Neurocomputing 500 (2022) 518–527

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Deep transfer learning-based gaze tracking for behavioral activity recognition

Javier de Lope [a,*], Manuel Graña [b]

[a] Department of Artificial Intelligence, Universidad Politécnica de Madrid (UPM), Madrid, Spain
[b] Computational Intelligence Group, University of the Basque Country (UPV/EHU), San Sebastian, Spain

ARTICLE INFO

Article history:
Received 7 March 2021
Revised 8 June 2021
Accepted 25 June 2021
Available online 23 May 2022

Keywords:
Deep transfer learning
Gaze tracking
Gaze ethogram
Human activity recognition

ABSTRACT

Computational Ethology studies focused on human beings is usually referred as Human Activity Recognition (HAR). Specifically, this paper belongs to a line of work on the identification of broad cognitive activities that users carry out with computers. The keystone of this kind of systems is the noninvasive detection of the subject's gaze fixations in selected display areas. Noninvasiveness is ensured by using the conventional laptop cameras without additional illumination or tracking devices. The gaze ethograms, composed as sequences of gaze fixations, are the basis to identify the user activities. To determine the gaze fixation display areas with the highest accuracy, this paper explores the use of a transfer learning approach applied to several well-known deep learning network (DLN) architectures whose input is the eye area extracted from the face image,and output is the identification of the gaze fixation area in the computer screen. Two different datasets are created and used in the validation experiments. We report encouraging results that may allow the general use of the system.

1. Introduction

Computational Ethology [1] has become a hot research field in the last few years. It integrates the information from several different sensors and activity measurement devices in order to characterize the behavior of living beings. Specifically, the computer-based analysis and recognition of human behavior, referred to as Human Activity Recognition (HAR) [2], receives plenty of attention and contributions. Basically, there are two types of sensors used in HAR research: cameras [3] and inertial sensors [4]. In computational neuroethology, these sensors are usually combined with neuronal activity data captured by using, for example, electroencephalography (EEG) equipment [5]. Much effort in HAR research is currently directed to the monitoring of aging people [14], and to the performance improvement in some sports [15]. The monitoring of elderly people is usually motivated by behavioral decline due to neurodegenerative diseases and its goals is to detect abnormal situations to raise alarms [6], for example, fall detection [16]. HAR studies are usually oriented to the identification of low level activities, for instance, the detection of abnormal behavioral situations in the elderly [6] by the use of 3D skeleton models of body postures [7], hence they do not deal with higher level behavior representations such as provided by ethograms.

An ethogram is a time plot of the low level actions carried out by the subject under observation that provides a high level behavioral representation. Ethograms have been used for animal phenotype characterization [8]. We are currently interested in the characterization of behavioral states of a laptop computer user by using the laptop camera and the microphone to determine the activity performed by the user by noninvasive computational methods. Previously, we have studied the performance of conventional machine learning approaches on such task [9]. In this paper we explore the use of deep learning techniques to recognize the subject's behavioral activity. Our hypothesis is that the subject's gaze fixations information allows to determine the specific activities in which the subject is engaged [9,61].

A *visual fixation* is the sustained gaze during a time interval in a specific direction which falls upon a single location in the visual stimulus. Its average duration in uncontrolled conditions is about 200 ms [12]. The *saccades* are quick, simultaneous movements of both eyes between two or more phases of fixation in the same direction [13]. *Blinking* is the semi-automatic rapid closing of the eyelids. Its rate is generally greater than a dozen blinks per minute, although it may decrease when the eyes are focused on an object for an extended period of time, for example, when reading.

[*] Corresponding author.
E-mail addresses: javier.delope@upm.es (J. de Lope), manuel.grana@ehu.eus (M. Graña).

https://doi.org/10.1016/j.neucom.2021.06.100
0925-2312/© 2022 The Author(s). Published by Elsevier B.V