipses and rectangles have been detected, as they are the most frequent regular shapes in the design of logos.

[0024] This segmentation of regular shapes is performed sequentially. Given that circles are specific cases of ellipses, circles are segmented first, excluding the contour of these objects once they are segmented, and then ellipses are detected. Finally rectangles are segmented. Contours which can lead the rectangle detection algorithm to confusion are thus extracted from the image. This is due to the fact that the contours of circles and ellipses can be considered as lines and therefore be analyzed as candidates for forming a rectangle. For each possible region of interest, an output is given in which the position of this region is indicated with respect to the image and the regular shape which defines it, in order to thus be able to use this information as yet another characteristic in the classification.

[0025] The Hough Transform is applied for the detection of circles. Said transform is applied in the binary image which is obtained after processing the image in the filtering module. The Hough transform is based on the idea that each line perpendicular to the tangent straight line of each point of a circumference passes through the center of said circumference. Thus, if the perpendicular lines of each point of the contour image are calculated, points with very high intensity values will appear in the places in which the centers of the circles are located.

[0026] Once all the possible points for the centers of the circumferences are accumulated, a threshold is applied so that only the possible centers in which a minimum number of points of the binarized image contribute. More or less perfect circles are detected depending on this threshold. For the case at hand, said threshold is not very restrictive since the contouring of the image may not be optimal and therefore imperfect circumferences are accepted.

[0027] For the detection of ellipses, a possible algorithm starts from a closed contour image defining the possible objects of the image. To define the location of the center of an ellipse, it is enough to choose three of its points the tangents of which are not parallel. These three points are paired in two groups and the intersection of their tangents is sought. If said point of intersection is joined with the mid-point of the straight line joining the two points, a straight line is defined in which the center of the ellipse is located. If this straight line is calculated for each of the two groups, the intersection between these two straight lines is the center of the ellipse. Based on this property of the ellipse, possible centers for the ellipses are defined, based on which it is determined wheth-

er the segmented object is an ellipse or not.

[0028] Finally, for the detection of rectangles, the definition of a rectangle as a parallelogram the sides of which form right angles with one another is taken as the starting point. The image is segmented based on this characteristic. The first step is to find straight lines that can form a rectangle. Once these straight lines are detected, it is necessary to check if these segments can form a rectangle, depending on the angle formed between them and on their situation.

[0029] The Hough transform, which converts segments of the original coordinate space into points in the Hough plane, can be used for the detection of the segments in the binary contour image. Once the possible segments of the image are detected, they are clustered depending on the angle value with respect to the x-axis. This clustering takes into account both the segments parallel to one another and perpendicular segments.

[0030] The segments are clustered by always taking the perpendicular segment located at the least distance until forming a quartet. Once the segments are clustered, it is analyzed which segments can belong to one and the same rectangle and which cannot. For each group, the distance between the centroid of the possible rectangle created by the candidate segments and the mid-point of the latter is calculated first. If said distance does not exceed a threshold, it is considered that the segments can form a rectangle. In the event that said threshold is exceeded, this group of segments does not create a rectangle.

[0031] Once the centroid is checked, the vector product of the contiguous segments is analyzed. If all the vector products have the same sign, these four segments are no longer candidate segments and are considered the sides of a rectangle. This check assures that the candidate segments delimit a closed area.

[0032] Once all the segment quartets are analyzed, trios are taken and the fourth segment is searched for in the image in order to be able to face possible errors in the detection of the segments. In the event that said line is not found, this fourth segment is determined from the other three. For the determination of whether they form a rectangle, the same process is performed as in the case of having the four segments. In order to be able to define the rectangle, the points of intersection of the segments are found and its sides are redefined.

[0033] Once the segmentation process has ended, feature parameters are extracted from the segmented image in the next step. Based on the results obtained for the extraction of feature parameters from each segment, the feature parameter extraction module creates a vector which will be the identifier of the logo for the step of classification. In a preferred embodiment of the present invention, the feature parameters extracted in this step are:

- The shape, indicating whether it is a circle, an ellipse or a rectangle.
- The color, by means of calculating the mean and

variance of the channels of a color space (for example, the YUV color space, which includes one luminance value and two chrominance values, is used in television). Said channels determine the color of each pixel of the image. In a preferred embodiment, the parameters relating to the color are extracted in a plurality of areas within the image or logo, said areas being separated depending on their entropy value, which is a statistical measure used to characterize the texture of an image, and which represents a measure of the disorder or complexity of said image. For example, an image which is entirely black or entirely white has zero entropy. The calculation of entropy is based on the levels of gray of each of the channels forming the image. The images with a greater number of levels have higher entropy.

- The entropy calculated for the entire segmented image.
- The area-perimeter relation, which is an invariant relation with respect to the various geometric transformations an image may experience.
- The extraction of feature points of interest, performed by means of the known SIFT algorithm, which is a method for detecting and extracting descriptors of local features from the images, detailed in detail in US 6,711,293 B1. These points of interest are invariant with respect to the rotation and change of scale of the image; and partially invariant with respect to the noise in the image and the changes in the lighting and perspective. These points of interest are the local extremes of the differences of the images filtered with Gaussian filters at several scales. Once these points are detected, descriptors are used to characterize each point. These descriptors are built by assigning an orientation and magnitude to the points next to the points of interest.

[0034] In a preferred embodiment, said descriptors are stored in vectors, which are subsequently subjected to PCA analysis. PCA analysis is a statistical technique for the synthesis of the information, or reduction of the dimension, explained, for example, in *"A Tutorial on Principal Component Analysis"*, by Jonathon Shlens. The objective is to reduce the number of variables to a lower number, losing the least amount of information possible. The new principal components or factors will be a linear combination of the original variables, and they will furthermore be independent of one another.

The dimension of the descriptor is reduced with PCA, thus reducing the processing time of the step of comparison. In a non-limiting example, it is possible to go from 128 elements in the descriptor obtained by means of the SIFT algorithm to 20 elements, achieving suitable characterization with a lower computational cost.

[0035] Based on the similarity of these feature parameters with the same feature parameters of reference logos, it is possible to determine in the classification module which logo is the possible logo detected or whether, in

contrast, it cannot be considered a logo. To that end a classification vector is created which consists of a numerical value vector incorporating the information of the described parameters (shape, entropy, color, area-perimeter relation, SIFT feature points), various weights being able to be assigned to said parameters when performing the classification. In an even more preferred embodiment, the greater weight is assigned to the value of the shape, since it is one of the identifiers of the logo.

[0036] The comparison with the parameters of the reference logos is performed using this vector. To that end a distance between the vector calculated for the segmented image and the vectors of the database 8 containing the same parameters calculated for the logos which are desired to be known is calculated. The smallest distance will determine the logo contained in the segmented image, provided that said distance does not exceed a certain pre-established threshold. If the distance exceeds said threshold, it is considered that the segmented image does not correspond to any logo.

[0037] The times each logo appears are counted in the indexing module and said information is stored in a result register for subsequently extracting the stored information.

[0038] Working in real time is important in the method according to the invention to avoid collapses given that it is intended to work continuously. To that end, the time elapsed between the step of capturing 1 frames and a final step of the method, indicative of the total run time, and which can be both the step of determining 5 whether the images correspond with logos, or the step of storing 6 the final result of the recognition of logos, is observed in the time management module; and the sampling frequency is modified such that the number of frames extracted with said sampling frequency can be processed in the time observed without the system collapsing.

[0039] In order not to lose information of any scene when the sampling frequency is too low, the video capture module preferably incorporates means for detecting changes in scene by means of the comparison of the degree of similarity between frames, such that at least one image from each scene is sampled, provided that said scene is longer than a certain threshold duration thereby enabling to discard scenes that are too short the information of which is not noticeable for the user.

[0040] Figure 2 shows a specific example of application of the method to a sampled frame 11 containing three logos 10A, 10B and 10C integrated in the image of the frame.

[0041] The pre-processed image 12 shows the same sampled frame 11 after applying a pre-processing which enhances the edges of the image and binarizes it. The processed image 13 has the same pre-processed image 12 after a further processing according to the techniques described in the present description which allow eliminating points on which the geometric figure detection algorithms are to be run for the purpose of reducing the computational load of said algorithms.

[0042] The segmented image 14 has the result of applying the segmentation algorithms to the processed image 13, thus locating the geometric figures contained in the original sampled frame 11. The content of said geometric figures is extracted in 15 such that the extraction of parameters and subsequent identification of the logos 10A, 10B and 10C can then be performed.

[0043] In view of this description and figures, a person skilled in the art will understand that the invention has been described according to some preferred embodiments thereof, but that multiple variations can be introduced in said preferred embodiments without departing from the object of the invention as it has been claimed.

Claims

1. Method of detection and recognition of logos in a video data stream comprising the step of:

- sampling (1) frames of the video data stream with a sampling frequency;

characterized in that it comprises the steps of:

- segmenting (3) areas with regular geometric shapes in each sampled frame;
 - for each segmented area, generating (4) a vector of feature parameters the elements of which are numerical values having feature parameters extracted from an image comprised in said segmented area in said sampled frame;
 - determining (5) whether said image comprised in said segmented area corresponds to a logo depending on the distance between said vector of feature parameters of said image and a set of vectors the elements of which are numerical values of said feature parameters extracted for each logo from a set of reference logos, said set of vectors being stored in a database (8).

2. Method according to claim 1, **characterized in that** it further comprises a step of pre-processing (2) of the sampled frames by means of edge enhancement and noise reduction filters, prior to the step of segmenting the areas with regular geometric shapes.

3. Method according to any of the previous claims, **characterized in that** it further includes a step of time management (7) which dynamically modifies the sampling frequency depending on a time elapsed between the step of sampling (1) frames and the step of determining (5) whether the images comprised in the segmented areas correspond to a logo.

4. Method according to any of claims 1 to 3, **characterized in that** it further comprises a step consisting of storing (6) in storage means (9), for each frame

in which it is determined that an image comprised in a segmented area of said frame corresponds to a logo, a name of said logo and a time code indicating a time in which said frame appears in the video data stream.

5
10
15
20
25
30
35
40
45
50
55

5. Method according to claim 4, **characterized in that** it further includes a step of time management (7) which dynamically modifies the sampling frequency depending on a time elapsed between the step of sampling (1) frames and the step of storing (6) in storage means (9) the name of the logo and the time code.

6. Method according to any of the previous claims, **characterized in that** in the step of sampling (1) frames, at least one frame of each scene of the video data stream is sampled, the duration of which is greater than a duration threshold.

7. Method according to any of the previous claims, **characterized in that** in the step of segmenting (3) areas with regular geometric shapes, said areas with regular geometric shapes are at least areas in the shape of circles, ellipses and a rectangles.

8. Method according to claim 7, **characterized in that** in the step of segmenting (3) areas with regular geometric shapes, the areas in the shape of circles are segmented first, the areas in the shape of ellipses second, and the areas in the shape of rectangles third.

9. Method according to any of the previous claims, **characterized in that** in the step of generating (4) a vector of feature parameters, said feature parameters of the images comprised in the segmented areas comprise an indicator of the geometric shape of the segmented area.

10. Method according to claim 9, **characterized in that** in order to calculate the similarity between the feature parameters of the images comprised in the segmented areas and reference values of said feature parameters for market brands comprised in a database (8), a weight is assigned to said indicator of the geometric shape of the segmented area that is greater than weights assigned to the remaining feature parameters of said image.

11. Method according to any of the previous claims, **characterized in that** in the step of generating (4) a vector of feature parameters, said feature parameters of the images comprised in the segmented areas comprise mean and variance values of parameters defining a color of pixels comprised in said segmented areas.

12. Method according to claim 11, **characterized in that** said mean and variance values are determined independently for a plurality of regions within the segmented area.

13. Method according to any of the previous claims, **characterized in that** in the step of generating (4) a vector of feature parameters, the feature parameters extracted from the images comprised in the segmented areas comprise a measure of the entropy of said image.

14. Method according to any of the previous claims, in the step of generating (4) a vector of feature parameters, **characterized in that** said feature parameters of the images comprised in the segmented areas comprise vectors containing descriptors of feature points of said images obtained by means of a SIFT algorithm, the dimension of said vectors containing descriptors being reduced by means of a principal component analysis.

15. Method according to any of the previous claims, **characterized in that** in the step of generating (4) a vector of feature parameters, said feature parameters extracted from the images comprised in the segmented areas comprise a parameter which relates the surface and the perimeter of the segmented area.

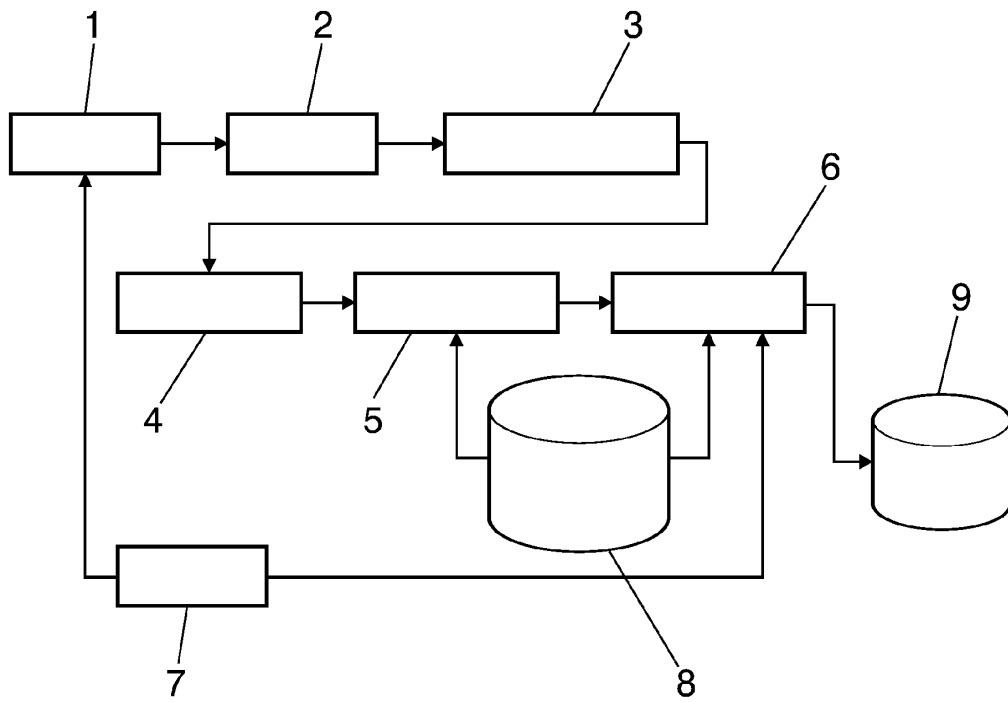


FIG. 1

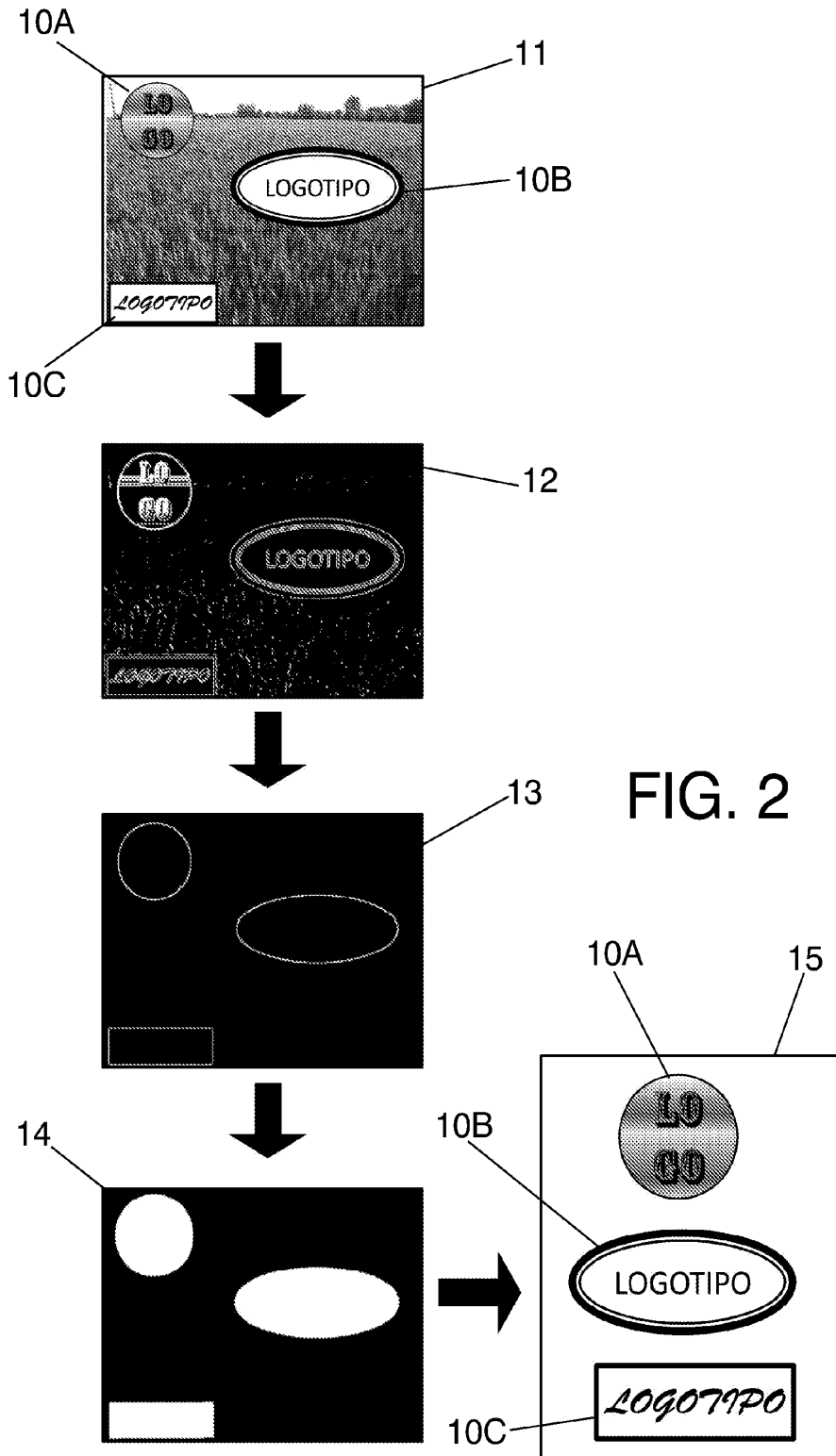


FIG. 2

2.3.2 Method for detecting the point of impact of a ball in sports events

- **Izenburua:** Method for detecting the point of impact of a ball in sports events
- **Egileak:** Igor G. Olaizola, Julián Flórez, J.C. San Román, Naiara Aginako, Mikel Labayen
- **Patente zenbakia:** ES2402728 (T3)
- **Igorpen data:** 2013-03-13

(19)



(11) EP 2 455 911 A1

(12)

EUROPEAN PATENT APPLICATION

(43) Date of publication:
23.05.2012 Bulletin 2012/21

(51) Int Cl.:
G06T 7/00 (2006.01) A63B 71/06 (2006.01)

(21) Application number: 10382310.0

(22) Date of filing: 23.11.2010

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB
GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO
PL PT RO RS SE SI SK SM TR
Designated Extension States:
BA ME

- Flórez Esnal, Julián
20009, San Sebastian (ES)
- San Román Otegui, Juan, Carlos
20600 Eibar (Guipúzcoa) (ES)
- Aginako Bengoa, Naiara
20009, San Sebastian (ES)
- Labayen Esnaola, Mikel
20009, San Sebastian (ES)

(71) Applicant: Vicomtech-Visual Interaction and
Communication Technologies Center
20009 San Sebastian (ES)

(74) Representative: Carpintero Lopez, Francisco et al
Herrero & Asociados, S.L.
Alcalá 35
28014 Madrid (ES)

(72) Inventors:
• García Olaizola, Igor
20009, San Sebastian (ES)

(54) Method for detecting the point of impact of a ball in sports events

(57) The invention relates to a method for determining the point of impact of a ball in a playing field during a controversial piece of play in a sports event and comprises the steps of: recording the contentious area during the game by means of a single camera, extracting the images corresponding to the controversial piece of play, selecting the area corresponding to the ball, calculating

the coordinates of the ball in pixels in each image, determining the point of intersection of the two straight lines joining the previous points and transforming the point of intersection into real coordinates. As a result of these steps it is possible to resolve the controversial piece of play with a single camera and without the aid of accessories such as radar signals, electric signals, etc.

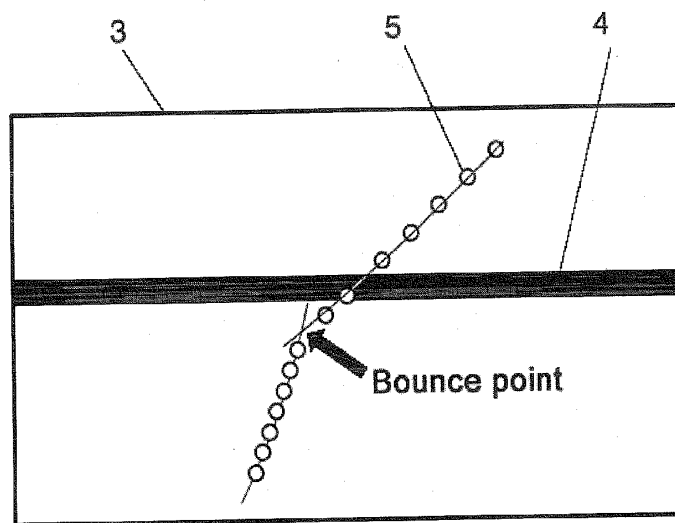


FIG. 2

EP 2 455 911 A1

DescriptionField of the Invention

[0001] The present invention applies to systems for aiding refereeing in ball sports. More specifically, it relates to a method for determining the point of impact of a ball in sports events in which the playing ground is limited by means of lines.

Background of the Invention

[0002] There are several systems adapted for detecting the point of impact of a ball in a playing ground. The purpose of these systems is to aid referees in making decisions in controversial pieces of play while the game itself takes place.

[0003] Devices based on transmitters and receivers of electromagnetic waves such as the one described in patent WO89/00066 are known. These systems allow a rather precise determination of the point of impact, but require structural changes in the ball and the playing field.

[0004] There are also visual devices such as the one described in patent W096725986, based on infrared cameras with which the point of impact of the ball can be seen. These systems have the drawback of not being able to be implemented in outdoor enclosures, since the rain, the wind and other atmospheric conditions and changes in lighting affect the reading of the position of impact.

Object of the Invention

[0005] The object of the present invention is to provide a system and method which enable the detection of the point of impact of a ball in the proximity of the lines delimiting the playing ground or in the line itself in real time and preventing the technical problems set forth above. To that end, the method of the invention proposes recording the contentious area during the game by means of a single camera, extracting the images corresponding to the controversial piece of play, selecting the area corresponding to the ball, calculating the coordinates of the ball in pixels in each image, plotting the points with the coordinates calculated in the previous step, determining the point of intersection of the two segments joining the previous points and finally transforming the point of intersection into real coordinates. For the calculation of the areas of interest and point of intersection of the segments, the movement and velocity vectors are preferably used. Preferably, it is all performed in a greyscale to simplify and cut down on resources.

[0006] The method has the additional advantage that the plane of the camera and the bounce plane of the ball can form any angle, since the necessary transformations for calculating the point of impact in real coordinates will be performed.

Brief Description of the Drawings

[0007] For the purpose of aiding to better understand the features of the invention according to a preferred practical embodiment thereof, a set of drawings has been attached to the following description, in which the following has been depicted with an illustrative character:

Figure 1 is a schematic depiction of the position of the camera used in the method of the invention with respect to playing ground and the line which limits it. Figure 2 is a depiction of the determination of the point of impact, which in turn is the point of intersection of the two lines which represent the positions of the ball in the proximity to the point of impact and after the point of impact. Figures 3a and 3b depict the steps in the calibration of the camera.

Detailed Description of the Invention

[0008] The system of the invention is made up of a camera capable of preferably capturing at least 50 frames per second, means for storing and processing the images, extracting the position of the ball in each frame, means for calculating the velocity of the ball in each moment from the previous data and means for calculating the position of the point of intersection of the lines joining the points which represent the position of the ball in each of the frames.

[0009] The capturing speed of the camera is determined from the velocity of the ball and also from the field of view which is being captured in each moment. If a greater field of view is captured, a lower capturing speed can be obtained.

[0010] For the implementation of the system for detecting the bounce point of the ball (Figure 1), a camera (1) is placed, preferably in the vicinity of the field line (4) in which the position of impact is to be detected. The closer the camera is placed the higher the resolution of the captured images, i.e., a smaller area will be represented with one and the same number of pixels, and therefore the accuracy of the result will increase.

[0011] Said camera must have the capacity to preferably detect at least 50 images per second and must have a configurable shutter speed. Both the image capturing speed and the shutter speed must be defined upon starting up the system. The shutter speed must be configured taking into account the lighting of the scene; the lower the amount of light, the lower the shutter speed and vice versa. Likewise, this parameter is limited by the need for the round shape of the ball to not be deformed when capturing the frame.

[0012] The capturing speed depends on the field of view of the camera (2) and of the maximum velocity of the ball. It must preferably at least 50 images per second in order to be able to define the position of the ball with sufficient accuracy and thus be able to determine the

bounce position with a minimum error, allowing the results to be able to be accepted as valid. The higher the image capturing speed, more information about the trajectory of the ball will be obtained and therefore, the bounce position will be more accurate, but it must be taken into account that the processing of the images will be more expensive and therefore it can depart from the real-time application which is to be performed. Therefore, when choosing the image capturing speed, not only the velocity of the ball must be taken into account, but also the necessary processing time. It is possible to determine beforehand which will be the maximum velocity of said ball and depending on this value to determine the necessary minimum capturing speed.

[0013] Likewise, the field of view of the camera must also be taken into account to determine the number of images which are captured per second. The greater the field of view, the lower the capturing speed the system will require. Even though this parameter must be taken into account, it is not as determining as the velocity of the ball.

[0014] Once the system has been started, the recording begins. The camera is recording the contentious area around the field line (3) and storing the information in a PC or any computer means in which the data will also be processed during the entire duration of the game. When there is a controversial piece of play, the user must indicate to the system that said piece of play has occurred and in that moment the system will begin analysing the images. To that end, it extracts the latest images from the memory, taking into account that the group of said latest images must contain the controversial piece of play. The number of images which must be extracted is defined at the beginning of the game depending on the number of images per second recorded by the camera.

[0015] Once the set of images is extracted, the first image of the sequence, which will be the reference image, is extracted. All the image processing which is done once the images are extracted is performed in what is defined as a region of interest of the image. This region of interest is a fraction of the image which is defined from a movement vector which is determined with the position of the ball in consecutive images. The calculation of said movement vector is summarised below.

[0016] Performing the image processing only in a fraction of the image speeds up the processing and therefore, a real-time operation is achieved.

[0017] After this point, all the references which are made to an image refer to the area of interest defined for each of the images of the sequence.

[0018] All the images of the sequence are converted from a colour space, which is determined by the camera output, into a greyscale space. The entire processing which is applied to the images will be performed in greyscale since the necessary information is the luminance information. For each of the images of the sequence, the difference is calculated pixel by pixel with the previous image (temporally) and with the reference image, such

that two difference images are obtained. These two images are converted into black and white images from a threshold value. The logic operation AND is performed for each pair of images, from which operation only the regions which are present in both images are extracted. The region corresponding to the ball is thus distinguished. The possibility of distinguishing the region of the ball in this manner is due to the fact that the velocity of the ball is considerably greater than the rest of the objects present in the scene.

[0019] Regions known as noise also have to be removed, for which once the logic operation AND is performed, the region of the ball for each of the images of the sequence is selected. Said region is chosen taking into account criteria of area, shape and position of the candidate regions.

[0020] The value of the area is determined by the value of the area of the ball in the previous image, except in the first iteration in which predefined thresholds are considered. In the event of obtaining similar area values for several candidate regions, the position of the region inside the image determines the most probable candidate region for being a ball. And finally, the regions which do not have an elliptical shape are discarded. The region which is considered to be the ball is thus extracted.

[0021] After having detected the ball in the image, the movement vector and the velocity vector of the ball are calculated taking into account the coordinates x'' and y'' in pixels with respect to the previous image and the time that has elapsed between one image and the next. In order to calculate the movement vector the difference between the coordinates (x'' and y'') of the centre of the ball is calculated for consecutive images.

[0022] In order to calculate only the velocity vector, the time concept must be added. This time is determined by the image capturing speed which has been defined in the camera upon starting up the system. In order to determine the velocity, the velocity in coordinate x' and the velocity in coordinate y' are calculated, the total velocity being the modulus of said vector.

$$t_{capt} = \frac{1}{v_{capt}}$$

$$v_x = \frac{x'''_n - x'''_{n-1}}{t_{capt}}$$

$$v_y = \frac{y'''_n - y'''_{n-1}}{t_{capt}}$$

$$v_{TOTAL} = \sqrt{v_x^2 + v_y^2}$$

[0023] The calculation of the movement vector allows speeding up the process for detecting the ball since, once the ball has been detected in the first images and the movement vector is obtained, an approximate region of interest in which to seek the ball can be defined in the following image. To that end, the central point of the ball is taken and, from the movement vector, the position in which the ball will be in the following image is estimated and the region of interest is extracted. It is thus not necessary to process the entire image but rather it is enough to take the region of interest and perform the search only in that region. The size of that region is determined by the size in the image of the ball, the velocity of the ball and the size of the image. The greater the size and the velocity of the ball, the larger the region of interest should be. Likewise, the larger the image the larger the region and vice versa. This region of interest is the region in which all the image processing is performed.

[0024] The process defined so far is repeated for each pair of images of the sequence. As shown in Figure 2, once the velocity vector has been extracted for the entire sequence of images, the sequence of positions of the ball, both in x'' and in y'' , is divided into two segments (5). In order to determine the limit of the segments, the difference in angle and modulus of the velocity vector is taken into account. The maximum value of the angle difference will determine that limit which will divide the points into two segments. If there are similar angle values, the modulus will determine that limit.

[0025] Once the coordinates of the position of the ball have been plotted in two segments, a least-square fitting is performed for each of the two segments. For the calculation of this fitting curve, the points which are above the minimum distance of the curve are discarded reiteratively. The trajectory of the ball for each of the two segments is thus determined. The point of intersection of the resulting curves is considered the bounce point of the ball.

[0026] The position of the bounce point is referenced with respect to the coordinates x'' and y'' of the image and the real distance of the bounce point with respect to the field line is to be known. To that end, the calibration of the camera is performed, which consists of performing a geometric transformation to be able to extract the coordinates x and y in the plane of the field line from the coordinates x'' and y'' in pixels.

[0027] The geometric transformation must be defined upon starting the system and consists of indicating which real coordinates some points of the image identified by the coordinates in pixels have. Thus, once the bounce point is identified, the position in the real plane can be calculated. This process is known as calibration of the camera.

[0028] The calibration consists of calculating a series

of parameters which allow placing the camera (position and tilt) with respect to a reference point (x,y,z) of the real world. A plane transformation matrix (H) and a matrix containing intrinsic parameters of the camera which relate to the measurements in the reference system of the scene (in millimetres) with the position of points in the image expressed in pixels are calculated. From this, a set of matrices (K,R,t) describing the rotation and translation parameters of the camera is extracted, allowing the coordinates (x,y) in the plane of the field line to be extracted.

[0029] The transformation matrix (H) must be defined upon starting the system and indicates with which real coordinates some points of the image of the camera correspond. Thus, once the bounce point is identified in the image, the position in the real plane can be calculated.

[0030] The resulting geometric transformation only works for all those points of the plane $Z=0$. To that end, it is necessary to define the bounce point in the image. The point will be defined by the coordinates (x'',y'') in pixels (see Figure 3b).

$$x'' = x''' + R * \sin(\alpha)$$

$$y'' = y''' + R * \cos(\alpha)$$

[0031] A small error is assumed since the exact point where the ball touches the ground is not shown in the image, since it is covered by the ball itself.

[0032] Once the point (x'',y'') is calculated, the point to the real plane is transformed by multiplying it by the plane transformation matrix (H).

$$(x', y') = H * (x'', y'')$$

$$H = \begin{pmatrix} a_{(1,1)} & a_{(1,2)} & a_{(1,3)} \\ a_{(2,1)} & a_{(2,2)} & a_{(2,3)} \\ a_{(3,1)} & a_{(3,2)} & a_{(3,3)} \end{pmatrix}$$

[0033] As seen in Figure 3b, the point (x',y') is not an exact projection of the centre of the ball, so it is moved in the direction of the optical vector of the camera a distance which depends on the position and tilt of the camera to the central point.

$$x = x' + v_x (K, R, \vec{t})$$

$$y = y' + v_y(K, R, \vec{t})$$

[0034] Once the real position of the bounce point, which is referenced in the field line has been calculated, it can be identified if the ball has bounced outside, inside or on the line itself.

10

Claims

1. Method for determining the point of impact of a ball in a playing field during a controversial piece of play in a sports event which is **characterised by** the following steps:
 - recording the contentious area during the game by means of a single camera
 - extracting the images corresponding to the controversial piece of play
 - selecting the area corresponding to the ball
 - calculating the coordinates of the ball in pixels in each image
 - plotting the points with the coordinates calculated in the previous step
 - determining the point of intersection of the two segments joining the previous points
 - transforming the point of intersection into real coordinates.
2. Method according to claim 1, wherein the shutter speed of the camera is at least 50 images per second.
3. Method according to any of the previous claims, wherein the velocity vector is calculated and used to determine the limit of the segments, by means of the difference in angle and modulus.
4. Method according to any of the previous claims, further comprising a step of calculating the movement vector of the ball for each of the images.
5. Method according to claim 4, including a step in which the region of interest is selected from the movement vector.
6. Method according to any of the previous claims, comprising a step of transforming the colour images to greyscale prior to the extraction of images.

55

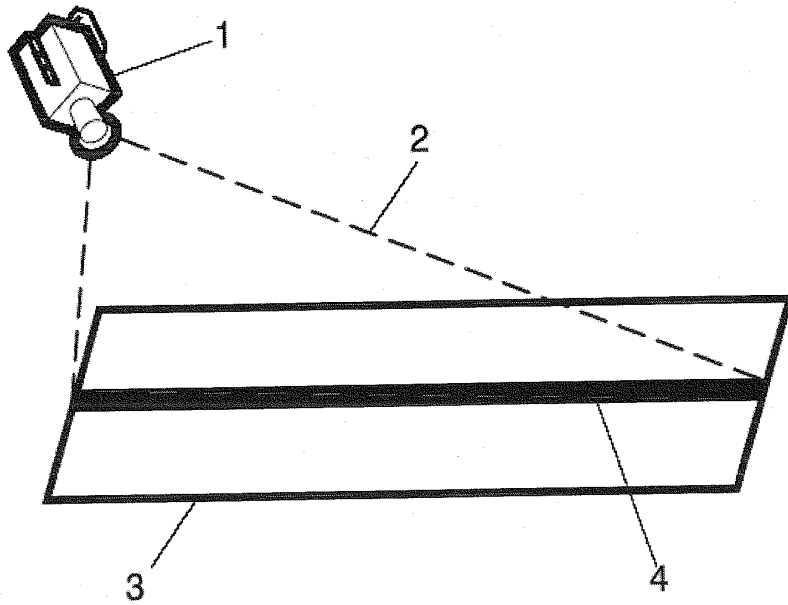


FIG. 1

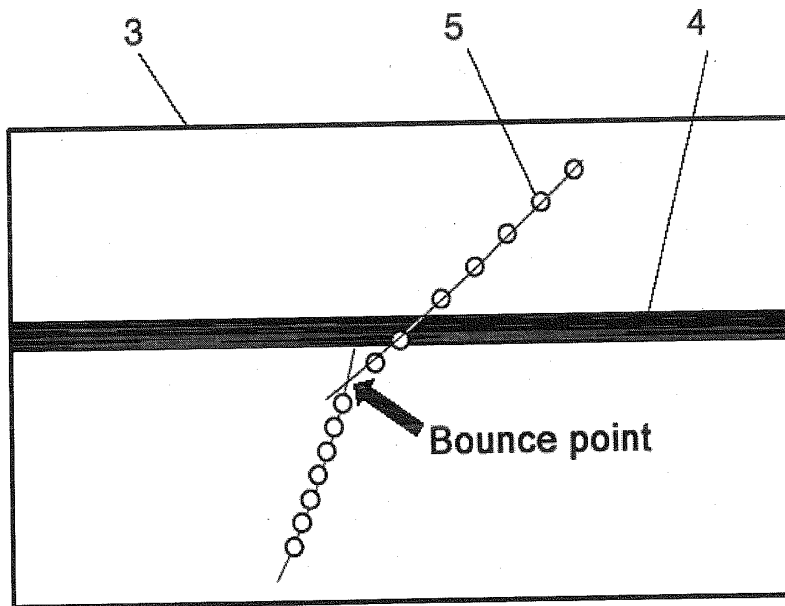


FIG. 2

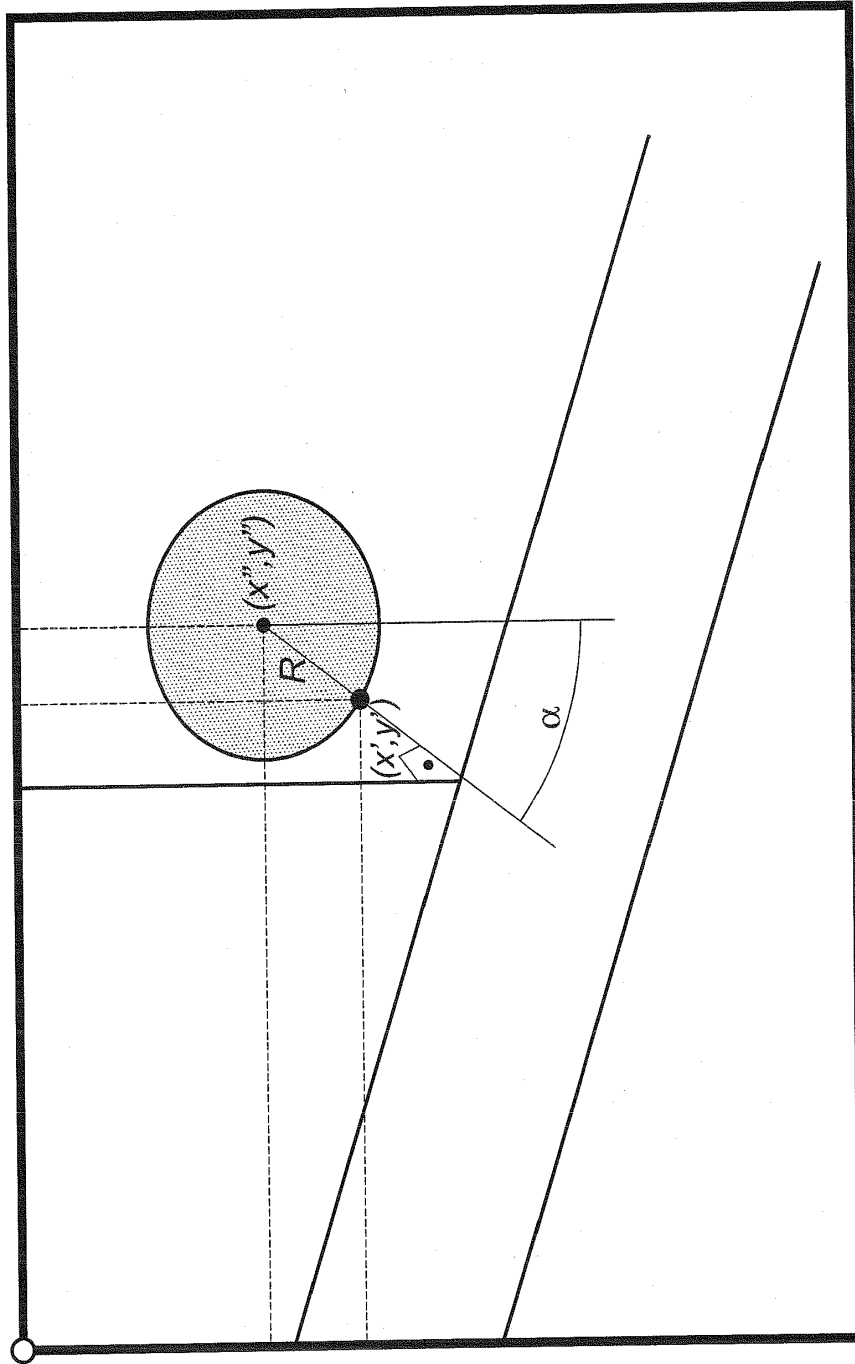


FIG. 3a

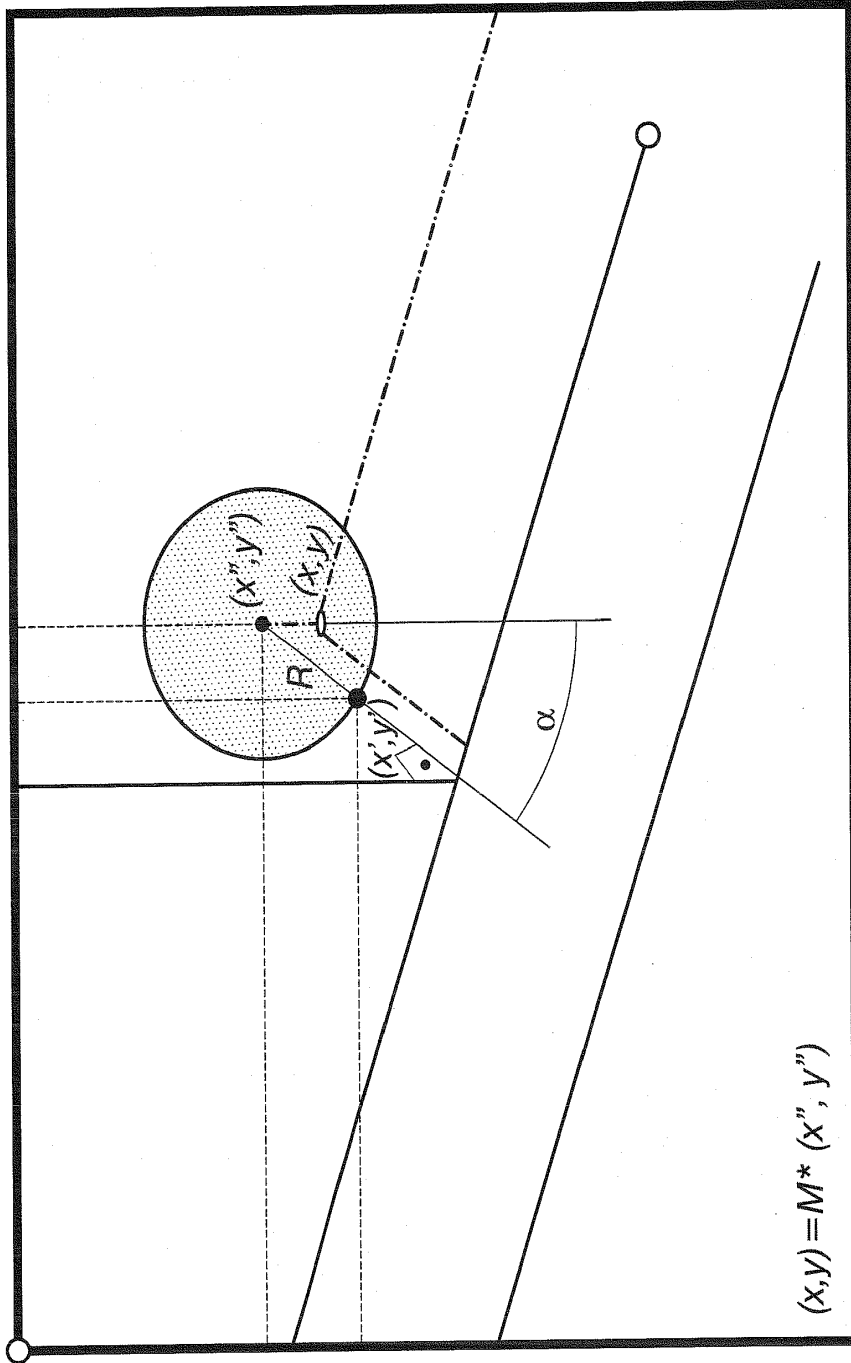


FIG. 3b

2.3.3 Accurate ball trajectory tracking and 3D visualization for computer-assisted sports broadcast

- **Izenburua:** Accurate ball trajectory tracking and 3D visualization for computer-assisted sports broadcast
- **Egileak:** Mikel Labayen, Igor G. Olaizola, Naiara Aginako, Julian Florez
- **Aldizkaria:** Multimedia Tools and Applications (MTAP)
- **Argitaletxea:** Springer
- **Zenbakia (Orrialdeak):** Vol. 73 (1819–1842)
- **Inpaktu-faktorea (urtea):** 1,346 (2014)
- **Kuartila:** Q2
- **Urtea:** 2014
- **DOI:** <http://dx.doi.org/10.1007/s11042-013-1558-x>

Accurate ball trajectory tracking and 3D visualization for computer-assisted sports broadcast

Mikel Labayen · Igor G. Olaizola ·
Naiara Aginako · Julian Florez

Published online: 24 August 2013
© Springer Science+Business Media New York 2013

Abstract The application of computer-aided controversial plays resolution in sport events significantly benefits organizers, referees and audience. Nowadays, especially in ball sports, very accurate technological solutions can be found. The main drawback of these systems is the need of complex and expensive hardware which makes them not affordable for less-known regional/traditional sports events. The lack of competitive systems with reduced hardware/software complexity and requirements motivates this research. Visual Analytics technologies permit system detecting the ball trajectory, solving with precision possible controversial plays. Ball is extracted from the video scene exploiting its shape features and velocity vector properties. Afterwards, its relative position to border line is calculated based on polynomial approximations. In order to enhance user visual experience, real-time rendering technologies are introduced to obtain virtual 3D reconstruction in quasi real-time. Comparing to other set ups, the main contribution of this work lays on the utilization of an unique camera per border line to extract 3D bounce point information. In addition, the system has no camera location/orientation limit, provided that line view is not occluded. Testing of the system has been done in real world scenarios, comparing the system output with referees' judgment. Visual results of the system have been broadcasted during Basque Pelota matches.

M. Labayen · I. G. Olaizola · N. Aginako (✉) · J. Florez
Department of Digital Television and Multimedia Services, Vicomtech - Ik4 Research Alliance,
San Sebastian-Donostia, Spain
e-mail: naginako@vicomtech.org

M. Labayen
e-mail: mlabayen@vicomtech.org

I. G. Olaizola
e-mail: iolaizola@vicomtech.org

J. Florez
e-mail: jflorez@vicomtech.org

Keywords Computer graphics • Camera calibration • Tracking • Segmentation • Sports events broadcast

1 Introduction

Object tracking has a prominent role within the field of computer vision. The proliferation of high performance computers, the availability of high quality video cameras at affordable prices, and the increasing need for automated video analysis has generated a great deal of interest in object tracking algorithms. Detection of target moving objects frame by frame, tracking and analysis to recognize their behavior are the usual pipeline in video analysis [2].

From application domain point of view, tracking systems are being introduced in sport game broadcasts, providing spectators with additional information. Due to the high performance equipment requirements, the renting of this kind of systems is quite expensive, making them unaffordable for small producers or broadcasters. This is exactly the Basque Pelota case. This regional/traditional game is produced by small producers and broadcasted by regional broadcasters. Their low budget does not allow to contract current setups to support controversial plays.

In this work, a system to assist referees solving controversial plays in sport games is described. The game has to be played with a ball and its playground must be delimited by lines. The developed software allows to reduce the set-up requirements, creating an accurate system that is affordable for a wider range of clients.

The set-up design, which is able to cover all border lines, is a challenge. This border line number can be high (e.g. tennis playground) driving the solution to multi-camera set-up. The modularity and scalability are important approaches for required solution. In this work The Basque Pelota test-case is presented in order to simplify the explanation. It is an ideal scenario to test the system first prototype because of its technical peculiarities. Since the playground is delimited by walls on 3 of its borders, it has only one border line to be covered. Once the application is validated in this game, this technology is being extended to other sports (i.e. tennis) which need a multi-camera distribution. For each camera image capture, image analysis and real-time rendering modules are reusable in this new modular and scalable set-up.

In the following Section 1.1, this article carries out a short analysis of the state of the art in controversial play resolution. Afterwards, in Section 2, a system overview is presented in terms of its objectives, description and specifications. Section 3 details the hardware and software (HW/SW) implementation of the core system, including camera calibration, image analysis and real time virtual 3D reconstruction processes. Finally, in Section 4 the document shows the results obtained from tests carried out in real scenarios and it ends up with Section 5 summarizing the conclusions.

1.1 Related works

In its simplest form, tracking can be defined as the problem of estimating the trajectory of an object in the image plane as it moves around a scene. In other words, a tracker assigns consistent labels to the tracked objects in the different frames of a video. Additionally, depending on the tracking domain, a tracker can also

provide object-centric information, such as orientation, area, or shape of an object. Therefore, the use of object tracking is pertinent in the tasks of [2]:

- Motion-based recognition, that is, human identification based on gait, automatic object detection, etc.
- Automated surveillance, that is, monitoring a scene to detect suspicious activities or unlikely events.
- Video indexing, that is, automatic annotation and retrieval of the videos in multimedia databases.
- Human-computer interaction, that is, gesture recognition, eye gaze tracking for data input to computers, etc.
- Traffic monitoring, that is, real-time gathering of traffic statistics to direct traffic flow.
- Vehicle navigation, that is, video-based path planning and obstacle avoidance capabilities.

This work focuses on motion-based object recognition in sport broadcasting. Tracking systems in the TV broadcast domain are not a recent approach at all. Most of the researched systems in this field are based on prediction algorithms based on Kalman [4, 17] or particle filters [10]. Extend state of the art material is available about methodologies dedicated e.g. to player and ball tracking in soccer [14, 18, 19] or tennis [5, 7, 13].

Some companies such as *Sportvision*¹ and *Virtual eye*² market systems which provide data content and enhancements for sports broadcasts and applications:

- The FoxTrax hockey puck tracking system [3] based on an infrared sensor. The circuit board inside a puck contained a shock sensor and infrared emitters. The puck emitted infrared pulses that were detected by both the 20 pulse detectors and the 10 modified IR cameras that were located in the rafters. Each IR camera processes the video locally and transmits the coordinates of candidate targets to the “Puck Truck”.
- *Strick zone* control by ball and player tracking. Three PCs connected to three video cameras track a pitched baseball’s flight toward the strike zone. Two cameras observe the baseball, while the third observes the batter to provide proper sizing for the strike zone.
- Playground lines drawings of *IST & TEN*³ in American football. This application uses a number of cameras shooting the field. Recent implementations require around four computers, one computer per camera plus a shared computer for chroma-keying and other tasks that can be run by a single operator.
- Cricket⁴ and golf ball tracking. Based on image computer graphics technology, 4 high-speed cameras (250 fps), two Infrared cameras and sophisticated computer rack are used to track the cricket ball. This set up needs at least a group of 4 operators to its management.
- Additional information for viewers as graphics and statistics in golf.

¹www.sportvision.com

²<http://virtualeye.tv/>

³http://www.ieeeahn.org/wiki/index.php/The_Making_of_Football%27s_Yellow_First-and-Ten_Line

⁴<https://www.youtube.com/watch?v=LjLe06H7EJg>

Due to their closed system, the algorithms on which they are based are in most of cases unknown.

*Hawk-Eye*⁵ markets the most important controversial play image-based analysis and 3D virtual replay reconstruction approach for situations in which a tennis ball sized object is used in the play. Although, it started as cricket ball tracker, it is well known because it is able to point the location of a ball bounce in a tennis court with high accuracy. *Hawk-Eye* uses 6 high speed specialized vision processing cameras which are positioned around the ground and calibrated. In addition the system uses two broadcast cameras and calibrates them so that the graphic is always overlaid in the right place. All cameras have anti-wobble software to deal with camera movement. According to information in its web [9], it is able to deliver a pinpoint accuracy of under 5 mm.

However, the complexity of the set-up, high-speed cameras are needed, and the equipment requirements make the system too expensive for less-known regional/traditional sports. Even more, it cost is around \$60.000 for one court which increases by 100 the cost of the system presented in this approach.

All these approaches use at least two or more high speed specialized vision processing cameras to determine the bounce distance from border line. In addition, they need operator team to control them. In order to reduce the existing solutions requirements, this work presents an alternative set-up, robust in terms of different possible camera operating location/orientation, based on unique broadcast-type camera per border line. This solution can be managed by one operator, even playground has more than one line under control, changing camera views from the system. The challenge in reduction of hardware complexity and achieving market solutions' accuracy motivates this work.

2 System overview

The industrial project called *Begira*, in which this research has been carried out, establishes the technical specifications that the developed system has to fulfil. Although the state of the art can offer specific solutions for some of the technical requirements, it cannot afford the consecution of all the technical specifications. Even more, the economical limitations are a also the variables that constrains this research.

2.1 System objectives

The system must be able to pinpoint accurately the distance of ball bounce from the line. The output 3D virtual video will simulate the last ball trajectory and will be inserted in TV PAL broadcast signal. The solution must be deployed on top of a simple HW/SW system to make it affordable for any producer, sport event broadcaster or less-known sport event organizer.

A system set-up design driven by flexibility in terms of size and operating location, can significantly reduce the costs rising this challenge as a major aim. In addition,

⁵www.hawkeyeinnovations.co.uk

the system needs to operate in quasi real-time, at least faster than the estimated time for video replay which is about 30 sec. Respect to pinpoint accuracy, defined by Basque Pelota referee committee, the estimation error should be under 1 cm for all cases and under 5 mm for 80 % of them. This error has been set taking into account the typical human eye incertitude in the appreciation of bounce point (from a distance and with millisecond duration), which is also subjective. As this limit was considered achievable after the demonstration of our first version of the system, it was determined as a requirement. Referee committee is aware of the difficulty of approaching these accuracy values, but consider them necessary in order to standardize the system.

2.2 System description an requirements

In this section, the system general workflow, as well as module specifications and functionalities are described (Fig. 1).

The prototype has four independent modules: a) the camera, b) the capturer/encoder, c) a laptop for storage and all processing tasks and d) a video adapter for TV PAL broadcast or playout. The camera captures images and transfers them in RAW format using an standard professional TV interface to the capturer/encoder module. This second module encodes frames using the H.264 codec and transfers it to the laptop where it stores them into a Transport Stream (.ts) video file. The image processing and later 3D virtual replay generation tasks are executed in the laptop. In the last module, the 3D virtual representation output video is adapted to broadcast quality video .

- **Camera** The cameras used for image capture must fulfil some characteristics in order to make the ball detection easier for segmentation and identification operations carried out in the next steps:
 - **Frame rate** The broadcast camera must provide enough images per second to track and predict the ball trajectory in each frame. The choice of this factor is defined taking into account the trade-off between the data processing time (in capturing/coding/storing) and the necessity of the amount of real images to be able to approach accurately the ball trajectory. To avoid missing frames, the time elapsed in recording/storing each frame must be less than reciprocal of the frame rate.
 - **Shutter speed and diaphragm aperture** In the case under study, the accuracy of the tracker can be improved if a target with stable shape, without blur effect, and with stable color (grayscale intensity) is acquired. Therefore, a high shutter speed camera is required. The choice of this factor is to be

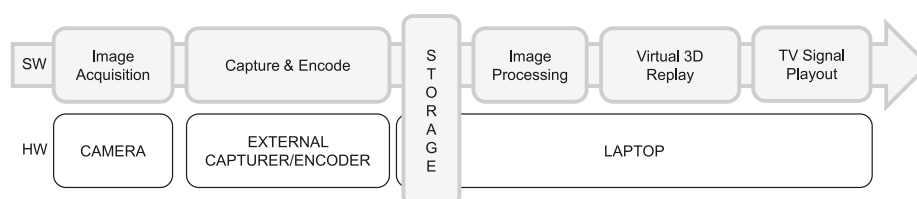


Fig. 1 System HW/SW workflow & modules

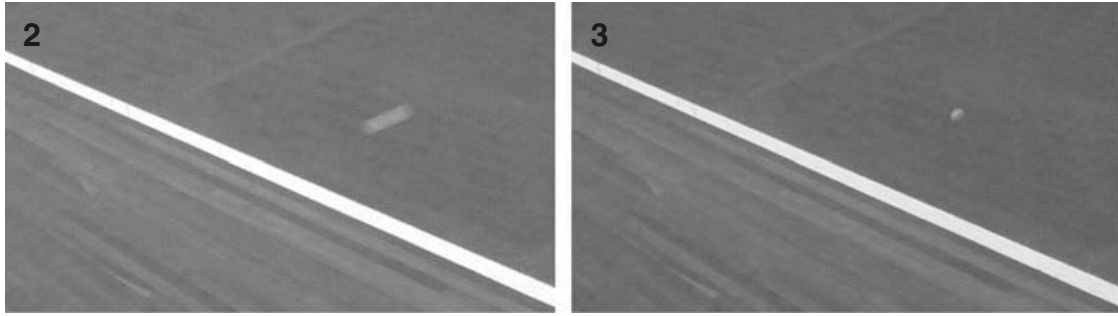


Fig. 2 Low shutter speed (auto). Electronic gain disabled. Blurry ball

Fig. 3 High shutter speed (1/500). Electronic gain enabled. Blur-free ball

defined taking into account the trade-off between the minimum illumination required in the segmentation process and the necessity to keep the shape of the ball stable. It will be set to the minimum that allows the ball to appear as a clear round object. The maximum speed of the ball will be relevant on it (see Figs. 2 and 3).

The minimum shutter speed and consequently the diaphragm aperture are set according to the illumination of the sport events place and to the expectable maximum ball speed in each game. In this work, these values are set for Basque Pelota courts (indoor, illuminated for TV broadcast).

- **Image resolution** The accuracy in ball bounce pinpointing is also related to the resolution of the captured images. The higher the resolution, the lower the pixels/distance ratio. Once again, the system performance is based on a trade off between lower processing time and higher accuracy in measure (see Table 1).

To approximately calculate Pixel/Distance ratios, we use as reference object the border line. Calculating the amount of pixels in the horizontal and vertical vertices of the image, we can compute border line width pixel amount and compare it to border line real width measurement.

- **Color space** The color space influences the ball segmentation process. Although multi-component color spaces can offer extra information in image object understanding, the bright white color of the ball and the dark color of the playground, offer high contrast which makes single component color spaces enough for segmentation purposes. This reduces the generated data amount for capture, encoding, storage and processing tasks.
- **Optical lens with fixed camera-to-playing field distance** The field of view of the camera must also be taken into account to determine lens distortion

Table 1 Pixel/Distance (pixel/cm) ratios calculated for different image resolutions and camera distance

Image resolutions	Distance from camera (m)		Reference image
	9 m	0.5 m	
1080p	0.45	0.065	
720p	0.7	0.09	
576p	0.85	0.15	

and system precision in terms of *image pixel/real distance*. The greater the field of view, the greater the covered scene area in which the ball trajectory can be analyzed. However, the greater the field of view, the greater the lens distortion and the lower the precision (pixel/distance). Even though this parameter must be taken into account, it is not as critical as others like velocity of the ball.

- **Capture & encoding module** The capturer/encoder module captures the RAW multi component video signal provided by the camera. After that, the signal is converted to a single component color space. Then, the signal is compressed and encoded in order to reduce the information data flow for the storage and processing tasks.
 - **Codec** The codec requirements have to solve the controversial relation between image quality and compression ratio. The goal is to obtain the maximum compression ratio, keeping the minimum image quality which ensures correct segmentation conditions after decoding.
 - **File container** The video file must be read and written at the same time. In addition, the read process must offer quasi random access capabilities for the retrieval of part of the whole recorded video starting from a specific frame. Moreover, most of multimedia container formats include timestamps and data just before file closing, becoming navigation more difficult. The chosen container must solve this problem.
- **Storage and processing laptop** The laptop and capture/encoding module are connected through USB 2.0. The laptop stores the encoded images in its hard disc, it retrieves and analyzes them and finally generates a 3D virtual replay of the action. A multi-tasking approach for quasi real-time performance establishes the hardware characteristics of the laptop. The system core software is stored and executed in this module. The algorithm robustness is directly related to the system set-up flexibility.
- **Video adapter to TV broadcast signal** The broadcasted output video signal must comply with the broadcaster graphical requirements and signal quality specifications at its mobile units. This module adapts the rendered video signal into a TV broadcast signal.

3 Implementation

In the first step of the implementation, state of the art and market study has been carried out to identify the existing HW/SW developments which best fit the needs of the system based on the requirements outlined above. Two main issues have been encountered at this point: on one hand, no specific HW/SW solution exists for the established requirements. On the other, available HW solutions deal with independent tasks identified in the system workflow & modules figure (Fig. 1). This context pushes the development of our own algorithms, as the outcome of a research process. The unique existing Open Source algorithms used in the implementation are Camera Calibration (OpenCV) and Polynomial Approximation (GNU).

The system as a whole has been integrated using Qt:⁶ a cross-platform application framework that is widely used for developing application software with graphical user interfaces (GUIs).

3.1 Image capture

From the beginning, this system was developed using conventional TV broadcaster equipment in order to reduce costs in later market adoption processes. The camera used in the tested prototype is a common professional HD handheld camera (Panasonic HVX200A⁷), widely used by many kinds of producers/broadcasters. It provides a RAW YUV(4:2:2) component signal at a maximum resolution of FULL-HD 1080i and a maximum frame rate of 50fps far away from the throughput and features of cameras required by other market solutions. The camera is set-up at HD 720p 50fps both providing enough resolution and frame rate for our approach.

The scene illumination conditions are then to be analyzed. The amount of available light is a combination of pelota court lights and of additional spotlights used in special competition broadcasting. Under these conditions, the scene often is not enough illuminated, providing resulting images (at 50fps and 1/500 shutter speed) which are low-contrast.

Taking into account the minimum illumination required to keep the color and constant shape characteristics of the ball in the segmentation process, the balanced compromise between shutter speed, which keeps constant the ball round shape, and diaphragm aperture and electronic light gain, which keep the scene contrast, is defined for each broadcasted event.

3.2 Image encoding and storage

The amount of data generated capturing HD 720p images at 50fps and described by the bit rate BR parameter makes necessary the use of compress/encode algorithms.

Resolution: (1280 * 720) pixels/frame

Frame Rate: 50 frames/sec

Bit Depth: 8 bits/pixel

Components: (1 (Y) + 0.5 (U) + 0.5 (V))

(Note: YUV 4:2:2 format)

$$BR = 1280 \times 720 \times 50 \times 8 \times 2 = 703.125 Gbps \quad (1)$$

The generated throughput would impose special storage, transmission bandwidth and equipment. This makes the system set-up more expensive and less compact. However, reduced system cost and dimensionality are central requirements from the beginning: to reduce the throughput the signal must be compressed. The dominant video codec today for web and mobile video (limited by the transmission channel bandwidth) is H.264 [6, 15]. H.264 compression preserves the video quality at high compression ratio better than other popular codecs widely available on the market [15, 16].

⁶www.qt.nokia.com

⁷www.panasonic.com/business/provideo/home.asp

Although the standard defines 17 sets of profiles, H.264 has three commonly-used: Baseline (lowest), Main, and High. Higher profiles (Main and High profiles) ensure the best signal quality-compression relation. Since the system needs high compression ratios with the best signal quality, the High profile is chosen.

H.264 is typically deployed into .MP4 file containers. However, a wide range of different containers can be used. One of the main difficulties of working with open videos is the random access within the content. Most seek function implementations require closed video files to function properly. However, in our case, the positioning at specific frame is performed while the video file is open and the encoder is appending information on it. To achieve this purpose, it is necessary to have time marks periodically embedded in the video file. Nevertheless, most video containers only include those marks just before the file is closed.

The container chosen to fit our requirements is therefore MPEG Transport Stream (.Ts) [12]. This Transport stream, devoted to content broadcasting, specifies a container format encapsulating packetized elementary streams, with error correction and stream synchronization features for maintaining transmission integrity when the signal is degraded. This allows to read specific video segments while writing into the same file.

The open source ffmpeg⁸ library has been used to compress, encode and encapsulate the video, as well as to retrieve video sections, and decode them. This package includes audio/video codec and audio/video container multiplexer and demultiplexer libraries.

3.3 Camera calibration

Camera calibration or resectioning is the process of finding the true parameters of the camera that produced a given photograph or video based on prior knowledge of the scene. The camera parameters are classified in extrinsic and intrinsic parameters.

Rotation and translation matrices (R, \vec{t}) contain the extrinsic parameters which denote the coordinate system transformations from 3D world coordinates to camera coordinates. On the other hand, the intrinsic parameter matrix (K) encompasses focal length, image format, and principal point (2).

$$K = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \quad (2)$$

Where,

f_x & f_y *Lens focal length*
 c_x & c_y *Principal point (the image center)*

The camera calibration is carried out using the so-called pinhole camera model, on which Opencv⁹ camera calibration routines are based. A scene view is formed

⁸www.ffmpeg.org

⁹<http://opencv.willowgarage.com/wiki/>

by projecting 3D points (x_p, y_p, z_p) into the image plane (x_i, y_i) using a perspective transformation.

$$P_i = \begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix}, \quad P_p = \begin{pmatrix} x_p \\ y_p \\ z_p \end{pmatrix} \quad (3)$$

From the homography matrix (H), the matrices (R, \vec{t}) describing the rotation and translation parameters of the camera can be extracted.

In the mathematical development below the captured image points are identified by P_i (image coordinate, pixel) and playground plane points, where the ball will bounce, by P_p (real world plane coordinate, cm). P_p^* is an auxiliary point (real world plane coordinate, cm).

Since OpenCV use homogenous coordinates:

$$\begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix} = K[R|t] \begin{pmatrix} x_p \\ y_p \\ z_p \\ 1 \end{pmatrix} \quad (4)$$

Where,

(x_p, y_p, z_p) Real world 3D coordinates

(x_i, y_i) Projection point coordinates

$$P_p^* = \begin{pmatrix} x_p^* \\ y_p^* \\ z_p^* \end{pmatrix} \quad (5)$$

$$H_{ip} = \begin{pmatrix} h_{(1,1)} & h_{(1,2)} & h_{(1,3)} \\ h_{(2,1)} & h_{(2,2)} & h_{(2,3)} \\ h_{(3,1)} & h_{(3,2)} & h_{(3,3)} \end{pmatrix} \quad (6)$$

$$P_p^* = H_{ip} \times P_i \quad (7)$$

$$H_{ip} = H_{pi}^{-1} \quad (8)$$

$$P_p = P_p^* / z_p^* = \begin{pmatrix} x_p^* / z_p^* \\ y_p^* / z_p^* \\ z_p^* / z_p^* \end{pmatrix} = \begin{pmatrix} x_p \\ y_p \\ 1 \end{pmatrix} \quad (9)$$

The homography matrix (H) needs to be calculated upon starting the system: it maps which pixels coordinates of captured image points P_i correspond with playground plane coordinate points P_p . Thus, once the bounce point in the captured image is identified, the position in the playground plane can be calculated.

The image points are selected using a calibration checkerboard as in Fig. 4. The playground plane points are predefined and they must correspond to the points of the checkerboard which are selected in the captured image. With this process, camera intrinsic parameters matrix (K) and plane homography matrix (H) are calculated. Consequently, the (R, \vec{t}) matrices are defined.

Fig. 4 Calibration checkerboard



The calibration information allows placing the camera (position and tilt) with respect to a reference point (x_p, y_p, z_p) of the pelota court world and determining its intrinsic distortion parameters. Undistort parameters and the geometric transformation, which establishes the relation between a captured image and the playground plane points parameters, are therefore established. Camera calibration makes the system robust in terms of different possible camera operating location. As a result, the system has no camera location/orientation limit, provided that line view is not occluded.

3.4 Image analysis and data processing

In this section, the image processing and accurate bounce point determination algorithms are explained. The development has been based on the open source OpenCV and GSL—GNU¹⁰ libraries.

Once the system has been started, the recording begins. The camera is acquiring the contentious area around the playground border line and storing the information in a laptop where data is also processed during the entire duration of the game. The data captured from the camera is stored as MPEG transport stream (.ts) and using H264 encoding. When a controversial play occurs, the operator triggers the system. To that end, it extracts the latest frames, which contain the controversial play.

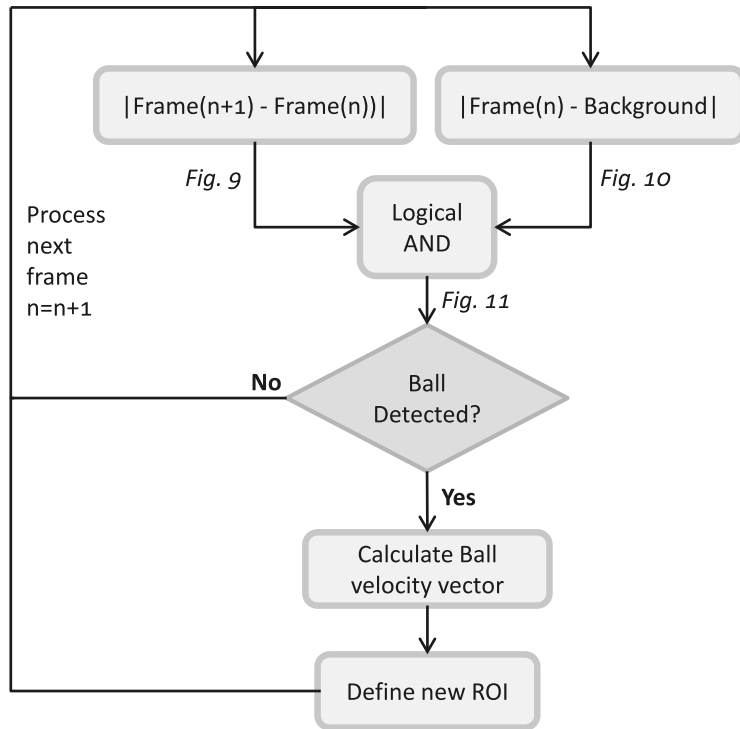
Once the set of images is extracted (Fig. 5), the first image of the sequence is set as the background image (Fig. 6). All the images of the sequence are converted from the color space, which is determined by the camera output, into a single component color space able to contrast the shape, movement and intensity descriptors to pinpoint the ball position in each frame.

After this, all the images are pre-processed using the camera intrinsic parameters matrix (K) to correct the distortion introduced by the optical lens.

After image preprocessing, the process for ball segmentation and tracking starts (Fig. 5) for each of the corrected images of the sequence (Fig. 7). Due to the knowledge of the probable initial ball position, this process is only applied for a

¹⁰www.gnu.org/software/gsl/

Fig. 5 Ball segmentation and tracking process



concrete image area. Image areas' difference is calculated pixel by pixel with respect to the previous image (temporally) (Fig. 8) and to the reference image (Fig. 6), such that two difference images (Figs. 9 and 10 respectively) are obtained. Broadcast camera position and orientation are statics from the beginning of the match. For this reason, the frame difference technique provides a background-free output.

These two subtraction image areas are transformed into black and white image areas via thresholding. The logic operation AND is performed for each pair of image areas (Fig. 11), so that only regions which are present in both images are extracted. Regions identified as noise also have to be removed by the logical AND composition operation. In order to discard noisy regions, estimated shape and area are used. Furthermore, velocity of the ball, considerably greater than that of rest of the objects present in the scene, is set as key characteristic for segmentation. This methodology is used for extracting the initial position of the ball. Once the initial position is determined, the tracking process of the ball is performed.

The tracking process is based on the calculation of a movement vector. This movement vector and the velocity vector of the ball are calculated taking into account its coordinates (x_i''', y_i''') in pixels with respect to the previous image and to the time that has elapsed between one image and the next. In order to calculate the movement vector, the difference between the coordinates (x_i''', y_i''') of the center of the ball is calculated for consecutive images.

The calculation of the movement vector allows predefining a ROI where the segmentation process occurs. Once the initial point of the ball has been extracted and the movement vector calculated, the system creates a ROI determining the prediction area for ball position. All the process steps of frames subtraction and AND logical operation will be made in the extracted ROI. Therefore, the process of ball detection speeds up.

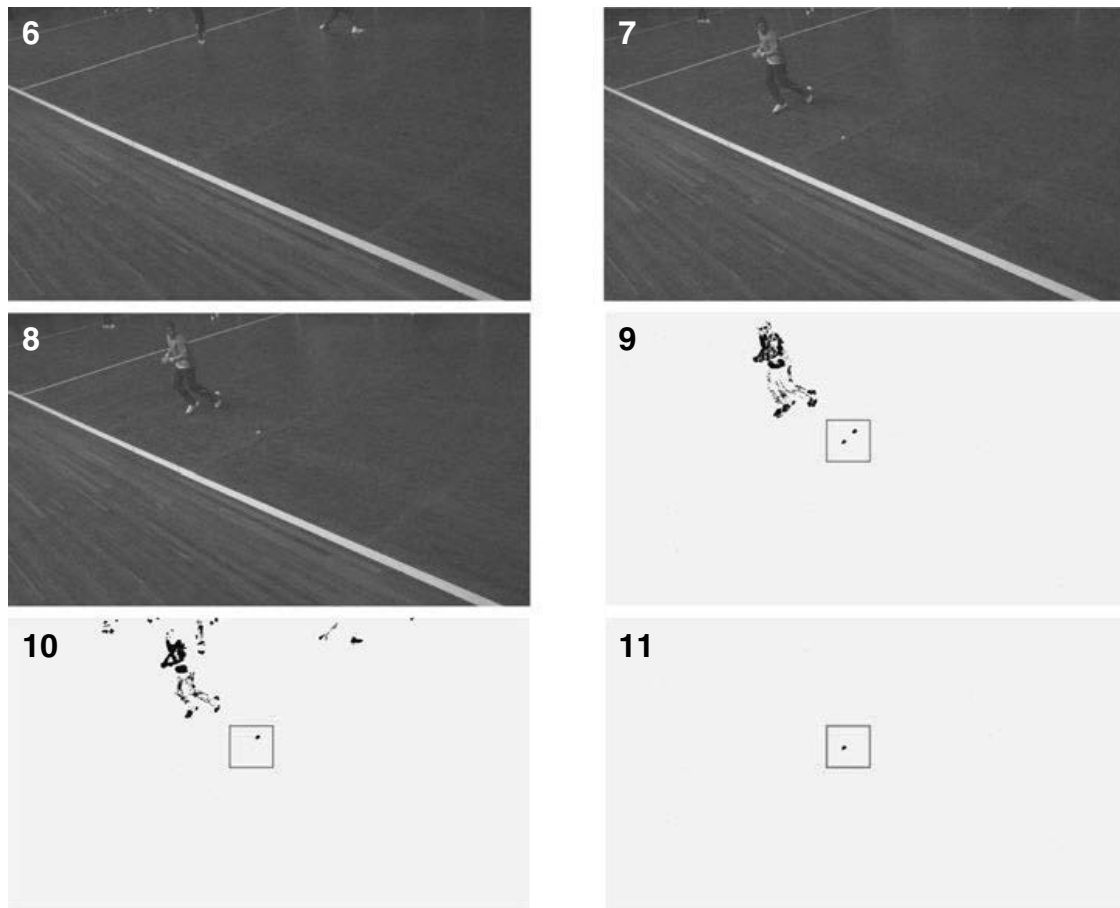


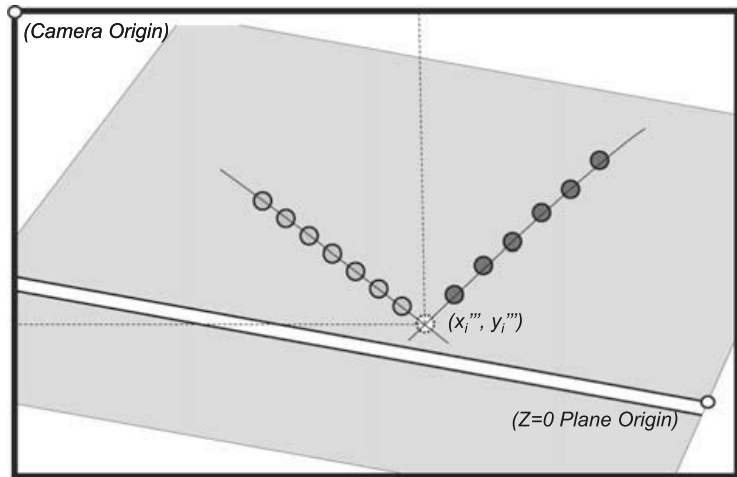
Fig. 6 Background reference image
Fig. 7 Frame(n+1)
Fig. 8 Extracted frame(n)
Fig. 9 $|\text{Frame}(n+1) - \text{Frame}(n)|$
Fig. 10 $|\text{Frame}(n) - \text{Reference background}|$
Fig. 11 Logical AND of (d) on (e) and resulting ROI

The combination of camera position/orientation and selected ROI size keep usually the players belonging regions (noise) out of ROI. However, if there are more than one ball candidate region after AND composition operation, elliptic shape, calculated area and predicted position are used to discard irrelevant ones.

Tracking prediction algorithms, like Kalman Filter or Particle Filters, have been implemented and tested but finally rejected because they do not offer any significant improvements comparing with less complex procedure assuming some approximation. This is due to the fact that ball trajectory can be considered quasi linear close to bounce point. In addition, the relation between capture frame rate and ball velocity makes vector module almost constant and smaller than ROI size. For this reason, the velocity vector information is enough to predict properly the future ROI position and ROI size to detect the ball even if changes its direction after bounce.

As shown in Fig. 12, once the velocity vector has been extracted for the entire sequence of frames, the sequence of positions of the ball (x_i''', y_i''') is divided into two segments. In order to define the limit of the segments, the difference in angle and modulus of the velocity vector is taken into account. The maximum value of the

Fig. 12 Point sequence split & polynomial approximation



angle difference determines the limit which divides the two segments. If the angle values are similar, the modulus is used to break the deadlock.

Once the coordinates of the ball position of the ball have been determined for the two segments, a least-square fitting is performed for each of the two segments [1]. For the calculation of this fitting curve, the points which are above a minimum distance to the curve are iteratively discarded.

if,

$$|x_i'''(n) - lsf(x_i'''(n))| > \sum_{n=1}^{length} \frac{|x_i'''(n) - lsf(x_i'''(n))|}{length} \Rightarrow (x_i'''(n), y_i'''(n)) \text{ point discarded.} \tag{10}$$

Where,

lsf *Least Square Fitted function*
length *Each segment length*

The trajectory of the ball for each of the two segments is thus determined. The point of the intersection of the resulting curves is considered the bounce point of the ball (Fig. 13).

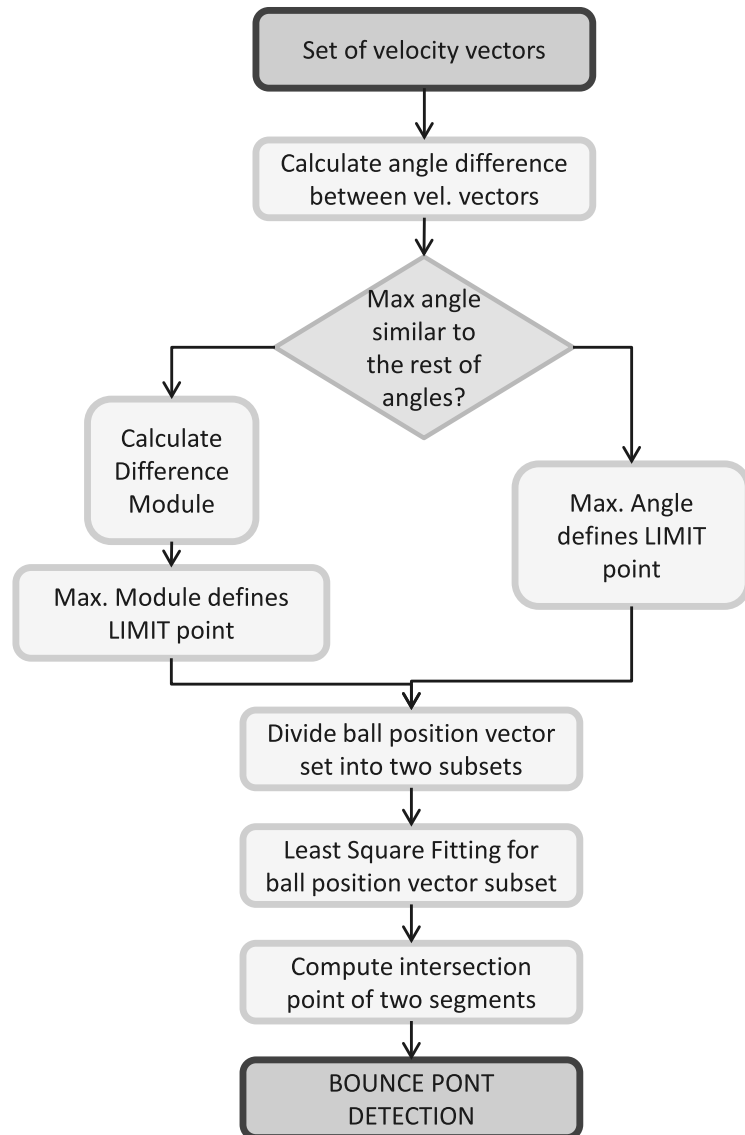
The position of the bounce point is now referenced with respect to the image coordinates (x_i''', y_i''') while the real distance of the bounce point to the playground border line is to be known. To that end, the geometric transformation obtained in the calibration procedure is applied to extract the coordinates (x_p, y_p) in the playground plane from the coordinates of captured image in pixels.

This geometric transformation, related to homography, can only be applied for the points of the playground plane ($z_p = 0$). To that end, it is necessary to define the bounce point in the image. The point is defined by the coordinates (x'', y'') in pixels (Figs. 14 and 15).

$$x_i'' = x_i''' + R * \sin(\alpha) \tag{11}$$

$$y_i'' = y_i''' + R * \cos(\alpha) \tag{12}$$

Fig. 13 Bounce point detection using velocity vectors



A small error is produced at the exact point where the ball touches the ground It has not been reflected in the figures, since the ball is superimposed.

$$x_p = x'_p + v_x(R, \vec{t}) = x_{p_{bounce}} \tag{13}$$

$$y_p = y'_p + v_y(R, \vec{t}) = y_{p_{bounce}} \tag{14}$$

Where,

$(x_{p_{bounce}}, y_{p_{bounce}})$ Bounce point at playground

$$z_p = \begin{cases} \left(\frac{z_{p_0}}{x_{p_{bounce}}} + \frac{g}{v_x^2} * (x_p - (2 * (x_p - x_{p_{bounce}}))) \right) * (x_p - x_{p_{bounce}}) & \text{if } x_p < x_{pb} \\ \frac{g}{v_x^2} * (x_{p_{bounce}}^2 - x_p^2) & \text{otherwise} \end{cases} \tag{15}$$

Where,

g Gravity constant

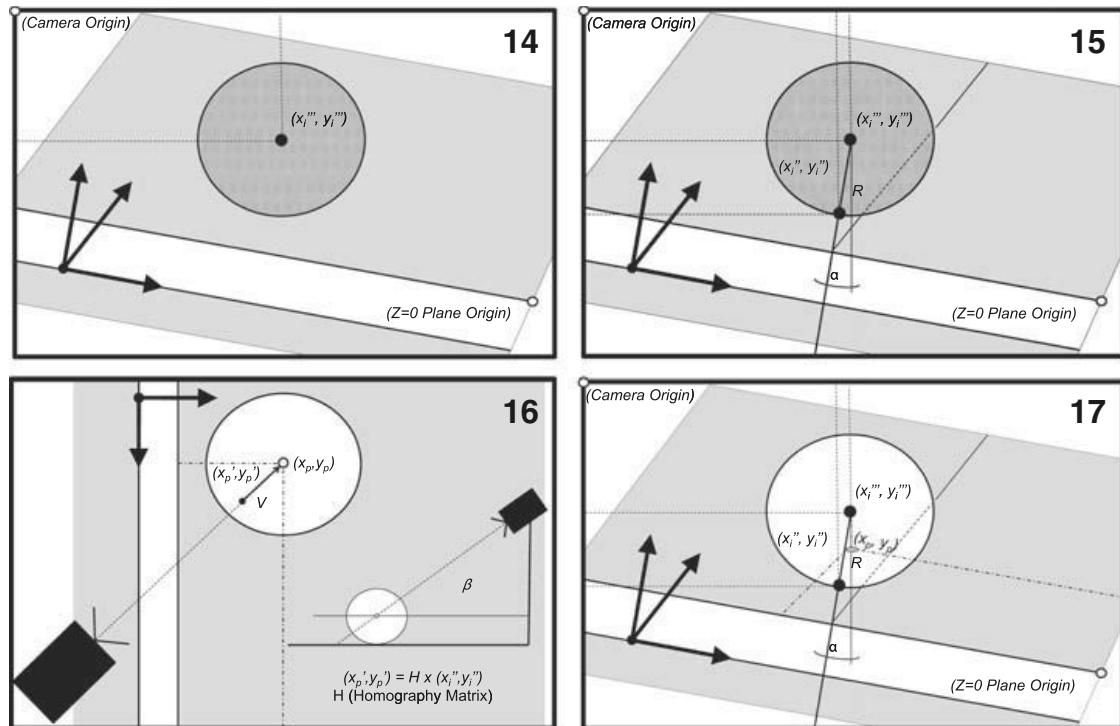


Fig. 14 The center of the ball (x_i''', y_i''') referenced to captured image origin

Fig. 15 The point (x_i'', y_i'') where the ball touch the ground referenced to image origin

Fig. 16 Transformed touch point (x_p', y_p') at ground ($z_p = 0$), referenced to ground plane origin

Fig. 17 Bounce point $(x_{p_{bounce}}, y_{p_{bounce}})$ referenced to ground plane origin

Once the point (x_i'', y_i'') is calculated, the point on the real ground plane is obtained by multiplying it by the plane transformation matrix (H) (Fig. 16 and Eq. 7).

As seen in Fig. 16, the point (x_p', y_p') is not an exact projection of the center of the ball, so it is moved in the direction of the optical vector of the camera with a distance which depends on the position and tilt of the camera to the central point.

Once the real position of the bounce point, which is referenced to the field line, has been calculated it can be determined whether the ball has bounced outside, inside or on the line itself.

Since the correct geometric transformation provided by the two plane homography only can determine the relation between captured images and playground plane points, the only actual ball 3D positioning can be carried out when it touches the ground (at bounce point). From this data, the rest of replay ball trajectory is simulated. Its (x_p, y_p) components are computed from the bounce point $(x_{p_{bounce}}, y_{p_{bounce}}, z_{p_{bounce}})$ and ball direction vectors defined taking into account some ball positions parameters in the processed images close to the bounce moment. The (z_p) component (15) is based on parabolic model taking into account (x_p, y_p) points, approximate ball velocity vector and the z_{p_0} determined from the ball position at the first analyzed image frame (Fig. 17).

3.5 Virtual 3D replay

The visual result of the image analysis is the controversial play virtual 3D replay. Here one of the most performance demanding issues is the rendering engine,

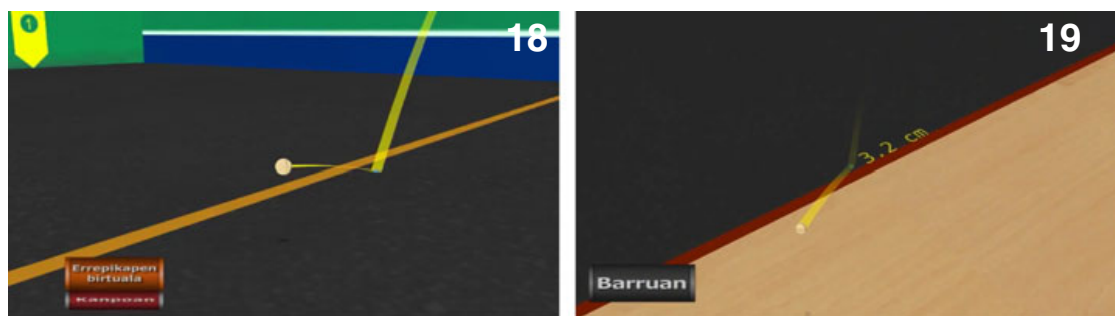


Fig. 18 3D virtual reconstruction of ball trajectory

Fig. 19 Virtual representation of bounce point

dedicated to the computational process of generating an image using 3D information. Firstly, this 3D shape information is converted into polygons and then into triangles. Secondly, these triangles are projected into a 2D image and, finally, each pixel inside the triangles is colored. The whole process takes too much time if no additional strategies or algorithms are used and live TV broadcast cannot be interrupted.

In order to address realtime 3D rendering, the approach is built on top of OpenSceneGraph¹¹ (OSG) library [11]. It is an open source, cross-platform graphics toolkit for the development of high performance graphic applications. It is based on the concept of a scene graph and uses OpenGL.¹²

OpenSceneGraph makes use of techniques that speed up the rendering computational process because the rendering motor deals with considerably reduced information: a Level of Detail (LOD) algorithm, culling techniques (frustum, occlusion and small feature culling) and a State Sorting strategy are employed to this end.

The basic LOD idea is to use simpler versions of an object as it makes less and less of a contribution to the rendered image. So, when an object is far away, less polygons will be used to define it, which reduces the number of triangles to be processed in the rendering. The criteria OSG uses to select a level of detail model depends on the distance of the object from eye point (range-based selection). And to stop the switching form one LOS to another being noticeable, a Continuous Level of Detail (CLOD) technique is used [8].

Culling techniques consist of removing portions of the scene that are not considered to contribute to the final image. The rest of the polygons are sent through the rendering pipeline. With the View frustum culling technique, all the polygon groups that are outside (the region of space in the modeled world visible form the eye point) are eliminated. When occlusion culling is used, all the objects hidden by groups of other objects are also eliminated from the sending-to-render process. And with Small Feature culling, small details that contribute little or nothing to the rendered images are not processed when the viewer is in motion [8].

State Sorting consists of sorting geometrical shapes with similar states into bins to minimize state changes in the rendering process [8].

¹¹www.openscenegraph.org

¹²www.opengl.org

The 3D visualization module takes as input the ball 3D trajectory (Figs. 18 and 19), the exact bounce point, its distance from the line, its shape and if the bounce point is *in* or *out*. According to this incoming data, the scenario is loaded and the ball trajectory simulated creating an output video file with the controversial play reconstruction. Although the output video is rendered by a conventional camera view for standard TV broadcast, it can be rendered with stereoscopic cameras for future 3DTV broadcasts.

The module configuration defines the variable parameters which describe the scenario: playground border lines width and color, ground material texture, rendered measure number and arrow colors, etc. This makes it easier to configure the virtual scenario for the different real pelota courts where the game takes place.

3.6 Signal adaptation for PAL-quality digital TV signal playout

With regards to the output signal, the system is able to provide HD 1080p digital video throughput. However, it is required to also be compatible with nowadays Standard Definition (SD) broadcaster TV signal standards. Rendered images are adapted to these restrictions. The output video is rendered with a Matrox4 CG2000¹³ video adaptor, since this hardware combines a 3D graphic accelerator with broadcast quality video I/O.

The system output signal can be adapted to lower quality formats (PAL 4:3, 16:9) if it is needed due to compatibility issues.

4 System evaluation

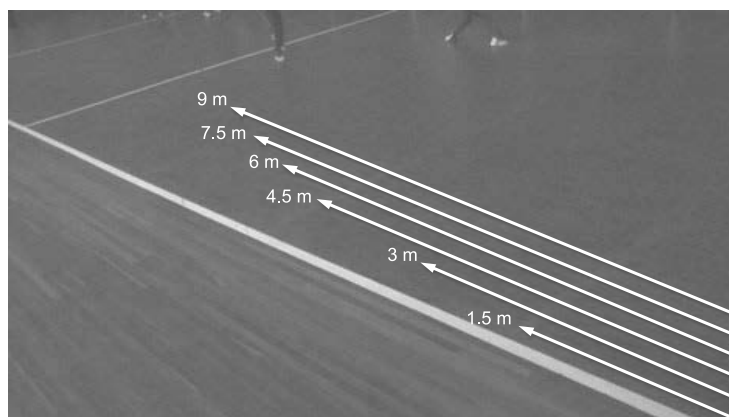
The initial assessment phase of the parameters described in Section 2.2 specifies in depth which parameters of the camera improve the further segmentation process.

According to this first assessment, the scene often is not illuminated enough for proper image acquisition. To improve the image quality, the balanced compromise between shutter speed, diaphragm aperture and electronic gain is set. Furthermore, the thresholds used in ball-player-background segmentation are set according to camera parameters and playground illumination. The evaluation of the system has been done before a professional TV broadcasted Basque Pelota match in three different stadiums. In sport broadcast, the courts must be well illuminated in order to maximize the contrast between foreground and background and make the ball visible over the playground for the audience and TV viewers. Accordingly, the illumination and its changes are under control during match time and our system take advantage of this stable environment.

The test set-up has been another considerable challenge. The ball physic behavior has been studied and tested, reaching that the ball bounce can be considered elastic, because its hardness, ruling out any deformation on its shape or track. In order to get to the conclusion that bounce area remains circular, tracing paper has been used. This tracing paper is set along the border line in order to determine real

¹³www.matrox.com

Fig. 20 Covered distances in the testing trials



distance measurements during testing period. Millimetric paper is set under the tracing paper to get the real distance from the bounce area center and the border line. Consequently, the real distance measurement error can not be above 0,5 mm to achieve market competence.

As mentioned before, the system has been tested in different playgrounds. For all tests, the selected parameters for the system are:

1. **Frame rate** = 50fps
2. **Shutter speed** = 1/500 sec
3. **Diaphragm aperture** = 1:1,7
4. **Image resolution** = 720p HD
5. **Camera lens focal length** = 35 mm
6. **ROI window size** = 100×100px
7. **Detection color space** = Gray scale
8. **File storage codec** = H.264

The camera was located 3 m from the ground and on one side of the line. The field of view allowed covering 9 m (as shown in Fig. 20), enough to cover the controversial action play zone. The camera pan and tilt were different for each test, determined to each playground. The system can be adapted for any kind of sport event, taking into account that the error can vary depending the referees requirements.

According to the results extracted from the tests made in one of the playgrounds (Table 2), the algorithm robustness is proved along different possible camera operating distances. Average error for both test measurements is 4,3 mm, which is below the target 5 mm deviation. The numerical error requirements listed in Section 2.1 are not fulfilled successfully, because the 80 % of measures should be below this 5 mm error threshold and only the 69,4 % of the errors satisfy this requirement. still, 80 % of the errors are below 7 mm.

Although Referee Committee considers this results as acceptable because the typical human eye incertitude in the appreciation from a distance, with millisecond duration, is considerably higher than that of the presented system, ongoing research is being developed in order to fulfill these requirements. One of the objectives of this ongoing research lays on the use of FullHD cameras instead of HD cameras. These cameras represent the same scene using a greater number of pixels and therefore the pixel-real distance ratio decreases, permitting a more precise calibration process and the minimization of error (in real cm measurement) when ball centre point is

Table 2 Accuracy test results for tests made in Ogueta(Vitoria) playground

Distance from camera [d] (m)	Test 1			Test 2		
	Real measure (cm)	System measure (cm)	Error (cm)	Real measure (cm)	System measure (cm)	Error (cm)
$d < 1.5$ m	8.40	8.65	0.25	-8.20	-8.34	0.14
$d < 1.5$ m	2.85	2.86	0.01	7.10	6.60	0.50
$d < 1.5$ m	3.75	4.12	0.37	0.80	0.57	0.23
$1.5 \text{ m} < d \leq 3$ m	0.75	-0.04	0.79	-0.50	-0.90	0.40
$1.5 \text{ m} < d \leq 3$ m	-3.75	-3.95	0.20	-5.20	-5.80	0.60
$1.5 \text{ m} < d \leq 3$ m	-0.10	-0.34	0.24	-7.80	-7.56	0.24
$3 \text{ m} < d \leq 4.5$ m	19.00	18.85	0.15	-3.00	-2.58	0.42
$3 \text{ m} < d \leq 4.5$ m	1.75	0.90	0.85	-0.70	-1.07	0.37
$3 \text{ m} < d \leq 4.5$ m	-3.20	-3.15	0.05	-3.55	-2.90	0.65
$4.5 \text{ m} < d \leq 6$ m	3.70	2.45	1.25	-4.75	-3.95	0.80
$4.5 \text{ m} < d \leq 6$ m	6.50	5.64	0.86	2.50	3.40	0.90
$4.5 \text{ m} < d \leq 6$ m	2.00	1.40	0.60	-6.20	-5.80	0.40
$6 \text{ m} < d \leq 7.5$ m	7.10	7.37	0.27	4.00	4.70	0.70
$6 \text{ m} < d \leq 7.5$ m	6.90	6.95	0.05	3.20	3.85	0.65
$6 \text{ m} < d \leq 7.5$ m	-1.70	-1.75	0.05	2.70	3.10	0.40
$7.5 \text{ m} < d \leq 9$ m	3.30	3.50	0.20	1.90	2.30	0.40
$7.5 \text{ m} < d \leq 9$ m	4.20	4.50	0.30	12.60	12.00	0.60
$7.5 \text{ m} < d \leq 9$ m	11.90	12.00	0.10	2.10	1.80	0.30

detected. The other major issue comes from the control of playground illumination, in order to improve the ball segmentation process.

The tests reveal that the precision in measurements is related to the accuracy in ball center pointing (in each captured frame) and to the accuracy in the homography matrix calculation, both of which are closely related to image resolution. Actually, the error in measurement is not constant across the field. Although the resolution of the image is constant, the real distance that a pixel represents (pixel/distance ratio) is different depending on camera location. The longer the distance between the line-point and the camera, the lower the (pixel/distance) ratio. Nevertheless, the experimental results show that this theoretical issue is not crucial for distances less than 9 m from the camera at HD 720p resolution.

In the springs of 2010, 2011 and 2012 the system was tested in the most important Basque Pelota competitions. Although the numerical measurements did not accomplish the goals of the Referee Committee, they considered the system ready to help them taking decisions during the match. The system worked as expected on professional platforms and the output signal was broadcasted live by the Basque public broadcaster (EiTB¹⁴) successfully. In 2010 it was watched by 219.000 spectators and the viewer share was 31,1 %.

The opportunity of broadcasting the virtual 3D repetition of the bounce permits to the spectator to get more information about the ongoing match. Due to the velocity of the ball and the limits of the broadcasting cameras, it's no viable to reproduce the last recorded frames and detect the bounce point of the ball. Only making a

¹⁴www.eitb.com

reconstruction of the followed track it's possible to determine this point. Therefore, the user experience is enhanced using the results of the described system.

5 Conclusions

In this work a low-cost automatic ball bounce detector and 3D virtual replay generator is proposed for sport event broadcast. The central engineering trade-off choice approach has been to reduce the bounce detection system set-up and hardware requirements to unique broadcast-type camera per border line as well as to reduce the system software to quasi real-time performance. The challenge of hardware complexity reduction keeping accuracy in results can be considered the main technical contribution of this work.

The algorithm introduces an additional advantage which makes it more flexible in terms of different possible camera operating location. Contrary to other approaches, the camera and the playground plane can form any angle, since the necessary transformations for calculating the point of impact in real coordinates are effective and accurate. For this reason, the system has no camera location constraint, provided that line view is not occluded.

The typical human eye incertitude in the appreciation of distant actions (a few meters from linesman to bounce events point), with millisecond duration (because of the ball speed) is considerably higher than the score of the presented system. Obtained results show that the measure errors are close to the demanded range in order to standardize the system for Basque Pelota events.

The use of 3D virtual reality for controversial action replays in sport event broadcasting enhances audiences and TV spectators' visual experience. Due to the reduction of the production costs, this contribution represents a new opportunity for less-known traditional/regional sport events to use this technology, as well as, for small producers, organizers and broadcasters to compete with well-known competition organizers and expensive broadcasting rights owners.

The experience of having developed a research effort applied to real world deployment for sports events has materialized a complete solution covering the whole production chain. Technical specifications and hardware requirements for a system that has to be included in a real world implementation are stronger than the ones required for a system with not so close relation with real world applications. Even more, several variables are no more under control of the researcher, which makes the work harder.

This work has been granted with the patent EP2455911 **Method for detecting the point of impact of a ball in sports events.**

Acknowledgements The authors would like to acknowledge the collaboration offered by G93 Telecomunicaciones¹⁵ (Audio-Visual, Computer and Graphic Services for Television) and EiTB (Basque public broadcaster) for the help offered in the system development, test and broadcast processes.

The authors are also grateful for the collaboration offered by ASPE¹⁶ (ASPE Jugadores de Pelota) in providing access to its professional pelota player training sessions and in advising in game

¹⁵www.g93.es

¹⁶www.aspepelota.com

rule issues as well as for the financial support offered by research project programs of the SPRI¹⁷ (Society for Industrial Promotion and Restructuring of Basque Country).

Finally, the authors would like to thank the rest of *Begira* research team: Maider Laka, Julen García and Aritz Legarretaetxebarria. Also, Javier Barandiaran and Iñigo Barandiaran for their advice and the colleagues of *Digital Television and Multimedia Services* department for the unconditional help offered.

References

- Ahn S (2004) Least squares orthogonal distance fitting of curves and surfaces in space. In: Lecture notes in computer science. Springer (2004). <http://books.google.es/books?id=we4cHJBFzLwC>
- Alper Yilmaz OJ, Shah M (2006) Object tracking: a survey. *ACM Comput Surv* 38(4). doi:10.1145/1177352.1177355
- Cavallaro R (1997) The foxtrax hockey puck tracking system. *IEEE Comput Graph Appl* 17:6–12
- Erik Cueva DZ, Rojas R (2005) Kalman filter for vision tracking. Freie Univ., Fachbereich Mathematik und Informatik
- Yan F, Christmas W, Kittler J (2005) A tennis ball tracking algorithm for automatic annotation of tennis match. In: British machine vision conference, vol 2, pp 619–628
- Sullivan GJ, Topiwala PN, Luthra N (2004) The h.264/avc advanced video coding standard: overview and introduction to the fidelity range extensions. In: SPIE 49th annual meeting optical science and technology, international society for optics and photonics, pp 454–474
- Gopal Pingali Agata Opalach YJ (2000) Ball tracking and virtual replays for innovative tennis broadcasts. In: Proceedings 15th international conference on pattern recognition, 2000, vol 4. IEEE, pp 152–156
- Haines E, Akenine-Moller T (2002) Real-time rendering, 2nd edn. AKPeters
- Innovations HE (2013) Hawk-eye accuracy and believability. <http://www.hawkeyeinnovations.co.uk/>
- Isard M, Blake A (1998) Condensation—conditional density propagation for visual tracking. *Int J Comput Vis* 29(1):5–28
- Inurrategi ML, Olaizola IG, Ugarte A, Macia I (2008) Tv sport broadcasts: real time virtual representation in 3d terrain models. In: 3DTV conference: the true vision—capture, transmission and display of 3D video, 2008. IEEE, pp 405–408
- Miller FP, Vandome AF, McBrewster J (2009) MPEG-2: lossy compression, video compression, audio compression (data), ATSC (standards), MPEG transport stream, MPEG-1 audio layer II, H. 262/MPEG-2 Part 2, MPEG-4, advanced audio coding. Alpha Press. <http://dl.acm.org/citation.cfm?id=1822909>
- Owens N, Harris C, Stennett C (2003) Hawk-eye tennis system. In: International conference on visual information engineering, VIE 2003. IET, pp 182–185
- Naidoo WC, Tapamo JR (2006) Soccer video analysis by ball, player and referee tracking. In: Proceedings of the 2006 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries. South African Institute for Computer Scientists and Information Technologists. pp 51–60
- Richardson IE (2003) The H.264 advanced video compression standard, 2nd edn. Vcodex Limited, UK
- Richardson IEG (2002) Video codec design: developing image and video compression systems. Wiley
- Wu W (2010) Tennis touching point detection based on high speed camera and Kalman filter. Clemson University
- Yu X, Leong HW, Xu C, Tian Qi (2006) Trajectory-based ball detection and tracking in broadcast soccer video. *IEEE Trans Multimedia* 8(6):1164–1178
- Yu X, Xu C, Leong HW, Tian, Qi, Tang Q, Wan KW (2003) Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video. In: Proceedings of the eleventh ACM international conference on Multimedia. ACM, pp 11–20

¹⁷www.spri.es



Mikel Labayen Between 2001 and 2005, Mikel studied Technical Telecommunication Engineering at the Universidad Publica de Navarra (UPNA, 2001–2005), specializing in Image and Sound. He undertook his undergraduate dissertation and final year project at University of Surrey, U.K. He completed his studies and graduated from the faculty of Telecommunication Engineering in between 2005–2007, also at the Universidad Pública de Navarra. This time he undertook his graduate dissertation and final year project in Vicomtech-Ik4 research center where he is now (since, 2007) a staff researcher in the field of Digital Television and Multimedia Services. Since 2010, he also teaches junior level courses at the University of the Basque Country at the Electronics and Telecommunications Department.



Igor G. Olaizola received a six years degree in Electronic and Control Engineering from the University of Navarra, Spain (2001). He developed his master thesis at Fraunhofer Institut für Integrierte Schaltungen (IIS), Erlangen Germany 2001 where worked for a year as research assistant on several projects related to MPEG standard audio decoding. He is member of Vicomtech technological center since 2002. Nowadays he is the head of the Digital Interactive TV and Multimedia Services department. In 2006 he also participated as a technology consultant in Vilau (www.vilau.net) company for one year. Moreover, he is working in his PhD based on automatic multimodal indexing and management of multimedia content at the Faculty of Informatics of San Sebastian, in the University of Basque Country.



Naiara Aginako Between 2000 and 2005 she studied Ingeniería Superior de Telecomunicación (Computer Engineering applied to Telecommunication) at the University of the Basque Country (UPV-EHU)(www.ehu.es). She was cooperating with the Electronics and Communications Department of the above mentioned Faculty between 2003 and 2005 where she undertook her undergraduate dissertation. Since October of 2005 she has been working and leading research projects in Vicomtech-IK4, as a member of the Digital Television and Multimedia Services department. She undertook her postgraduate course in Informatics Engineering at the University of the Basque Country (UPV-EHU) during 2008–2010. Since 2006, she also teaches junior level courses at the University of the Basque Country at the Electronics and Telecommunications Department.



Julián Florez studied Industrial Engineering in the University of Navarra (1980), and obtained his PhD in the University of Manchester, Institute of Science and Technology UMIST (1985), in the field of Adaptive Control. From 1985 to 1990, he worked as Researcher in the Center of Study and Technical Research of Gipuzkoa (CEIT), where he collaborated in several research projects related to Electrical and Industrial Engineering with a marked industrial focus. From 1985 to 1994, he was Associate Professor in the School of Industrial Engineering of the University of Navarra, and since 1994, he is Professor at the same university. From 1990 to 1997, Dr. Flórez worked as Senior Researcher in CEIT, where he was in charge of a Department of Industrial Applications. From 1997 to 2001, he worked as Director of Corporate Development of AVANZIT-SGT (Servicios Generales de Teledifusión) in the fields of Information Systems, Communications and Broadcasting infrastructure. He has a strong background in Digital Television infrastructures and was tightly involved in the deployment of one of the biggest Digital TV organizations in Spain and Europe: Quiero TV. Since 2001, he is Principal Researcher in Vicomtech-Ik4. He holds some patents and has written more than 40 research papers in different areas of Industrial and Electrical Engineering.

2.3.4 Machine Learning for Video Action Recognition: a Computer Vision Approach

- **Izenburua:** Machine Learning for Video Action Recognition: a Computer Vision Approach
- **Egileak:** Naiara Aginako, Goretti EcheGARAY, Igor G. Olaizola, Basilio Sierra
- **Aldizkaria:** Machine Vision and Applications (AURKEZTUA)
- **Argitaletxea:** SPRINGER
- **Orrialdeak:** 1-15
- **Inpaktu-faktorea (urtea):** 1,272 (2015)
- **Kuartila:** Q2
- **Urtea:** 2017

Machine Learning for Video Action Recognition: a Computer Vision Approach

Naiara Aginako · Goretti Echeagaray ·
Igor G. Olaizola · Julián Flórez · Basilio
Sierra

Received: date / Accepted: date

Abstract The automatic detection of video action is still a challenging research task. In this paper, we consider a first atomic approach to classify a single action in a short video sequence. The presented method combines four different concepts: global image descriptors, image transformation algorithms, Machine Learning paradigms for supervised classification and Feature Subset Selection (FSS) techniques. Using an image characterization method called DITEC which is based on the Trace Transform, the information contained in a video is handled as an image. This allows us to apply Image Processing solutions for the analysis of the video, more concretely, of the occurring action. Key features are extracted to nourish Machine Learning classifiers in order to predict the action. The final step is to use a Feature Subset Selection (FSS) standard method to select the most accurate attributes for the identification of the action. The idea of understanding videos as images widens the possibilities for the analysis of temporal behavior of actions within a video.

Keywords Computer Vision · Image and Video Processing · Machine Learning · Action Recognition

1 Introduction

Action recognition in videos is a challenging task which is nowadays being tackled by several researches from different perspectives. Some of them apply Computer Vision approaches [32] [23] [8], others Machine Learning Classifiers [33] and a combination of both of them is also being used by other authors.

This paper presents a new action recognition method for activities being performed in a video. Our method consists on four consecutive steps: the

Naiara Aginako
Faculty of Engineering EHU-UPV (Donostia)
Tel.: +34-943017229
E-mail: naiara.aginako@ehu.eus

extraction of global image descriptors, application of image processing algorithms, using Machine Learning paradigms for supervised classification and implementation of Feature Subset Selection (FSS) methods for attribute selection. The main novelty relies on the compression of the temporal behavior of the videos in an information matrix and the consideration of this matrix as an image. To that end, first, an image characterization method called DITEC [20] is applied to each video frame. This method is based on the Trace Transform [15]. Thus, each frame is reduced to a previously fixed number of characterizing attributes. This number has been assigned after accomplishing different experiments with the aim of obtaining the best image scene classification results [13].

As a consequence, each video can be understood as a matrix containing this fixed number of attributes as columns and the number of video frames as rows. Therefore, the temporal behavior of the video is compressed within these attributes. If this matrix is considered as an image, videos could be handled and analyzed as an image based Computer Vision problem. Concretely, image processing algorithms can be applied to these images, and the prediction of the action can be proposed to be done using a Machine Learning classifier. Finally, a FSS standard method is applied to select the most accurate attributes for the classification.

The test dataset of the proposed approach is based on the well-known UT-Interaction video action recognition dataset. More specifically, it is based on the dataset used in the ICPR 2010 Contest on Semantic Description of Human Activities challenge. Mention that the obtained results improve the results of the first classified participant.

The rest of the paper is structured as follows. Section 2 introduces related work. In Section 3 the new approach is presented, focusing on the main four steps. In Section 4 the experimental framework and its results are detailed. Finally, Section 5 lists the conclusions and future work ideas.

2 Related work

The exponential growth of video content has led to a great need for methods of intelligent video analysis and understanding. Among them, human action recognition plays a central role in many applications such as surveillance, injury rehabilitation or sport training. The majority of the research accomplished in this topic is focused on different methods of extracting and representing salient features from video actions. From the image analysis perspective, these methods can be classified into two categories: local and global based features extraction.

To date, most of the work in human action recognition has been done on a selection of a few datasets with an static and simple background, such as Weizmann [35] and KTH [27]. These datasets are not representative of the real world action recognition problems but they could be considered the starting point for the development of automatic retrieval systems. Most of the best

results obtained using these datasets have been achieved using global features [29] mainly due to the simplicity of the background. Indeed, there is few noise to interfere in the classification process.

As stated in [15], Trace Transform used as a global feature is a very powerful tool for image classification because it permits the construction of invariant features for the identification of the images. Even more, [11] proposes a novel method based on the Trace Transform for Human Action Recognition tested on the aforementioned datasets.

Moreover, popular classification models used for image and video classification include, among others, different decision trees [30], Support Vector Machines (SVM) [10], [26], [16], [21], Neural Networks [5] and Hidden Markov Models (HMM) [19], [28]. Being the best classifier often depends on the application and corpus [22].

In order to combine the benefits of different classifiers, classifiers' fusion is starting to become a common practice and several examples can be found in the literature [2]. Arruti et al. [4] use four Machine Learning paradigms (IB, ID3, C4.5, NB) and evolutionary algorithms to select feature subsets that noticeably optimize the automatic emotion recognition success rate. Schuller et al. [26] combine SVMs, decision trees and Bayesian classifiers to yield higher classification accuracy.

3 Proposed approach

The proposed approach relies on the following ideas: first, each video is converted into a matrix that will be considered as an image. Second, image classification solutions are applied to deal with the action recognition task.

In order to convert a video into a matrix that compresses its main information, a global descriptor is applied to each frame. Therefore, each frame is represented as a vector of descriptors with a predefined length. Consequently, considering these descriptors as the columns of a matrix, and frames as lines, we obtain a matrix representation for each video. The main contribution of the presented work relies on the understanding of this information matrix as a gray scale image. One of the characteristics of this image is that pixel values are not real values between 0-255 but float values that are instead the attributes of the DITEC global descriptor.

Remark that one of the drawbacks of this idea is the necessity of resizing all the matrices to a fixed number of lines for the classification step. This is due to the different length of the chosen videos. This action could result in a loss of key information for the recognition task.

Once a fixed size matrix has been created from the analyzed video, this is already treated as a grayscale image with the particularity mentioned above. Thus, different image transformations are computed to this image in order to determine the most suitable one for the recognition task.

Four are the main steps of the action recognition process, and hence the concepts, to be taken into consideration:

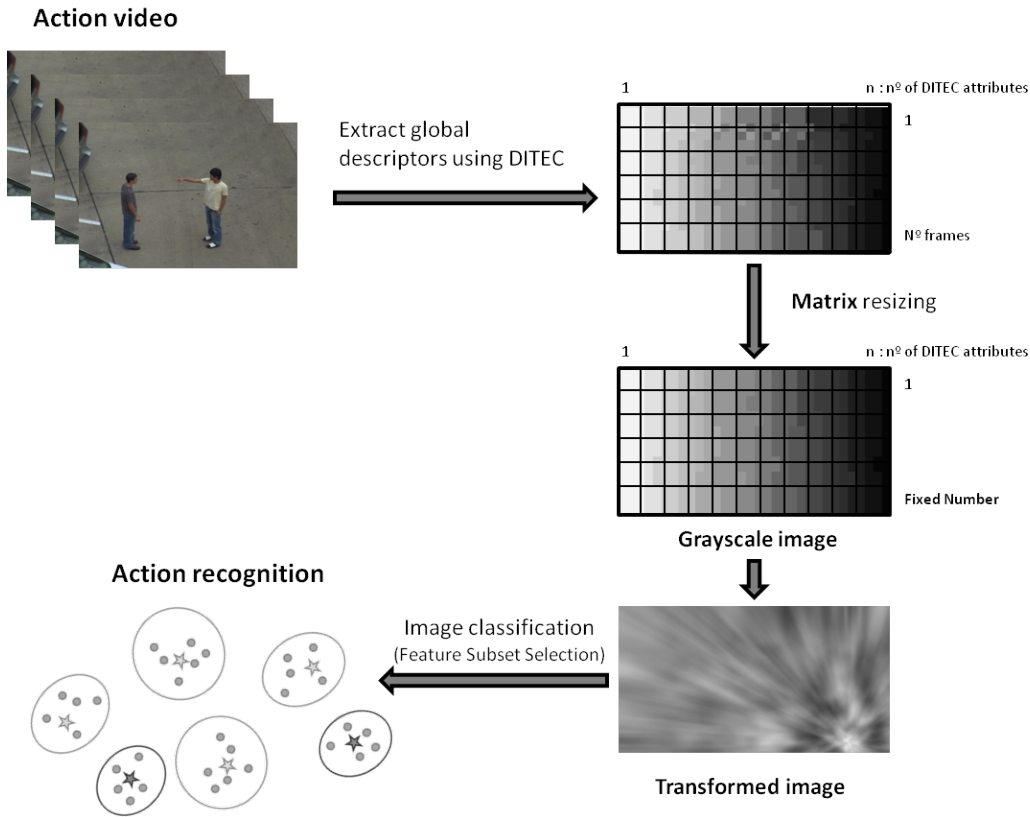


Fig. 1 Proposed approach

3.1 DITEC

The main objective of the DITEC method in the presented solution is the extraction of global descriptors that characterize each video frame. DITEC is based on statistical modeling of the Trace Transform. Indeed, Trace Transform has already been proposed for several Computer Vision applications and its effectiveness has been proven.

Concretely, two are the main steps for this approach: data transformation and feature extraction.

- **Data transformation:** Trace Transform applied to the pre-processed image I . In this case image I will be each video frame. The result will depend on the chosen functional and on the selected geometric parameters. The outcome T of the Trace Transform of an image is a two-dimensional signal represented by means of sinusoids with a particular amplitude, phase, frequency and intensity.
- **Feature extraction:** summarization of the extracted features T , compressed and adapted into a manageable object-based descriptor. The wave features contained in the resulting image must be characterized. In order to do this, the 2D trace signal Tk is transformed to the frequency domain

and a two-dimensional DCT (Discrete Cosine Transform) is applied. Then, the DCT is compressed to a vector of two components (average value and kurtosis of all the orthogonal elements of the main diagonal).

The 2D forward DCT is given by:

$$X_{k_1 k_2} = \alpha_{k_1} \alpha_{k_2} \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} x_{n_1 n_2} \cos \left[\frac{\pi k_1 (2n_1 + 1)}{2N_1} \right] \cos \left[\frac{\pi k_2 (2n_2 + 1)}{2N_2} \right] \quad (1)$$

where:

$$\alpha \in \begin{cases} \frac{1}{\sqrt{N_i}} & k_i = 0 \\ \sqrt{\frac{2}{N_i}} & k_i \neq 0 \end{cases} \quad (2)$$

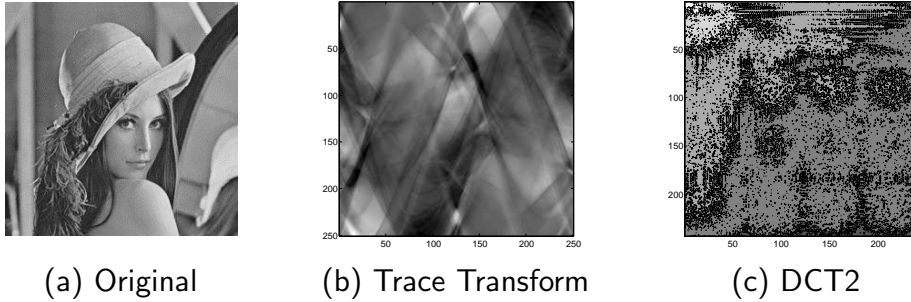


Fig. 2 Trace Transform and subsequent Discrete Cosine Transform of Lenna. (Y channel of YCbCr color space)

To summarize, DITEC permits describing each video frame as a vector of a fixed and constant number of attributes. In the experimental phase the length of the vector has been set at 40, as previous research on the method has demonstrated the adequacy of this length taking into account the performance and the preciseness. The matrix composed by all DITEC vectors extracted from the video is considered an image in the following processes.

3.2 Image Transformations

In order to obtain different image descriptors, diverse transformations have been applied to the original gray scale pictures. In this way, we acquire multiple

aspects for the same picture and different values for the same pixel positions that are necessary for the classification step.

This work has performed some of the most common transformations in order to show the benefits of the proposed approach making use of simple algorithms. Table 1 lists the collection of transformations used, as well as a brief description of each one. It is worth stressing that any other CV transformations could also be used.

Table 1 Used image transformations

<i>Transform</i>	Command	Effect
<i>Transf. 1</i>	Convolve	Apply a convolution kernel to the image
<i>Transf. 2</i>	Despeckle	Reduce speckle within an image
<i>Transf. 3</i>	Edge	Apply a filter to detect edges in the image
<i>Transf. 4</i>	Enhance	Apply a digital filter to enhance a noisy image
<i>Transf. 5</i>	Equalize	Perform histogram equalization to an image
<i>Transf. 6</i>	Gamma	Perform a gamma correction
<i>Transf. 7</i>	Gaussian	Reduce image noise and reduce detail levels
<i>Transf. 8</i>	Lat	Local adaptive thresholding
<i>Transf. 9</i>	Linear-Str.	Linear with saturation histogram stretch
<i>Transf. 10</i>	Median	Apply a median filter to the image
<i>Transf. 11</i>	Modulate	Vary the brightness, saturation and hue
<i>Transf. 12</i>	Negate	Replace each pixel with its complementary color
<i>Transf. 13</i>	Radial-blur	Radial blur the image
<i>Transf. 14</i>	Raise	Lighten/darken image edges to create a 3-D effect
<i>Transf. 15</i>	Selective-blur	Selectively blur pixels within a contrast threshold
<i>Transf. 16</i>	Sharpen	Sharpen the image
<i>Transf. 17</i>	Shade	Shade the image using a distant light source
<i>Transf. 18</i>	Shave	Shave pixels from the image edges
<i>Transf. 19</i>	Trim	Trim image edges
<i>Transf. 20</i>	Unsharp	Unsharpen the image

3.3 Machine Learning

In the training set used to generate the classification model for the supervised learning task, the y label value associated to each x sample is known. For this analysis, paradigms which come from Artificial Intelligence and belong to *Machine Learning* or *ML* family are used. For this task, we have selected five different standard classifiers. A brief description of each of them is included below:

– **Classification Trees**

The C4.5 [24] represents a classification model by a decision tree. The tree is constructed in a top-down way, dividing the training set and beginning with the selection of the best variable in the root of the tree.

– **Bayesian Networks (BN)**

A Bayesian Network [6], Belief Network or Directed Acyclic Graphical Model is a probabilistic graphical model that represents a set of random

variables and their conditional independencies via a directed acyclic graph (DAG).

– **Naive Bayes (NB)**

The Naive-Bayes rule [7] uses the Bayes theorem to predict the class for each case, assuming that the predictive genes are independent given the category. To classify a new sample characterized by d genes $\mathbf{X} = (X_1, X_2, \dots, X_d)$, the NB classifier applies the following rule:

$$c_{NB} = \arg \max_{c_j \in C} p(c_j) \prod_{i=1}^d p(x_i | c_j)$$

where c_{NB} denotes the class label predicted by the Naive-Bayes classifier and the possible classes of the problem are grouped in $C = \{c_1, \dots, c_l\}$.

In simple terms, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. In spite of their naive design and apparently over-simplified assumptions, Naive Bayes classifiers often work much better in many complex real-world situations than one might expect.

– **K-Nearest Neighbors (K-NN)**

The NN classification decision method gives to x the category θ'_n of its nearest neighbor x'_n . In case of a tie in the nearest neighbor, the decision rule is modified in order to break it. A mistake is made if $\theta'_n \neq \theta$.

An immediate extension to this decision rule is the so called k -NN approach [3], where the candidate x is assigned the most frequent class in the k nearest neighbors of x . We have used the IB algorithm [1].

– **Support Vector Machines (SVM)**

SVM are a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an n -dimensional space, a SVM constructs a separating hyperplane in that space that maximizes the margin between the two data sets. This margin is calculated by constructing two parallel hyperplanes, one on each side of the separating hyperplane, which are "pushed up against" the two data sets. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighboring datapoints of both classes, since in general the larger the margin the lower the generalization error of the classifier [9].

Selected classifiers belong to different families (trees, probabilistic,...) and are widely used. The idea is to deal with this variety of classifiers aiming to select the one that gives the best results. This is the usual mode of affording a classification problem in the Machine Learning community [18].

3.4 Feature Subset Selection (FSS)

The goal of Machine Learning is to induce a classifier that allows us to classify new examples that are only characterized by their descriptive features. In

this way, a 'general rule' is induced to classify new examples using a learning algorithm. However, not all descriptive features are always useful for this purpose. For that reason, the Feature Subset Selection (FSS) approaches select the best subset of candidate features. This leads to reduce the data acquisition cost, improve the comprehensibility of the final classification model, fasten the induction of the final classification model and improve classification accuracy [14].

The used approach searches for a good variable subset, independent to the selected classifier, considering the relationship between the predicting variables and the class. Conversely, a ranking of the variables is obtained according to the used criteria –a mathematical formula– and some of those variables which are at the top of that ranking are selected.

4 Experiments

In this section, the whole experimental design is described. Firstly, the case study and the dataset are given, followed by the definition of the experimental setup.

4.1 Case study

As already mentioned, in this work a well known set of videos that was presented in ICPR 2010 Contest on Semantic Description of Human Activities (SDHA 2010) ¹ completes the test dataset. Choosing this dataset allows the reproduction of the experiments and therefore guarantees their re-usability.

Within this Contest in the "High-level Human Interaction Recognition Challenge", participants are expected to recognize ongoing human activities from continuous videos. The intention is to motivate researchers to explore the recognition of complex human activities from videos taken in realistic settings. Each video contains several human-human interactions (e.g. hand shaking) occurring sequentially and/or concurrently.

Concretely, the UT-Interaction dataset contains videos of continuous executions of 6 classes of human-human interactions (see Figure 3): Hand Shaking (3a), Hugging (3b), Kicking (3c), Pointing (3d), Punching (3e) and Pushing (3f). Ground truth labels for these interactions are provided, including time intervals and bounding boxes.

Videos are divided into two sets of 10 video sequences each: Set 1 is composed of videos taken on a parking lot. They have a slightly different zoom rate and their backgrounds are mostly static with little camera jitter. Videos from set 2 are taken in a lawn in a windy day. The background contains slight movements (e.g. tree moves) and they have more camera jitters. Some sequences show only two interacting persons; however, in other sequences the scene presents interacting persons AND pedestrians. Each set has a different

¹ <http://cvrc.ece.utexas.edu/SDHA2010/Human.Interaction.html>

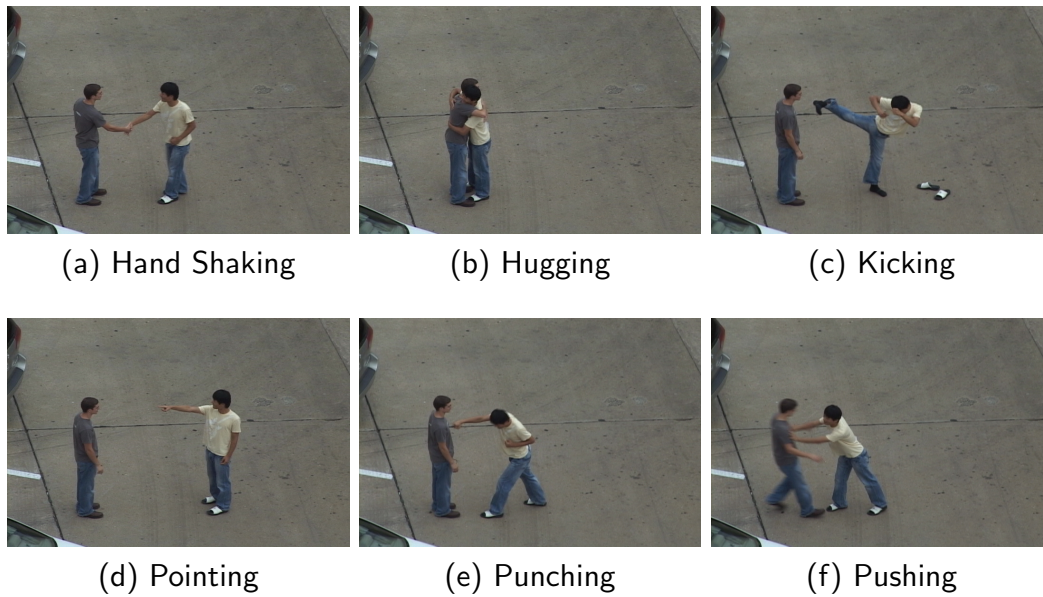


Fig. 3 Six different actions are performed in the videos

background, scale and illumination. Therefore, there are in overall 20 videos of approximately 1 minute length. Video resolution is 720×480 , 30fps, color and the height of a person occupies about 200 pixels.

But for the "High-level Human Interaction Recognition Challenge", the original 20 videos have been cropped into 120 videos focusing on when and where the action happens. The resolution of these segmented videos depends on the bounding box of the segmented area. Although they are short videos, it is worth noticing that the number of frames differs from 43 frames (the shortest one) to 175 frames (the largest). These are the videos used for testing the presented approach, as the objective of the method is the action recognition and not it's detection.

Results of the experiments performed in the challenge have been reported on [25]. Average results achieved by each method are shown in Table 2. The first column represents the name of the methods applied by each of the participants. These results are used to make the final comparison with the results of the proposed approach.

4.2 Experimental results analysis

In all the experiments the 10-fold cross-validation [31] was applied to get a validated classification accuracy (well classified rate). This accuracy reflects the number of times the actions are correctly recognized for each video instance.

Tables 3 to 6 present the results obtained for each of the Machine Learning paradigms used. The best accuracy obtained per classifier is highlighted in bold. The results suggest that Support Vector Machine (SVM) is the classifier which performs better. It obtains a 90% of well classified videos, which means

	Shake	Hug	Kick	Point	Punch	Push	Total
Laptev + kNN	0.24	0.435	0.665	0.93	0.535	0.395	0.53
Laptev + Bayes.	0.37	0.695	0.545	0.9	0.41	0.46	0.56
Laptev + SVM	0.49	0.715	0.63	0.85	0.535	0.495	0.62
Latpev + SVM (best)	0.5	0.75	0.75	0.85	0.55	0.6	0.67
Cuboid + kNN	0.605	0.8	0.45	0.915	0.485	0.485	0.62
Cuboid + Bayes.	0.375	0.77	0.72	0.95	0.36	0.43	0.60
Cuboid + SVM	0.665	0.815	0.635	0.91	0.575	0.545	0.69
Cuboid + SVM (best)	0.8	0.85	0.75	0.95	0.7	0.6	0.78
Team BIWI	0.6	0.95	1	1	0.75	0.65	0.83

Table 2 Challenge average results

that only 12 of the 120 videos have been incorrectly classified. The rest of the best accuracy values range from 49.17% (BN) to 80% (NB).

Columns indicate the new size of the image that represents the video. This resizing is necessary so that classification methods have the same number of attributes for all the videos.

Notice that the best results for SVM and Naive Bayes have been obtained with the same Image Transformation -Lattice- and the same image size -100x40-.

Figure 4 shows the best results. As seen in Table 6, the validated accuracy –using Leave One Out as validation approach– is 90%, which means that 108 among the 120 videos have been correctly classified in the actions they belong to. Average results are the following:

- **True Positive rate: 0.9**
- **False Positive Rate: 0.02**
- **Precision: 0.901**
- **Recall: 0.9**
- **F-Measure: 0.9**
- **AUC: 0.976**

Obtained results, separated according to class and action, are compared with the challenge best ones in Table 9. As it can be seen, the best contestant paradigm in the challenge classifies perfectly two among the six classes, conversely *Kick* and *Point*, and it obtains a very good result for another class (*Hug*). However, the accuracy decreases significantly in the other three actions. In the best results of the proposed approach, obtained by using the SVM classifier, the six classes are uniformly well classified. It outperforms the Team BIWI result in three classes (*Shake*, *Punch* and *Push*), a tie is obtained in one of them (*Hug*) and a slightly worse result is obtained in the remaining two classes (*Kick* and *Point*). In overall, the proposed approach obtains a better accuracy than the challenge’s winner, with an increase that goes from 0.83 to 0.90; this implies that 8 more among the 120 videos are well classified compared to the previous best result.

Regarding FSS implementation, after performing some previous experiments using WEKA, we obtained that the best filter for this experiment is

	100x40	30x50	40x100	40x50	50x30	50x40	50x50
Original	25%	35.83%	27.50%	28.33%	22.50%	31.67%	31.67%
Convolve	25%	35.83%	27.50%	28.33%	22.50%	31.67%	31.67%
Despeckle	33.33%	29.17%	43.33%	29.17%	33.33%	35.83%	35.83%
Edge	0%	0%	0%	0%	0%	0%	0%
Enhance	43.33%	35%	35%	35%	32.50%	31.67%	20%
Equalize	40%	40%	0%	10.83%	22.50%	34.17%	30%
Gamma	30.83%	47.50%	32.50%	35.83%	37.50%	32.50%	35%
Gaussian	31.67%	45%	36.67%	35%	41.67%	35%	20%
Lat	0%	0%	0%	0%	0%	0%	0%
Linear-Str.	27.50%	10%	20.83%	42.50%	27.50%	15.83%	26.67%
Median	40%	20%	30.83%	27.50%	37.50%	20.83%	40%
Modulate	31.67%	40.83%	8.33%	8.33%	34.17%	39.17%	36.67%
Negate	25%	35.83%	25%	28.33%	23.33%	31.67%	33.33%
Radial-blur	31.67%	35%	42.50%	32.50%	33.33%	22.50%	32.50%
Raise	32.50%	37.50%	31.67%	49.17%	27.50%	28.33%	33.33%
Selective-blur	25%	35.83%	27.50%	28.33%	22.50%	31.67%	31.67%
Sharpen	4.17%	25.83%	22.50%	26.67%	34.17%	37.50%	24.17%
Shade	0%	0%	0%	0%	28.33%	30.83%	21.67%
Trim	25%	35.83%	27.50%	28.33%	22.50%	31.67%	31.67%
Unsharp	32.50%	12.50%	28.33%	32.50%	28.33%	34.17%	28.33%

Table 3 Bayesian Networks (BN)

	100x40	30x50	40x100	40x50	50x30	50x40	50x50
Original	66.67%	60.83%	66.67%	64.17%	61.67%	65.83%	63.33%
Convolve	66.67%	60.83%	66.67%	64.17%	61.67%	65.83%	63.33%
Despeckle	69.17%	55.83%	64.17%	61.67%	62.50%	60.83%	63.33%
Edge	73.33%	65%	64.17%	62.50%	60.83%	62.50%	68.33%
Enhance	70%	61.67%	65%	60%	67.50%	66.67%	60.83%
Equalize	67.50%	55%	65.83%	61.67%	60.83%	57.50%	61.67%
Gamma	69.17%	63.33%	61.67%	62.50%	67.50%	68.33%	66.67%
Gaussian	60.83%	55.83%	57.50%	53.33%	48.33%	52.50%	50.83%
Lat	80%	53.33%	68.33%	65.83%	68.33%	72.50%	66.67%
Linear-Str.	75%	60.83%	60%	60%	68.33%	66.67%	60.83%
Median	67.50%	65%	66.67%	64.17%	62.50%	64.17%	61.67%
Modulate	67.50%	62.50%	64.17%	58.33%	61.67%	67.50%	70%
Negate	66.67%	60%	61.67%	64.17%	60.83%	66.67%	63.33%
Radial-blur	66.67%	64.17%	56.67%	54.17%	63.33%	63.33%	60.83%
Raise	67.50%	62.50%	62.50%	64.17%	65.83%	65%	65%
Selective-blur	66.67%	60.83%	66.67%	64.17%	61.67%	65.83%	63.33%
Sharpen	70%	57.50%	55.83%	62.50%	65%	63.33%	65%
Shade	60.83%	55.83%	55.83%	60.83%	59.17%	63.33%	62.50%
Trim	66.67%	60.83%	66.67%	64.17%	61.67%	65.83%	63.33%
Unsharp	72.50%	63.33%	66.67%	63.33%	64.17%	62.50%	65.83%

Table 4 Naive Bayes

the so-called SVMAttributeEval [12]. As indicated in the WEKA project documentation, this method evaluates the worth of an attribute using a SVM classifier. Attributes are ranked by the square of the weight assigned by the SVM. Attribute selection for multi-class problems is handled by ranking attributes for each class separately using a one-vs-all (OVA) method and then dealing with them from the top of each pile to give a final ranking.

	100x40	30x50	40x100	40x50	50x30	50x40	50x50
Original	57.50%	48.33%	50.83%	47.50%	50%	52.50%	52.50%
Convolve	57.50%	48.33%	50.83%	47.50%	50%	52.50%	52.50%
Despeckle	50.83%	51.67%	50%	55%	47.50%	51.67%	54.17%
Edge	57.50%	47.50%	58.33%	44.17%	55%	45%	48.33%
Enhance	45.83%	45.83%	45.83%	53.33%	56.67%	51.67%	51.67%
Equalize	50%	41.67%	37.50%	42.50%	49.17%	49.17%	51.67%
Gamma	53.33%	40.83%	48.33%	48.33%	52.50%	45%	49.17%
Gaussian	47.50%	48.33%	53.33%	50.83%	45%	50%	45.83%
Lat	58.33%	34.17%	51.67%	54.17%	56.67%	46.67%	50%
Linear-Str.	60%	42.50%	50%	44.17%	48.33%	47.50%	54.17%
Median	53.33%	50%	53.33%	57.50%	55.83%	56.67%	54.17%
Modulate	53.33%	59.17%	52.50%	49.17%	47.50%	48.33%	53.33%
Negate	57.50%	48.33%	52.50%	47.50%	50%	52.50%	52.50%
Radial-blur	50%	54.17%	54.17%	53.33%	55%	55.83%	50.83%
Raise	47.50%	50.83%	55%	54.17%	52.50%	43.33%	43.33%
Selective-blur	57.50%	48.33%	50.83%	47.50%	50%	52.50%	52.50%
Sharpen	56.67%	48.33%	47.50%	52.50%	50%	47.50%	54.17%
Shade	50%	48.33%	51.67%	39.17%	50%	46.67%	40.83%
Trim	57.50%	48.33%	50.83%	47.50%	50%	52.50%	52.50%
Unsharp	50%	45.83%	45.83%	55.83%	55%	45.83%	50.83%

Table 5 K-Nearest Neighbors (K-NN)

	100x40	30x50	40x100	40x50	50x30	50x40	50x50
Original	69.17%	69.17%	76.67%	67.50%	65%	71.67%	68.33%
Convolve	69.17%	69.17%	76.67%	67.50%	65%	71.67%	68.33%
Despeckle	77.50%	65.83%	65%	65%	64.17%	69.17%	70%
Edge	86.67%	69.17%	69.17%	65.83%	59.17%	72.50%	80%
Enhance	82.50%	70.83%	63.33%	66.67%	75%	65%	73.33%
Equalize	70.83%	64.17%	62.50%	65%	65%	69.17%	65.83%
Gamma	77.50%	65.83%	65%	65.83%	74.17%	65%	68.33%
Gaussian	63.33%	59.17%	64.17%	55%	50%	54.17%	52.50%
Lat	90%	65%	77.50%	71.67%	74.17%	82.50%	77.50%
Linear-Str.	70.83%	60.83%	71.67%	70.83%	70%	69.17%	72.50%
Median	70.83%	71.67%	74.17%	70%	72.50%	73.33%	79.17%
Modulate	75%	67.50%	68.33%	67.50%	75.83%	75%	70.83%
Negate	69.17%	69.17%	75%	67.50%	65%	71.67%	68.33%
Radial-blur	64.17%	57.50%	67.50%	60%	57.50%	65.83%	67.50%
Raise	69.17%	70%	66.67%	72.50%	65.83%	68.33%	70.83%
Selective-blur	69.17%	69.17%	76.67%	67.50%	65%	71.67%	68.33%
Sharpen	79.17%	60.83%	71.67%	67.50%	67.50%	78.33%	68.33%
Shade	79.17%	70%	71.67%	75%	74.17%	74.17%	70%
Trim	69.17%	69.17%	76.67%	67.50%	65%	71.67%	68.33%
Unsharp	76.67%	63.33%	75.83%	74.17%	73.33%	72.50%	70.83%

Table 6 Support Vector Machines (SVM)

Other approaches have also been tested, but the results were very poor. In our opinion, this is due to the fact that there are six classes, and the OVA selection criteria [17] performs better for these multi-class classification problems. Nevertheless, other approaches could also be tried, such as Singular Value Decomposition [34] or evolutionary computation wrapper approaches,

	100x40	30x50	40x100	40x50	50x30	50x40	50x50
Original	36.67%	36.67%	34.17%	35%	25.83%	33.33%	39.17%
Convolve	36.67%	36.67%	34.17%	35%	25.83%	33.33%	39.17%
Despeckle	43.33%	24.17%	34.17%	25.83%	35%	38.33%	26.67%
Edge	28.33%	48.33%	20%	37.50%	35%	35.83%	35%
Enhance	47.50%	40.83%	37.50%	32.50%	31.67%	27.50%	37.50%
Equalize	29.17%	23.33%	43.33%	35%	28.33%	25%	30%
Gamma	36.67%	32.50%	32.50%	37.50%	40%	43.33%	32.50%
Gaussian	30.83%	33.33%	29.17%	30%	33.33%	30%	35%
Lat	27.50%	20%	28.33%	33.33%	34.17%	37.50%	28.33%
Linear-Str.	24.17%	34.17%	29.17%	28.33%	31.67%	25%	27.50%
Median	27.50%	38.33%	31.67%	40%	35%	30%	33.33%
Modulate	32.50%	33.33%	50%	43.33%	35.83%	26.67%	29.17%
Negate	37.50%	35.83%	28.33%	35.83%	25%	33.33%	37.50%
Radial-blur	43.33%	28.33%	37.50%	29.17%	35%	28.33%	35%
Raise	26.67%	37.50%	38.33%	28.33%	39.17%	25.83%	30%
Selective-blur	36.67%	36.67%	34.17%	35%	25.83%	33.33%	39.17%
Sharpen	19.17%	35%	22.50%	26.67%	27.50%	35%	18.33%
Shade	46.67%	33.33%	28.33%	25.83%	28.33%	30%	15.83%
Trim	36.67%	36.67%	34.17%	35%	25.83%	33.33%	39.17%
Unsharp	30.83%	17.50%	37.50%	35%	31.67%	32.50%	24.17%

Table 7 J48

Correctly Classified Instances 108 (90%)						
Incorrectly Classified Instances 12 (10%)						
Kappa statistic 0.88						
Total Number of Instances 120						
=== Detailed Accuracy By Class ===						
TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.9	0.03	0.857	0.9	0.878	0.964	a
0.95	0.01	0.95	0.95	0.95	0.987	b
0.9	0.01	0.947	0.9	0.923	0.984	c
0.95	0.01	0.95	0.95	0.95	0.994	d
0.85	0.04	0.81	0.85	0.829	0.953	e
0.85	0.02	0.895	0.85	0.872	0.976	f

0.9	0.02	0.901	0.9	0.9	0.976	<-- W. Av.
Accuracies per class:						
0.9	0.95	0.9	0.95	0.85	0.85	

Fig. 4 Best result statistics given by WEKA

in which we have some experience and hence are to be investigated in the near future.

	Shake	Hug	Kick	Point	Punch	Push
Shake	18	0	0	1	1	0
Hug	1	19	0	0	0	0
Kick	0	1	18	0	1	0
Point	0	0	1	19	0	0
Punch	1	0	0	0	17	2
Push	1	0	0	0	2	17

Table 8 Best results (SVM classifier) confusion matrix

	Shake	Hug	Kick	Point	Punch	Push	Total
BN	0.55	0.7	0.3	0.7	0.3	0.4	0.49
NB	0.85	0.8	0.75	0.95	0.8	0.65	0.80
J48	0.5	0.8	0.5	0.75	0.2	0.25	0.50
KNN	0.65	0.7	0.7	0.85	0.4	0.25	0.51
SVM	0.9	0.95	0.9	0.95	0.85	0.85	0.90
Team BIWI	0.6	0.95	1	1	0.75	0.65	0.83

Table 9 Final results of the proposed approach; Challenge best result is shown in the last row for comparison purposes

5 Conclusions and future work

Giving computers the ability to recognize human actions in videos is still an ongoing research topic, which is now considered one of the most relevant challenges in Video and Image Based classification.

In this paper a novel approach is presented and evaluated. The main idea and first step of the proposed approach is to obtain an image that represents the whole video. Therefore, the obtained image is the temporal representation of a video. In our case, DITEC is used to do that compression but further studies on this matter should be accomplished in order to deeply analyze the consequences and benefits of using it. Even more, this study will probably bring into light relevant aspects of the temporal behavior of actions, valuable to tackle the action detection task in videos.

Once the image is obtained, it is used in a classification approach which requires previous image related operations and transformations. It is worth mentioning that, apart from the ones used in this study, several other approaches related to image based classification could be used. This is to be investigated in the near future.

Machine Learning classifiers are used to test the accuracy of the proposed approach; as a future work a new multi-classifier system [36] is to be adapted and learned to deal with video action classification.

Feature Subset Selection is a well known technique used to improve the obtained results. We have used a standard approach, but finding a new way to select an appropriate subset of predictor variables by means of an evolutionary paradigm [4] would be an interesting challenge.

A future work of this research will be to perform new experiments on different databases in order to test out the efficiency of the presented new classification approach in other data set conditions and domains. Other Ma-

chine Learning paradigms are to be tested as well; we aim at improving the results using multi-classifier models.

Acknowledgements This research work was developed under the umbrella of P-REACT European project (project reference: 607881). It was also partially funded by the Spanish Ministry of Economy and Competitiveness within the project TIN2015-64395-R (MINECO/FEDER) and by the Department of Education, Universities and Research of the Basque Government (grant IT900-16). Robotika eta Sistema Autonomoen Ikerketa Taldea (RSAIT) is part of the BAILab unit for research and teaching supported by the University of the Basque Country (UFI11/45).

References

1. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. *Machine Learning* **6**, 37–66 (1991)
2. Alhamdoosh, M., Wang, D.: Fast decorrelated neural network ensembles with random weights. *Information Sciences* **264**, 104–117 (2014)
3. Altman, N.S.: An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* **46**(3), 175–185 (1992)
4. Arruti, A., Cearreta, I., Álvarez, A., Lazkano, E., Sierra, B.: Feature selection for speech emotion recognition in spanish and basque: On the use of machine learning to improve human-computer interaction. *PloS one* **9**(10), e108,975 (2014)
5. Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E.: Desperately seeking emotions or: Actors, wizards, and human beings. In: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion* (2000)
6. Buntine, W.: Theory refinement on bayesian networks. In: *Proceedings of the Seventh conference on Uncertainty in Artificial Intelligence*, pp. 52–60. Morgan Kaufmann Publishers Inc. (1991)
7. Cestnik, B.: Estimating probabilities: a crucial task in machine learning. In: *ECAI*, vol. 90, pp. 147–149 (1990)
8. Cheng, G., Wan, Y., Saudagar, A.N., Namuduri, K., Buckles, B.P.: Advances in human action recognition: a survey. *arXiv preprint arXiv:1501.05964* (2015)
9. Cortes, C., Vapnik, V.: Support-vector networks. *Machine learning* **20**(3), 273–297 (1995)
10. Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., Truong, K.: The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing* **In press** (2015)
11. Goudelis, G., Karpouzis, K., Kollias, S.: Exploring trace transform for robust human action recognition. *Pattern Recognition* **46**(12), 3238–3248 (2013)
12. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. *Machine Learning* **46**, 389–422 (2002)
13. Igor G. Olaizola Iigo Barandiaran, B.S.M.G.: Ditec: Experimental analysis of an image characterization method based on the trace transform. In: *VISAPP 2013 .- International Conference on Computer Vision Theory and Applications, Volume: 1, vol. 1. INSTICC, Barcelona, Spain* (2013)
14. Inza, I., Larrañaga, P., R. Etxeberria, B.S.: Feature subset selection by Bayesian networks based optimization. *Artificial Intelligence* **123**(1–2), 157–184 (2000)
15. Kadyrov, A., Petrou, M.: The trace transform and its applications. *IEEE Transactions on pattern analysis and machine intelligence* **23**(8), 811–828 (2001)
16. Lee, C.C., Mower, E., Busso, C., Lee, S., Narayanan, S.: Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication* **53**(9), 1162–1171 (2011)
17. Mendialdua, I., Martínez-Otzeta, J.M., Rodríguez-Rodríguez, I., Ruiz-Vazquez, T., Sierra, B.: Dynamic selection of the best base classifier in one versus one. *Knowledge-Based Systems* **85**, 298–306 (2015)

18. Mitchell, T.: *Machine Learning*. McGraw Hill (1997)
19. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden markov models. *Speech communication* **41**(4), 603–623 (2003)
20. Olaizola, I.G., Quartulli, M., Florez, J., Sierra, B.: Trace transform based method for color image domain identification. *IEEE Transactions on Multimedia* **16**(3), 679–685 (2014)
21. Pan, Y., Shen, P., Shen, L.: Speech emotion recognition using support vector machine. *International Journal of Smart Home* **6**(2), 101–107 (2012)
22. Pfister, T., Robinson, P.: Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis. *Affective Computing*, *IEEE Transactions on* **2**(2), 66–78 (2011)
23. Poppe, R.: A survey on vision-based human action recognition. *Image and vision computing* **28**(6), 976–990 (2010)
24. Quinlan, J.R.: *C4.5: programs for machine learning*. Elsevier (2014)
25. Ryoo, M., Chen, C.C., Aggarwal, J., Roy-Chowdhury, A.: An overview of contest on semantic description of human activities (sdha) 2010. In: *Recognizing Patterns in Signals, Speech, Images and Videos*, pp. 270–285. Springer (2010)
26. Schuller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M., Rigoll, G.: Speaker independent speech emotion recognition by ensemble classification. In: *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 864–867. IEEE (2005)
27. Schldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: *Proc. Int. Conf. Pattern Recognition (ICPR'04)*. Cambridge, U.K (2004)
28. Shahin, I.: Speaker identification in emotional talking environments based on CSPHMM2s. *Engineering Applications of Artificial Intelligence* **26**(7), 1652–1659. DOI 10.1016/j.engappai.2013.03.013. URL <http://linkinghub.elsevier.com/retrieve/pii/S0952197613000559>
29. Shao, L., Chen, X.: Histogram of body poses and spectral regression discriminant analysis for human action categorization. In: *Proceedings of the British Machine Vision Conference*, pp. 88.1–88.11. BMVA Press (2010). Doi:10.5244/C.24.88
30. Sobol-Shikler, T., Robinson, P.: Classification of complex information: Inference of co-occurring affective states from their expressions in speech. *Pattern Analysis and Machine Intelligence*, *IEEE Transactions on* **32**(7), 1284–1297 (2010)
31. Stone, M.: Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 111–147 (1974)
32. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding* **104**(23), 249 – 257 (2006). DOI <http://dx.doi.org/10.1016/j.cviu.2006.07.013>. URL <http://www.sciencedirect.com/science/article/pii/S1077314206001081>. Special Issue on Modeling People: Vision-based understanding of a persons shape, appearance, movement and behaviour
33. Yang, X., Tian, Y.: Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: *CVPR Workshops*, pp. 14–19. IEEE Computer Society (2012)
34. Zelaia, A., Arregi, O., Sierra, B.: Combining singular value decomposition and a multi-classifier: A new approach to support coreference resolution. *Eng. Appl. of AI* **46**, 279–286 (2015)
35. Zelnik-Manor, L., Irani, M.: Event-based analysis of video. In: *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 2, pp. II–123. IEEE (2001)
36. Ivarez, A., Sierra, B., Arruti, A., Lpez-Gil, J.M., Garay-Vitoria, N.: Classifier subset selection for the stacked generalization method applied to emotion recognition in speech. *Sensors* **16**(1), 21 (2016)

ONDORIOAK ETA ETORKIZUNERAKO LANA

3.1 Ondorioak

Ikerketa honetan lortutako ekarpenek konputagailu bidezko ikusmenaren barruan kokatuta dagoen irudi eta bideoen ulerpenaren zientzian aurrerapauso bat emateko balio izan dute. Lan honek aditzera eman duen bezala, diseinatu eta garatu diren metodo berriak alor desberdinetan aplikatu daitezke: ingurugiroa, segurtasuna, kirola, teledetekzioa, eta abar. Alor konkretu bateko edukiarekin lan egiteak, hau da, domeinu itxi eta zehatz baten barruan lan egiteak, irtenbidearen diseinu eta garapena arazoari moldatzea ahalbidetzen du eta ondorioz, emaitza askoz hobegoak lortzen dira.

Ikerketa lan hau I+G proiektuen esparruan garatu da. Honen ondorioz, egindako ikerketa industriari aplikagarriak diren irtenbideetara bideratu da batez ere. Honetaz gain, industriaren beharrak zeintzuk diren eta ikusmenak industriari eskaintzeko duenaz jabetzeko aukera ekarri du egoera honek. Berezitasun honek lortu diren ekarpenak eta jarraitutako bidea era batean moldatu ditu.

Ikerketaren lehen pausua, irudien behe-mailako ezaugarriak ateratzeko metodoen ezagupenean, konparaketan eta erabileran zetzan. Hauen bidez, posible da aplikazio errealetan agertzen diren arazo sinpleei erantzuna ematea. Baina iruditik atera nahi den informazioa handitzen den neurrian edota irudiaren konplexutasunaren arabera, metodo horiek ez dira nahikoak. Honen ondorioz, ikasketa automatikoko sailkatzaileen erabilera nahitaezkoa bihurtzen da. Lan honetan egiaztatu den bezala, sailkatzaileek asko hobetzen dute bai irudi eta bai bideo analisisian lortutako emaitzen doitasuna.

Bideoari dagokionez, argi dago etorkizunean erabiliko den oinarritzko edukia dela eta ondorioz, honen analitika, ikerketa munduan funtsezko izaten jarraituko du. Txosten honetan bideoen analisirako oinarritzko pausu batzuk aurkeztu dira: batetik bideoetan agertzen diren objektuen identifikaziorako aplikaziora moldatutako metodo desberdinak eta bestetik ekintzen sailkapenerako metodologia. Bideoen denbora

ezaugarriak bere baitan duen informazioaren garrantzia ustiari beharrekotako faktore bat dela ikasteko balio izan du aurrera eramandako ikerketak.

Argi dago alor honetan ikerketarako aukera asko daudela oraindik; txosten honetan aurkeztutakoa bide egokian emandako pausu batzuk dira bakarrik, baina beharrezko pausuak.

3.2 Etorkizunerako lana

Orokorki hitz eginda, konputagailu bidezko ikusmenaren barruan gaur egun dauden erronka handienak bi multzotan bana daitezke: batetik Internet, hau da, inongo zehaztapenik gabeko alor guztietako eduki multimedia eta bestetik, Big Data, hau da, eduki kopuru erraldoiak. Interneten kasuan, erronkarik handiena edukien heterogeneotasuna da, finkatutako domeinu gabeko edukiak hain zuzen ere. Big Dataren kasuan berriz, nahiz eta domeinua finkatuta egon daitekeen, datu kopuru erraldoiak mugatzen du irtenbide posibleen garapena.

Lan honetan aurkeztutako ekarpenak domeinu itxi batean agertzen diren konputagailu bidezko ikusmenaren aplikazioak dira. Txosten honetan aipatu den bezala Internet, domeinu irekiko neurrik gabeko arazo bat aurkezten du. Egungo ikerketa gehienak eduki honen domeinua finkatu eta analisia domeinuari moldatzen dituzten tekniketari oinarritzen dira. Beste ikerketa askok berriz, zuzenean irudien behe-mailako ezaugarrietatik irudien ezagutzarako metodoak aurkezten dituzte.

Bestetik, azken urteetan multimedia eduki kopuruaren hazkundera kontutan izanik, egun hain ezaguna den Big Data kontzeptuaren inguruan eman daitekeen ikerketa aipatu behar da. Arazoa ez da soilik, eduki horren ulerpenera lortzea, baizik eta eduki kopuru hori maneiatu, aztertu eta sailkatzeko erremintak diseinatu eta garatzea. Beraz, arazo honi aurre egiteko, ikerketa lan honetan garatutako metodoen egokitasuna aztertu behar da etorkizunean.

Etorkizun hurbilerako zehaztutako lanak honako hauek dira:

- Irudien sailkapenerako aurkeztu den desberdintasun faktore berria irisen irudiak erabiliz oso emaitza onak lortu direnez, . honen aplikazioa beste alor batzuetara zabaltzea.
- Bideoen denbora ezaugarriak bere baitan duen informazioa ateratzeko eta aztertze metodo berri baten garapena. Irudien analisirako baliagarriak diren ezaugarrien moldaketa informazioaren ulerpenerako.

- Domeinuaren identifikaziorako, irudien ezagutzarako eta etiketatu semantikorako erabili diren metodoen egokitasunaren azterketa eta hauen moldaketa.

Ikerketa bizirik dagoen izaki bat izanik, egunero agertzen dira lanean jarraitzeko bide ezberdinak. Lortutako emaitzek eta erronka berriek marrazten dute bide hori eta honek zehaztuko ditu etorkizunean eman beharreko hurrengo pausuak.

ERANSKINAK

I+G proiektuak

Aurretik aipatu den bezala, txosten honetan aurkeztutako ikerketa lana Vicomtech-
IK4n aurrera eramandako proiektuetan oinarritu da. Proiektu gehienek izaera
industrialak eragina izan du ikerketa prozesuan zehar hartutako erabakietan eta
baita lortutako emaitzetan ere. Hurrengo lerroetan ikerketa prozesu honen gako izan
diren proiektuak aurkezten dira. Beraien deskribapena eta helburuez gain, lortutako
emaitzak ere aditzera ematen dira ikerketa lan honekin izan duten harremanaren
adierazgarri gisa.

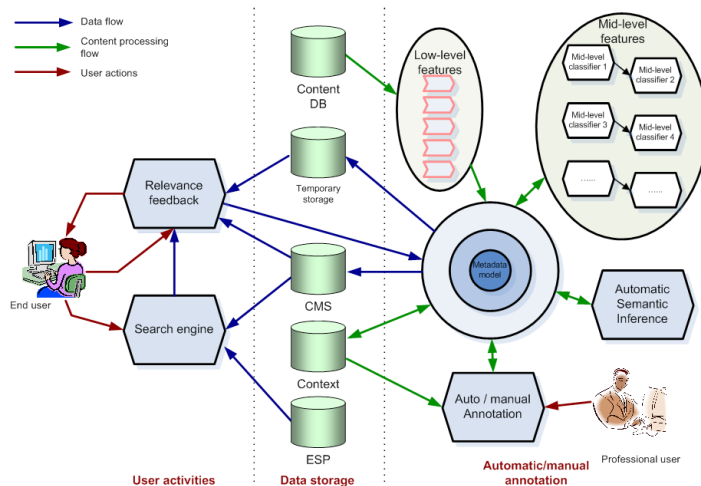
A.1 RUSHES

- **Izenburua:** Retrieval of multimedia semantic units for enhanced reusability
- **Kodea:** 045189
- **Erakunde finantzatzailea:** Europako Batzordea
- **Datak:** 2007-2009

Deskribapena eta helburuak

RUSHES proiektuaren helburu nagusia editatu gabeko eduki multimedia berres-
kuratzeko eta berrerabiltzeko sistema baten diseinua, ezarpena eta balioztatzea
da (ikus A.1 irudia). Horretarako beharrezkoa da ezagutza jatorrizko multimedia
edukitik ateratzea. Irudi prozesamenduaren kasuan, irudiaren behe-mailako analisia
egiten da deskriptore esanguratsuenak ateratzeko eta irudiaren ezagutza semantikoa
ondorioztatu ahal izateko.

Proiektu europarra izanik, alor bakoitzean iaioak diren ikerketa taldeekin lan egi,
partaide bakoitzaren ezagutzak bateratu eta sistema erreal batean martxan jartzeko
aukera eman du. *RUSHES* plataforma behe-mailako ezaugarrietan oinarritzen da
semantic-gap edo brexa semantikoari aurre egiteko.



Irudia A.1.: RUSHES plataformaren prozesuen fluxua

Lortutako emaitzak

Ikerketa lan honen barruan, eduki multimediaren azterketari eta ulerpenari dago-kionez hurrengo emaitzak lortu dira: natura eta ez-natura diren irudien bereizketa, bideo baten barruan elkarrizketak aurkitzea, bideo baten barruan eszena desberdinak identifikatu eta itsas-lerroaren identifikazioa helikopteroko bideoetatik abiatuz. Diseinatutako plataformaren arazo handiena kontzeptu berrien barnerapen prozesuan dago: kontzeptu berriak identifikatzeko orduan behe-mailako ezaugarri berriak definitu behar dira eta hau sistemaren eskalabilitatearen ikuspegitik ezinezkoa bihurtzen da.

A.2 SIAM

- **Izenburua:** Diseño y Desarrollo de un Sistema de Análisis Multimedia de Contenido Audiovisual en Plataformas Web Colaborativas
- **Enpresa finantzatzailea:** Hispavista
- **Datak:** 2009-2010

Deskribapena eta helburuak

Proiektu honen helburu nagusia multimedia edukiak aztertzeko erremintak garatzea da. Erreminta hauek erabiltzaileek sortutako eduki kantitate handiak ustiatzeko

aukera ematen dute. Etiketa semantikoak era automatikoan ezartzen zaizkie edukiei eta era honetan, edukien berrerabilpena ahalbidetzen da.

Lortutako emaitzak

SIAM proiektuak irudi eta bideo prozesaketan oinarritutako ikerketa aurrera eramatea ahalbidetu du. Bertan lortutako emaitzarik aipagarriena bideo bateko eszenak bereizteko algoritmoa izan da. Kasu honetan eszena, aldaketa bortizik gabeko bideo zati bat bezala ulertzen da. Eszena bakoitza identifikatu ondoren, hauek bideoaren azpi-domeinutzat har daitezke. Ondorioz, eszena bakoitzari dagozkion irudi multzoari irudi prozesamenduko algoritmo desberdinak aplikatzen zaizkio, halaber bakoitzari moldatutako algoritmoak. Horrela, askoz emaitza hobegoak lortzen dira.

A.3 GRAFEMA

- **Izenburua:** Multimedia edukien kudeaketarako sistema
- **Erakunde finantzatzailea:** Gipuzkoako Foru Aldundia
- **Datak:** 2012

Deskribapena eta helburuak

GRAFEMA proiektuaren helburu nagusia multimedia edukiak gorde, anotatu eta berreskuratzeko plataforma bat sortzea da. Kasu honetan, plataformaren arkitekturaren diseinua eta garapena izango dute garrantzia handiago eta ez edukitik informazioa ateratzeko algoritmoek.

Lortutako emaitzak

*GRAFEMA*ren plataformak multimedia edukien bilaketarako prozesu iteratiboek duten erabilgarritasuna probatzeko balio izan du. Prozesua laburbilduz, behe-mailako ezaugarriak ateratzen dira edukietatik (kasu honetan era askotako multimedia edukiak aztertu dira: testuak, irudiak, bideoak, eta abar), ezaugarri horien sailkapena egiten da ikasketa automatikoko metodoak erabiliz eta azkenik eredu semantiko bat sortzen da. Bilaketa egiteko orduan, eredu semantikoen arteko distantziak kalkulatu eta gertukoena aukeratzen da ontzat.

A.4 SKEYE

- **Izenburua:** Sistema de análisis meteorológico basado en imágenes del cielo tomadas desde tierra
- **Enpresa finantzatzailea:** Dominion
- **Datak:** 2007-2008

Deskribapena eta helburuak

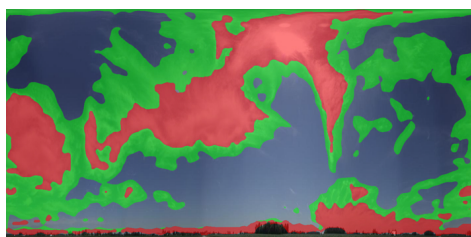
Gaur egun behatoki-meteorologikoez ematen duten informazio bakarra neurtzen dituzten balioak dira. *Skeye* proposatzen duen sistema, zeruaren irudietatik abiatuz, uneoro lainoen-estaldura faktorea automatikoki kalkulatzeko gai da. Bere helburua irudien behe-mailako informaziotik abiatuz adierazpen semantiko bat lortzea eta informazio hori sistema meteorologiko baten barruan kudeatzeko kapazitatea izatea da.

Lortutako emaitzak

Skeye proiektuan egindako irudi prozesaketak, irudiaren behe-mailako ezaugarrietatik, goi-mailako ulerpenerantz pauso bat ematea ahalbidetzen du. Horretarako beharrezkoa da arau simple batzuen ezarpena. Kasu honetan, pixelen kolorea eta testura aztertuz, zerua lau kategoriatan banatzen da: zerua, hodeiak, eguzkia eta zerua ez dena (ikusi A.2 irudia). Era honetan, klase bakoitzeko pixel kopurua kontutan izanda, lainoen-estaldura faktorea automatikoki kalkulatu du sistemak.



(a) Irudi originala



(b) Lainoen segmentazioa

Irudia A.2.: *Skeye* proiektuan lortutako zeru irudien segmentazioa

Horretaz gain, *Skeye* plataformaren bidez irudiaren zeruertza bilatzen du eta aurkitu ezean norabide horretan behe-lainoa dagoela adierazten du. Horretarako, alde aurretik eskuz zeruertza kokapena adierazi dio erabiltzaileak.

Proiektu honetan garatu diren algoritmoak domeinu honetako arazoei aurre egiteko definitu dira eta ondorioz ezin dira berrerabili beste irudi prozesamenduko arazoetan. Nahiz eta irudi prozesamenduko teknika orokorrak erabili, ezarritako balditzen ondorioz garatutako algoritmoak ez dute beste domeinu batetara egokitzeko aukera ematen.

A.5 SIRA

- **Izenburua:** Diseño y desarrollo de un sistema de reconocimiento de marcas comerciales en emisiones televisivas
- **Enpresa finantzatzailea:** Vilaumedia
- **Datak:** 2005-2007

Deskribapena eta helburuak

Proiektu honen helburu nagusia telebistako edukietan logoak aurkitzea da. Modu honetan, telebistako edukietan agertzen den publizitatearen kontaketa automatikoa egitea lortzen da. Garatutako algoritmoak logoek orokorrean izaten dituzten oinarritzko ezaugarrietan oinarritzen dira hauek detektatu ahal izateko. Behin logoak detektatu direnean, hauen jarraipena egin behar da bideoa osatzen duten fotograma guztietan.

Lortutako emaitzak

Nahiz eta bideoak izan proiektu honetan aztergai, hauek irudien jarraipen bat bezala kontsideratu dira. Hau da, ez da kontutan izan denbora aldagaia, baizik eta irudi bakoitza irudi independente bat balitz bezala kontsideratu da eta irudien prozesaketarako algoritmoak garatu dira. Logoen hurrengo ezaugarriak kontutan hartu dira: forma erregularrak, kolore deigarriak eta testua. Lortutako emaitzak egokiak izan dira aurretik finkatu diren helburuak kontutan izanik, hala ere, logoen jarraipenean egin ahal izateko eta oklusio kasuei aurre egiteko irtenbideak gehitzea falta da.

A.6 CAPER

- **Izenburua:** Collaborative information, Acquisition, Processing, Exploitation and Reporting for the prevention of organised crime

- **Kodea:** 261712
- **Erakunde finantzatzailea:** Europako Batzordea
- **Datak:** 2011-2014

Deskribapena eta helburuak

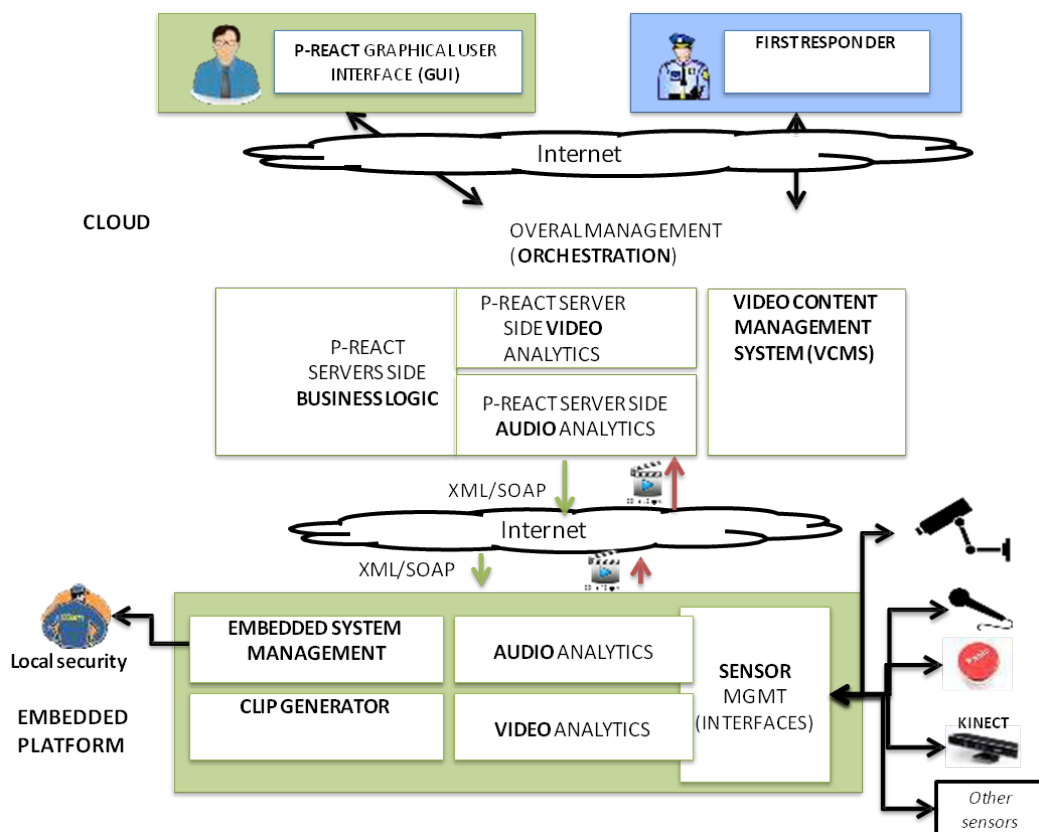
*CAPER*en helburu nagusia antolatutako krimena detektatu eta aurre hartzeko informazioa elkarbanatu ahal izateko plataforma komun bat sortzea da. Plataforma hau garatzeko beharrezkoak diren osagarri teknikoak multimedia edukiak prozesatzeko gai izan behar dira: zehatzago, audioa, testua, irudia, bideoa eta baita eduki biometrikoa ere aztertze algoritmoak barnean hartzeko kapazitatea behar du.

Lortutako emaitzak

CAPER proiektuaren barruan egindako ikerketa nagusia, irudien klasifikaziorako eta berrerabilpenerako metodoak garatzean oinarritu da. Alde batetik, pertsonen identifikaziorako algoritmoak garatu dira. Horretarako ezagutza biometrikoko algoritmoak garatu dira. Bestetik, irudietatik behe-mailako ezaugarri nagusienak atera dira, sistemak ezaugarri horien konparaketatik antzeko irudiak bilatzeko kapazitatea izateko. Azken ataza honetan lortutako emaitzak ez dira oso onak izan, domeinu oso zabala izanik, doitasun balio oso baxuak lortzen baitira.

A.7 P-REACT

- **Izenburua:** Petty cRiminality diminution through sEarch and Analysis in multi-source video Capturing and archiving plaTform
- **Kodea:** 607881
- **Erakunde finantzatzailea:** Europako Batzordea
- **Datak:** 2014-2016



Irudia A.3.: P-REACT plataforma

Deskribapena eta helburuak

Delitu txikiak egunerokotasuneko gauza dira gaur egun. Hauek ekiditeko P-REACT proiektuak negozio txikiak babesteko plataforma bat sortzea du helburua (ikusi A.3 irudia). Plataforma hau sarean dauden sentzore desberdinek (bideo kamerak, kamara terminoak, mikrofonoak) eta zentralizatutako sistema batek sortuko dute. Plataforma honen automatizaziorako beharrezkoak diren irudi eta bideo prozesaketa algoritmoak garatzea da helburu nagusienetakoa.

Lortutako emaitzak

Bideozaintzarako kamerek grabatutako bideoetatik abiatuta, bertan gertatzen diren ekintzak identifikatzeko metodoak garatu dira. Ekintzen detekzio automatiko honek kamara kokatuta dagoen lekuan gertatzen denaren berri ematen du eta ekintza susmagarri bat izatekotan, P-REACT plataformak alarma automatiko bat sortzen du. Bideoen azterketa automatiko honek, bideo zaintzaileen lana asko errazten du. Egindako lana aurretik finkatutako ekintza konkretu batzuk identifikatzean datza. Bideoen ekintza ezagutza oraindik mundu zientifikoan ebatzi gabe dagoen ataza bat da.

A.8 BEGIRA

- **Izenburua:** Diseño y desarrollo de un sistema seguimiento preciso de objetos en transmisiones deportivas
- **Enpresa finantzatzailea:** G93
- **Datak:** 2006-2009

Deskribapena eta helburuak

Begiraren helburua telebistatik zuzenean igortzen diren pilota partiduetan pilotaren ibilbidea eta bote puntua birtualki bistaritzen duen kamara bakarreko sistema sortzea da. Horretarako, bideoa denbora errealean prozesatu behar da bideoaren fotograma edo irudi bakoitzean pilotaren kokapena zein zehaztu ahal izateko. Irudi prozesaketaren barruan aurkitzen diren algoritmoak erabiltzen dira, bai pilota bideoan aurkitzeko bai honen jarraipena egiteko fotogrametan zehar.

Lortutako emaitzak

Begira proiektuan lortutako emaitzak behe-mailako irudi prozesaketa soilarekin lortu daitezkeen aurrerapenen adierazgarri dira. Irudi bakoitzaren azterketa egin ondoren, pilota segmentatu, bideoaren irudietan zehar bere ibilbidea jarraitu, lurreko marra aurkitu eta pilotaren bote puntua identifikatzeko kapaza den sistema eraiki da. Gainera, prozesaketa guztia ia denbora errealean egiten da. Aipatzekoa da *Begira* proiektuan lortutako emaitzak Euskal Irrati Telebistaren (EITB) pilota partiduen zuzeneko emanaldietan igorri egin zirela eta garatutako sistema argitalpen zientifikoetan argitaratzeaz gain, patente europar baten bidez babestuta dagoela.

Beste argitalpen batzuk

Ikerketa lan honekin lotura duten beste argitalpen batzuk:

- **Izenburua:** Local descriptors fusion for mobile iris verification
 - **Egileak:** Naiara Aginako, J.M. Martinez-Otzerta, Basilio Sierra, Modesto Castrillón-Santana, Javier Lorenzo-Navarro
 - **Proceedings:** Proceedings of International Conference on Pattern Recognition, ICPR 2016 (Cancún, México)
 - **Orriak:** 1-5
 - **Urtea:** 2016
 - **Laburpena:** *This paper summarizes the proposal submitted by the joint team conformed by researchers from UPV and ULPGC to the Mobile Iris CHallenge Evaluation II. The approach makes use of a state-of-the-art iris segmentation technique, to later extract features making use of local descriptors. Those suitable to the problem are selected after evaluating a collection of 15 local descriptors, covering a range of different grid configuration setups. A Machine Learning approach is used, learning a supervised classifier to deal with the descriptors data. A classifier is obtained for each descriptor, and the best ones are combined in a multi-classifier system. The final step fuses the classifier outputs obtained for 5 different local descriptors, to compute the dissimilarity measure for a pair of iris images.*

- **Izenburua:** Machine learning approach to dissimilarity computation: Iris matching
 - **Egileak:** Naiara Aginako, J.M. Martinez-Otzerta, Igor Rodriguez, Elena Lazkano, Basilio Sierra
 - **Proceedings:** Proceedings of International Conference on Pattern Recognition, ICPR 2016 (Cancún, México)
 - **Orriak:** 1-5
 - **Urtea:** 2016
 - **Laburpena:** *This paper presents a novel approach for iris dissimilarity computation based on Machine Learning paradigms and Computer Vision transformations. Based on the training dataset given by the MICHE II Cha-*

llenge organizers, a set of classifiers has been constructed and tested, aiming at classifying a single image. The main novelty of this paper remains in the used approach to iris dissimilarity computation: given two iris images, both of them are classified using the same paradigm, obtaining the a posteriori probability for each of the considered class values. Hence, two distributions are obtained, one for each iris image, and the dissimilarity is computed as the distance between these two distributions. Experimental results indicate the appropriateness of this new approach, even though more research and experiments are needed to obtain some improvements and to accelerate the classification process.

3.
 - **Izenburua:** Large scale thematic mapping by supervised machine learning on big data distributed cluster computing frameworks
 - **Egileak:** Javier Lozano, Naiara Aginako, Marco Quartulli, Igor G. Olaizola, Ekaitz Zulueta, Pedro Iriondo
 - **Proceedings:** Proceedings of 2015 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2015 (Milán, Italia)
 - **Orriak:** 1504-1507
 - **Argitaletxea:** IEEE
 - **Urtea:** 2015
 - **DOI:** <http://dx.doi.org/10.1109/IGARSS.2015.7326065>
 - **Laburpena:** *The Petabyte-scale data volumes in Earth Observation (EO) archives are not efficiently manageable with serial processes running on large isolated servers. Distributed storage and processing based on 'big data' cloud computing frameworks needs to be considered as a part of the solution.*
-

4.
 - **Izenburua:** Scalable Machine Learning for Fast Thematic Mapping in Web Servers
 - **Egileak:** Javier Lozano, Naiara Aginako, Marco Quartulli, Ekaitz Zulueta, Igor García Olaizola
 - **Proceedings:** Proceedings of the 2014 conference on Big Data from Space, BiDS14 (Frascati, Italia)
 - **Orriak:** 38-41
 - **Argitaletxea:** Publications Office of the European Union
 - **Urtea:** 2014
 - **DOI:** <http://dx.doi.org/10.2788/1823>

- **Laburpena:** *We present a web mining processing service that returns supervised probabilistic classifications of Earth Observation (EO) data in tiled form, with the aim to create user-selection based thematic maps from remotely sensed raster imagery. User interfaces supporting interactive navigation and model training and tuning are implemented in open HTML5 standards, while software interfaces among components conform to OGC standards. Near real time operation in the servers is attained by exploiting efficient data structures for high dimensional indexing and search.*
-

- 5.
- **Izenburua:** Multimedia analysis of video sources
 - **Egileak:** Juan Arraiza, Montse Cuadros, Naiara Aginako, Matteo Raffaelli, Olga Kaehm, Nasser Damer, J.P. Neto
 - **Proceedings:** 2014 International Conference on Signal Processing and Multimedia Applications, SIGMAP 2014 (Viena, Austria)
 - **Orriak:** 346-352
 - **Argitaletxea:** IEEE
 - **Urtea:** 2014
 - **DOI:** <http://ieeexplore.ieee.org/document/7514535/>
 - **Laburpena:** *Law Enforcement Agencies (LEAs) spend increasing efforts and resources on monitoring open sources, searching for suspicious behaviours and crime clues. The task of efficiently and effectively monitoring open sources is strongly linked to the capability of automatically retrieving and analyzing multimedia data. This paper presents a multimodal analytics system, created in cooperation with European LEAs. In particular it is described how the video analytics subsystem produces a workflow of multimedia data analysis processes. After a first analysis of video files, images are extracted in order to perform image comparison, classification and face recognition. In addition, audio content is extracted to perform speaker recognition and multilingual analysis of text transcripts. The integration of multimedia analysis results allows LEAs to extract pertinent knowledge from the gathered information.*
-

- 6.
- **Izenburua:** Image Analysis Platform for Data Management in the Meteorological Domain
 - **Egileak:** Igor G. Olaizola, Naiara Aginako, Mikel Labayen
 - **Proceedings:** 2009 Fourth International Workshop on Semantic Media Adaptation and Personalization, SMAP 2009 (Donostia, Spain)

- **Orriak:** 89-94
 - **Argitaletxea:** IEEE
 - **Urtea:** 2009
 - **DOI:** <http://dx.doi.org/10.1109/SMAP.2009.29>
 - **Laburpena:** *This paper proposes an architecture to provide semantic media information to the current existing meteorological models and prediction techniques. Satellite images have been used by meteorologists during the last 50 years, but we present a new method to take advantage of local images taken from the earth. Networked terrestrial weather stations can offer valuable image information, both of local and wide areas adding details that cannot be captured by satellites. Based on the results of two projects carried out together with the Basque Meteorology Agency (Euskalmet), we propose a method to port from image data to semantic meteorological information and an architecture to integrate the existing weather data and knowledge structures with multimedia semantics. The validation of the analysis system has been carried out using sky images taken in visual spectrum and the results have demonstrated the great potential of such platforms that could be extended to other data sources in order to apply multimedia semantic technologies in application fields like meteorology.*
-

- 7.
- **Izenburua:** COST292 experimental framework for TRECVID 2008
 - **Egileak:** Q. Zhang, G. Toliás, B. Mansencal, A. Saracoglu, N. Aginako, A. Alatan, L. A. Alexandre, Y. Avrithis, J. Benois-Pineau, K. Chandramouli, M. Corvaglia, U. Damnjanovic, A. Dimou, E. Esen, N. Fatemi, J. Goya
 - **Proceedings:** Proceedings of 6th TRECVID Workshop, TRECVID 2008 (Gaithersburg, USA)
 - **Orriak:** 1-15
 - **Urtea:** 2008
 - **Laburpena:** *In this paper, we give an overview of the four tasks submitted to TRECVID 2008 by COST292. The high-level feature extraction framework comprises four systems. The first system transforms a set of low-level descriptors into the semantic space using Latent Semantic Analysis and utilises neural networks for feature detection. The second system uses a multi-modal classifier based on SVMs and several descriptors. The third system uses three image classifiers based on ant colony optimisation, particle swarm optimisation and a multi-objective learning algorithm. The fourth system uses a Gaussian model for singing detection and a person detection algorithm. The search task is based on an interactive retrieval application combining retrieval functionalities in various modalities with a user interfa-*

ce supporting automatic and interactive search over all queries submitted. The rushes task submission is based on a spectral clustering approach for removing similar scenes based on eigenvalues of frame similarity matrix and a redundancy removal strategy which depends on semantic features extraction such as camera motion and faces. Finally, the submission to the copy detection task is conducted by two different systems. The first system consists of a video module and an audio module. The second system is based on mid-level features that are related to the temporal structure of videos.

Bibliografía

- [Aco85] Malcolm Acock. „Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. By David Marr“. In: *The Modern Schoolman* 62.2 (1985), pp. 141–142 (cit. on p. 1).
- [Agi+14] Naiara Aginako, Javier Lozano, Marco Quartulli, Basilio Sierra, and Igor G. Olaizola. „Identification of plant species on large botanical image datasets“. In: *1st International Workshop on Environmental Multimedia Retrieval co-located with ACM International Conference on Multimedia Retrieval (ICMR 2014)*. Vol. 1222. 2014, pp. 38–44 (cit. on pp. 19, 22, 72).
- [Agi+17a] Naiara Aginako, Goretti Echegaray, Igor G. Olaizola, Julian Florez, and Basilio Sierra. „Machine Learning for Video Action Recognition: a Computer Vision Approach“. In: *Machine Vision and Applications* (2017) (cit. on pp. 20, 21, 106).
- [Agi+17b] Naiara Aginako, J.M. Martinez-Otzerta, Igor Rodriguez, Elena Lazkano, and Basilio Sierra. „Iris matching by means of machine learning paradigms: a new approach to dissimilarity computation“. In: *Pattern Recognition Letters x.x* (2017), p. xx (cit. on pp. 19, 20, 71).
- [Agi+17c] Naiara Aginako, J.M. Martinez-Otzerta, Basilio Sierra, Modesto Castrillón-Santana, and Javier Lorenzo-Navarro. „Periocular and iris local descriptors for identity verification in mobile applications“. In: *Pattern Recognition Letters x.x* (2017), p. xx (cit. on pp. 19, 21, 72).
- [And+12] Felipe S. P. Andrade, Jurandy Almeida, Hélio Pedrini, and Ricardo da S.Torres. „Fusion of Local and Global Descriptors for Content-Based Image and Video Retrieval“. In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 17th Iberoamerican Congress, CIARP 2012, Buenos Aires, Argentina, September 3-6, 2012. Proceedings*. Ed. by Luis Alvarez, Marta Mejail, Luis Gomez, and Julio Jacobo. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 845–853 (cit. on p. 13).
- [ASC13] Jon Arrospeide, Luis Salgado, and Massimo Camplani. „Image-based on-road vehicle detection using cost-effective Histograms of Oriented Gradients“. In: *Journal of Visual Communication and Image Representation* 24.7 (2013), pp. 1182–1190 (cit. on p. 9).
- [BCM03] M. Bicego, U. Castellani, and V. Murino. „Using hidden Markov models and wavelets for face recognition“. In: *12th International Conference on Image Analysis and Processing, 2003.Proceedings*. Sept. 2003, pp. 52–56 (cit. on p. 15).

- [Bie87] Irving Biederman. „Recognition-by-components: a theory of human image understanding.“ In: *Psychological review* 94.2 (1987), p. 115 (cit. on p. 1).
- [BMP02] S. Belongie, J. Malik, and J. Puzicha. „Shape matching and object recognition using shape contexts“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.4 (Apr. 2002), pp. 509–522 (cit. on p. 13).
- [BT10] Alessandro Bergamo and Lorenzo Torresani. „Exploiting weakly-labeled Web images to improve object classification: a domain adaptation approach“. In: *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*. 2010, pp. 181–189 (cit. on p. 3).
- [BZM06] Anna Bosch, Andrew Zisserman, and Xavier Muñoz. „Scene Classification Via pLSA“. In: *Computer Vision – ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV*. Ed. by Aleš Leonardis, Horst Bischof, and Axel Pinz. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 517–530 (cit. on p. 3).
- [Car15] Carla Marshall. *By 2019, 80% of the World’s Internet Traffic Will Be Video [Cisco Study]*. https://people.cs.umass.edu/~elm/Teaching/Docs/IntroCV_1_19_11.pdf. Irakurrita: 2017-01-18. 2015 (cit. on p. 1).
- [Cow+01] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. „Emotion recognition in human-computer interaction“. In: *IEEE Signal Processing Magazine* 18.1 (Jan. 2001), pp. 32–80 (cit. on p. 16).
- [dAn+14] E. d’Angelo, L. Jacques, A. Alahi, and P. Vandergheynst. „From Bits to Images: Inversion of Local Binary Descriptors“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.5 (May 2014), pp. 874–887 (cit. on p. 8).
- [Der90] R. Deriche. „Fast algorithms for low-level vision“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.1 (Jan. 1990), pp. 78–87 (cit. on p. 3).
- [DKN04] Thomas Deselaers, Daniel Keysers, and Hermann Ney. „Features for image retrieval: A quantitative comparison“. In: *Joint Pattern Recognition Symposium*. Springer. 2004, pp. 228–236 (cit. on p. 13).
- [DKN08] Thomas Deselaers, Daniel Keysers, and Hermann Ney. „Features for image retrieval: an experimental comparison“. In: *Information retrieval* 11.2 (2008), pp. 77–107 (cit. on p. 13).
- [DT05] N. Dalal and B. Triggs. „Histograms of oriented gradients for human detection“. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*. Vol. 1. June 2005, 886–893 vol. 1 (cit. on p. 9).
- [Eis14] Mohamed Eisa. „Combined Local and Global Features for Improving the Shape Retrieval“. In: *International Journal of Computer Science Issues (IJCSI)* 11.3 (2014), p. 12 (cit. on p. 13).
- [Eri11] Erik G. Learned-Miller. *Introduction to Computer Vision*. <http://tubularinsights.com/2019-internet-video-traffic/>. Irakurrita: 2017-01-18. 2011 (cit. on p. 7).

- [FPZ03] Robert Fergus, Pietro Perona, and Andrew Zisserman. „Object class recognition by unsupervised scale-invariant learning“. In: *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*. Vol. 2. IEEE. 2003, pp. II–II (cit. on p. 13).
- [GFT98] Bilge Günsel, A. Müfit Ferman, and A. Murat Tekalp. „Temporal video segmentation using unsupervised clustering and semantic object tracking“. In: *Journal of Electronic Imaging* 7.3 (1998), pp. 592–604 (cit. on p. 15).
- [GJB] Hervé Goeau, Alexis Joly, and Pierre Bonnet. *LifeCLEF 2014- Plant retrieval task*. <http://www.imageclef.org/2014/lifeclef/plant> (cit. on p. 72).
- [GN08] P Geetha and Vasumathi Narayanan. „A survey of content-based video retrieval“. In: (2008) (cit. on p. 16).
- [Guo+14] Yulan Guo, Mohammed Bennamoun, Ferdous Sohel, Min Lu, and Jianwei Wan. „3d object recognition in cluttered scenes with local surface features: A survey“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.11 (2014), pp. 2270–2287 (cit. on p. 13).
- [GZZ10] Zhenhua Guo, Lei Zhang, and David Zhang. „Rotation invariant texture classification using {LBP} variance (LBPV) with global matching“. In: *Pattern Recognition* 43.3 (2010), pp. 706–719 (cit. on p. 9).
- [HTZ08] Xuelong Hu, Yingcheng Tang, and Zhenghua Zhang. „Video object matching based on SIFT algorithm“. In: *2008 International Conference on Neural Networks and Signal Processing*. June 2008, pp. 412–415 (cit. on p. 9).
- [Hu+04] Weiming Hu, Tieniu Tan, Liang Wang, and S. Maybank. „A survey on visual surveillance of object motion and behaviors“. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 34.3 (Aug. 2004), pp. 334–352 (cit. on p. 16).
- [Hu+15] Guosheng Hu, Yongxin Yang, Dong Yi, Josef Kittler, William Christmas, Stan Z. Li, and Timothy Hospedales. „When Face Recognition Meets With Deep Learning: An Evaluation of Convolutional Neural Networks for Face Recognition“. In: *The IEEE International Conference on Computer Vision (ICCV) Workshops*. Dec. 2015 (cit. on p. 15).
- [HVL08] Niels Haering, Péter L. Venetianer, and Alan Lipton. „The evolution of video surveillance: an overview“. In: *Machine Vision and Applications* 19.5 (2008), pp. 279–290 (cit. on p. 16).
- [ICP16] ICPR16. *23rd International Conference on Pattern Recognition (ICPR2016)*. <http://www.icpr2016.org>. Dec. 8, 2016 (cit. on p. 72).
- [Jin+03] Feng Jing, Mingjing Li, Lei Zhang, Hong-Jiang Zhang, and Bo Zhang. „Learning in region-based image retrieval“. In: *International Conference on Image and Video Retrieval*. Springer. 2003, pp. 206–215 (cit. on p. 4).
- [Jon+02] K. Jonsson, J. Kittler, Y.P. Li, and J. Matas. „Support vector machines for face authentication“. In: *Image and Vision Computing* 20.5–6 (2002), pp. 369–375 (cit. on p. 15).
- [Jos16] Josh James. *Data Never Sleeps 4.0*. <https://www.domo.com/blog/data-never-sleeps-4-0/>. Irakurrita: 2017-01-26. 2016 (cit. on p. 2).

- [JS07] Alejandro Jaimes and Nicu Sebe. „Multimodal human–computer interaction: A survey“. In: *Computer Vision and Image Understanding* 108.1–2 (2007). Special Issue on Vision for Human-Computer Interaction, pp. 116–134 (cit. on p. 16).
- [Kim+10] In Su Kim, Hong Seok Choi, Kwang Moo Yi, Jin Young Choi, and Seong G. Kong. „Intelligent visual surveillance — A survey“. In: *International Journal of Control, Automation and Systems* 8.5 (2010), pp. 926–939 (cit. on p. 16).
- [KLH02] Sanjiv Kumar, Alexander C Loui, and Martial Hebert. „Probabilistic classification of image regions using an observation-constrained generative approach“. In: (2002) (cit. on p. 14).
- [Kne+98] Stefan Knerr, Emmanuel Augustin, Olivier Baret, and David Price. „Hidden Markov model based word recognition and its application to legal amount reading on French checks“. In: *Computer Vision and Image Understanding* 70.3 (1998), pp. 404–419 (cit. on p. 15).
- [Lab+14] Mikel Labayen, Igor G Olaizola, Naiara Aginako, and Julián Flórez. „Accurate ball trajectory tracking and 3D visualization for computer-assisted sports broadcast“. In: *Multimedia tools and applications* 73.3 (2014), pp. 1819–1842 (cit. on pp. 20, 21, 105).
- [LAO08] Mikel Labayen, Naiara Aginako, and Igor G. Olaizola. „Weather analysis system based on sky images taken from the Earth“. In: *IET Conference Proceedings*. Institution of Engineering and Technology, Jan. 2008, 146–151(5) (cit. on pp. 19, 22, 26).
- [Law+97] S. Lawrence, C. L. Giles, Ah Chung Tsoi, and A. D. Back. „Face recognition: a convolutional neural-network approach“. In: *IEEE Transactions on Neural Networks* 8.1 (Jan. 1997), pp. 98–113 (cit. on p. 15).
- [LC03] Xiaoming Liu and Tsuhan Cheng. „Video-based face recognition using adaptive hidden Markov models“. In: *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings*. Vol. 1. June 2003, pp. 340–345 (cit. on p. 15).
- [Liu+07] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. „A survey of content-based image retrieval with high-level semantics“. In: *Pattern Recognition* 40.1 (2007), pp. 262–282 (cit. on p. 3).
- [LK06] A.C. Loui and S. Kumar. *Method for image region classification using unsupervised and supervised learning*. US Patent 7,039,239. May 2006 (cit. on p. 14).
- [LMH04] Andreas D. Lattner, Andrea Miene, and Otthein Herzog. „A Combination of Machine Learning and Image Processing Technologies for the Classification of Image Regions“. In: *Adaptive Multimedia Retrieval: First International Workshop, AMR 2003, Hamburg, Germany, September 15-16, 2003, Revised Selected and Invited Papers*. Ed. by Andreas Nürnberger and Marcin Detyniecki. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 185–199 (cit. on p. 14).
- [LMP04] Riccardo Leonardi, Pierangelo Migliorati, and Maria Prandini. „Semantic indexing of soccer audio-visual sequences: a multimodal approach based on controlled Markov chains“. In: *IEEE Transactions on Circuits and Systems for Video Technology* 14.5 (2004), pp. 634–643 (cit. on p. 15).

- [Low04] David G. Lowe. „Distinctive Image Features from Scale-Invariant Keypoints“. In: *International Journal of Computer Vision* 60.2 (2004), pp. 91–110 (cit. on p. 13).
- [Low99] D. G. Lowe. „Object recognition from local scale-invariant features“. In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*. Vol. 2. 1999, 1150–1157 vol.2 (cit. on p. 9).
- [Loz+15a] Javier Lozano, Naiara Aginako, Marco Quartulli, Igor G. Olaizola, and Ekaitz Zulueta. „Web-Based Supervised Thematic Mapping“. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8.5 (May 2015), pp. 2165–2176 (cit. on pp. 18, 20).
- [Loz+15b] Javier Lozano, Naiara Aginako, Marco Quartulli, Igor G. Olaizola, Ekaitz Zulueta, and Pedro Iriondo. „Large scale thematic mapping by supervised machine learning on big data distributed cluster computing frameworks“. In: *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. 2015, pp. 1504–1507 (cit. on p. 25).
- [LSP06a] S. Lazebnik, C. Schmid, and J. Ponce. „Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories“. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 2. 2006, pp. 2169–2178 (cit. on p. 3).
- [LSP06b] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. „Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories“. In: *Computer vision and pattern recognition, 2006 IEEE computer society conference on*. Vol. 2. IEEE. 2006, pp. 2169–2178 (cit. on p. 13).
- [MAJ13] Anand Mishra, Karteek Alahari, and C.V. Jawahar. „Image Retrieval Using Textual Cues“. In: *The IEEE International Conference on Computer Vision (ICCV)*. Dec. 2013 (cit. on p. 4).
- [MIC] MICHE. *Mobile Iris CHallenge Evaluation II (MICHE II)*. http://biplab.unisa.it/MICHE_Contest_ICPR2016/index.php (cit. on p. 72).
- [Mit97] Tom M Mitchell. „Does machine learning really work?“. In: *AI magazine* 18.3 (1997), p. 11 (cit. on p. 10).
- [MKS03] Vasileios Mezaris, Ioannis Kompatsiaris, and Michael G Strintzis. „An ontology approach to object-based image retrieval“. In: *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*. Vol. 2. IEEE. 2003, pp. II–511 (cit. on p. 4).
- [MLS05] Krystian Mikolajczyk, Bastian Leibe, and Bernt Schiele. „Local features for object class recognition“. In: *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*. Vol. 2. IEEE. 2005, pp. 1792–1799 (cit. on p. 13).
- [Nac+08] S. U. Naci, Uros Damnjanovic, Boris Mansencal, Jenny Benois-Pineau, Christian Kaes, Marzia Corvaglia, Eliana Rossi, and Naiara Aginako. „The COST292 Experimental Framework for Rushes Summarization Task in TRECVID 2008“. In: *Proceedings of the 2Nd ACM TRECVID Video Summarization Workshop*. TVS '08. Vancouver, British Columbia, Canada: ACM, 2008, pp. 40–44 (cit. on pp. 19, 21, 26).

- [NH98] A. V. Nefian and M. H. Hayes. „Hidden Markov models for face recognition“. In: *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*. Vol. 5. May 1998, 2721–2724 vol.5 (cit. on p. 15).
- [NIS] NIST. *TREC Video Retrieval Evaluation*. <http://trecvid.nist.gov/> (cit. on p. 26).
- [NJT06] Eric Nowak, Frédéric Jurie, and Bill Triggs. „Sampling Strategies for Bag-of-features Image Classification“. In: *Proceedings of the 9th European Conference on Computer Vision - Volume Part IV. ECCV'06*. Graz, Austria: Springer-Verlag, 2006, pp. 490–503 (cit. on p. 13).
- [NL84] A. M. Nazif and M. D. Levine. „Low Level Image Segmentation: An Expert System“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6.5* (Sept. 1984), pp. 555–577 (cit. on p. 3).
- [OAL12] Igor G. Olaizola, Naiara Aginako, and Mikel Labayen. *Method of detection and recognition of logos in a video stream*. European Patent(PCT). ES2395448 (T3). 2013-02-12 (cit. on pp. 20, 21, 105).
- [OD11] Stephen O'Hara and Bruce A. Draper. „Introduction to the Bag of Features Paradigm for Image Classification and Retrieval“. In: *CoRR abs/1101.3354* (2011) (cit. on p. 13).
- [Ola+08] Igor G. Olaizola, Julián Flórez, J.C. San Román, Naiara Aginako, and Mikel Labayen. *Method for detecting the point of impact of a ball in sports events*. European Patent(PCT). ES2402728 (T3). 2013-05-08 (cit. on pp. 20, 21, 105).
- [Ola+14] Igor G Olaizola, Marco Quartulli, Julian Florez, and Basilio Sierra. „Trace transform based method for color image domain identification“. In: *IEEE Transactions on Multimedia* 16.3 (2014), pp. 679–685 (cit. on p. 3).
- [Ola12] Igor G. Olaizola. „A framework for content based semantic information extraction from multimedia contents“. PhD thesis. Computer Science and Artificial Intelligence department. Computer Science Faculty. University of the Basque Country (EHU-UPV), 2012 (cit. on p. 6).
- [Oli+95] C. Olivier, T. Paquet, M. Avila, and Y. Lecourtier. „Recognition of handwritten words using stochastic models“. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 1. Aug. 1995, 19–23 vol.1 (cit. on p. 15).
- [OT01] Aude Oliva and Antonio Torralba. „Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope“. In: *International Journal of Computer Vision* 42.3 (2001), pp. 145–175 (cit. on p. 3).
- [Pet00] Milan Petkovic. „Content-based video retrieval“. In: (2000) (cit. on p. 16).
- [PG17] Luca Piras and Giorgio Giacinto. „Information Fusion in Content Based Image Retrieval: A Comprehensive Overview“. In: *Information Fusion* (2017) (cit. on p. 4).
- [Pha+14] V. Pham, T. Bluche, C. Kermorvant, and J. Louradour. „Dropout Improves Recurrent Neural Networks for Handwriting Recognition“. In: *2014 14th International Conference on Frontiers in Handwriting Recognition*. Sept. 2014, pp. 285–290 (cit. on p. 15).

- [Phi+98] P Jonathon Phillips et al. „Support vector machines applied to face recognition“. In: (1998) (cit. on p. 15).
- [PV13] K Poulouse Jacob and ER Vimina. „Content based image retrieval using low level features of automatically extracted regions of interest“. In: (2013) (cit. on p. 13).
- [PW12] Oluwatoyin P Popoola and Kejun Wang. „Video-based abnormal human behavior recognition—A review“. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.6 (2012), pp. 865–878 (cit. on p. 16).
- [Qia+16] X. Qian, X. Tan, Y. Zhang, R. Hong, and M. Wang. „Enhancing Sketch-Based Image Retrieval by Re-Ranking and Relevance Feedback“. In: *IEEE Transactions on Image Processing* 25.1 (Jan. 2016), pp. 195–208 (cit. on p. 4).
- [RA15] Siddharth S. Rautaray and Anupam Agrawal. „Vision based hand gesture recognition for human computer interaction: a survey“. In: *Artificial Intelligence Review* 43.1 (2015), pp. 1–54 (cit. on p. 16).
- [RBK98] H. A. Rowley, S. Baluja, and T. Kanade. „Neural network-based face detection“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.1 (Jan. 1998), pp. 23–38 (cit. on p. 15).
- [RM16] Research and Markets. *Video Analytics Market - Global Forecast to 2021*. <http://www.researchandmarkets.com/>. 2016 (cit. on p. 16).
- [RV16] M. Ravinder and T. Venugopal. „Content-Based Cricket Video Shot Classification Using Bag-of-Visual-Features“. In: *Artificial Intelligence and Evolutionary Computations in Engineering Systems: Proceedings of ICAIECES 2015*. Ed. by Subhransu Sekhar Dash, M. Arun Bhaskar, Bijaya Ketan Panigrahi, and Swagatam Das. New Delhi: Springer India, 2016, pp. 599–606 (cit. on p. 13).
- [Sam59] Arthur L Samuel. „Some studies in machine learning using the game of checkers“. In: *IBM Journal of research and development* 3.3 (1959), pp. 210–229 (cit. on p. 10).
- [Scl+99] Stan Sclaroff, Marco La Cascia, Saratendu Sethi, and Leonid Taycher. „Unifying Textual and Visual Cues for Content-Based Image Retrieval on the World Wide Web“. In: *Computer Vision and Image Understanding* 75.1 (1999), pp. 86–98 (cit. on p. 4).
- [SDN16] D. Suryani, P. Doetsch, and H. Ney. „On the Benefits of Convolutional Neural Network Combinations in Offline Handwriting Recognition“. In: *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. Oct. 2016, pp. 193–198 (cit. on p. 15).
- [Seg+09] Alvaro Segura, Aitor Moreno, Igor G. Olaizola, Naiara Aginako, Mikel Labayen, Jorge Posada, Jose Antonio Aranda, and Rubén García De Andoin. „Visual Processing of Geographic and Environmental Information in the Basque Country: Two Basque Case Studies“. In: *GeoSpatial Visual Analytics: Geographical Information Processing and Visual Analytics for Environmental Security*. Ed. by Raffaele De Amicis, Radovan Stojanovic, and Giuseppe Conti. Dordrecht: Springer Netherlands, 2009, pp. 199–207 (cit. on pp. 19, 21, 26).
- [SH94] F. S. Samaria and A. C. Harter. „Parameterisation of a stochastic model for human face identification“. In: *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*. Dec. 1994, pp. 138–142 (cit. on p. 15).

- [Sha] Roman Shapovalov. *Computer Blindness*. <http://computerblindness.blogspot.com.es/2010/06/object-detection-vs-semantic.html> (cit. on p. 14).
- [SHS10] I. Saleemi, L. Hartung, and M. Shah. „Scene understanding by statistical modeling of motion patterns“. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. June 2010, pp. 2069–2076 (cit. on p. 3).
- [SM97] Cordelia Schmid and Roger Mohr. „Local grayvalue invariants for image retrieval“. In: *IEEE transactions on pattern analysis and machine intelligence* 19.5 (1997), pp. 530–535 (cit. on p. 13).
- [SSZ06] Josef Sivic, Frederik Schaffalitzky, and Andrew Zisserman. „Object Level Grouping for Video Shots“. In: *International Journal of Computer Vision* 67.2 (2006), pp. 189–210 (cit. on p. 13).
- [SZ94] S. W. Smoliar and HongJiang Zhang. „Content based video indexing and retrieval“. In: *IEEE MultiMedia* 1.2 (Summer 1994), pp. 62–72 (cit. on p. 16).
- [Sze10] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010 (cit. on p. 1, 9).
- [Tao+07] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. „General tensor discriminant analysis and gabor features for gait recognition“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.10 (2007) (cit. on p. 13).
- [TG80] Anne M. Treisman and Garry Gelade. „A feature-integration theory of attention“. In: *Cognitive psychology* 12.1 (1980), pp. 97–136 (cit. on p. 1).
- [TM08] Tinne Tuytelaars and Krystian Mikolajczyk. „Local Invariant Feature Detectors: A Survey“. In: *Foundations and Trends® in Computer Graphics and Vision* 3.3 (2008), pp. 177–280 (cit. on p. 9).
- [TV04] Tinne Tuytelaars and Luc Van Gool. „Matching Widely Separated Views Based on Affine Invariant Regions“. In: *International Journal of Computer Vision* 59.1 (2004), pp. 61–85 (cit. on p. 13).
- [TZ04] W. Tavanapong and Junyu Zhou. „Shot clustering techniques for story browsing“. In: *IEEE Transactions on Multimedia* 6.4 (Aug. 2004), pp. 517–527 (cit. on p. 15).
- [VVV88] P.W Verbeek, H.A Vrooman, and L.J Van Vliet. „Low-level image processing by max-min filters“. In: *Signal Processing* 15.3 (1988), pp. 249–258 (cit. on p. 3).
- [Wan+11] Yue Wang, ZuJun Hou, Karianto Leman, Nam Trung Pham, TeckWee Chua, and Richard Chang. „Combination of Local and Global Features for Near-duplicate Detection“. In: *Proceedings of the 17th International Conference on Advances in Multimedia Modeling - Volume Part I. MMM'11*. Taipei, Taiwan: Springer-Verlag, 2011, pp. 328–338 (cit. on p. 13).
- [Wan+16] J. Wang, J. Zheng, S. Zhang, J. He, X. Liang, and S. Feng. „A Face Recognition System Based on Local Binary Patterns and Support Vector Machine for Home Security Service Robot“. In: *2016 9th International Symposium on Computational Intelligence and Design (ISCID)*. Vol. 2. Dec. 2016, pp. 303–307 (cit. on p. 15).
- [WW09] Y. Wu and Y. Wu. „Shape-Based Image Retrieval Using Combining Global and Local Shape Features“. In: *2009 2nd International Congress on Image and Signal Processing*. Oct. 2009, pp. 1–5 (cit. on p. 13).

- [Yin+07] P. Yin, A. Criminisi, J. Winn, and I. Essa. „Tree-based Classifiers for Bilayer Video Segmentation“. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. June 2007, pp. 1–8 (cit. on p. 15).
- [YYD15] Joe Yue-Hei Ng, Fan Yang, and Larry S Davis. „Exploiting local features from deep networks for image retrieval“. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2015, pp. 53–61 (cit. on p. 13).
- [YYL98] Minerva Yeung, Boon-Lock Yeo, and Bede Liu. „Segmentation of Video by Clustering and Graph Analysis“. In: *Computer Vision and Image Understanding* 71.1 (1998), pp. 94–109 (cit. on p. 15).
- [Zam+14] F. Zamora-Martinez, V. Frinken, S. España-Boquera, M.J. Castro-Bleda, A. Fischer, and H. Bunke. „Neural network language models for off-line handwriting recognition“. In: *Pattern Recognition* 47.4 (2014), pp. 1642–1652 (cit. on p. 15).
- [ZH03] Xiang Sean Zhou and Thomas S. Huang. „Relevance feedback in image retrieval: A comprehensive review“, journal="Multimedia Systems". In: 8.6 (2003), pp. 536–544 (cit. on p. 4).
- [ZL02] Yu Jin Zhang and HB Lu. „A hierarchical organization scheme for video data“. In: *Pattern Recognition* 35.11 (2002), pp. 2381–2387 (cit. on p. 15).
- [ZZZ16] Bei Zhao, Yanfei Zhong, and Liangpei Zhang. „A spectral–structural bag-of-features scene classifier for very high spatial resolution remote sensing imagery“. In: *{ISPRS} Journal of Photogrammetry and Remote Sensing* 116 (2016), pp. 73–85 (cit. on p. 13).

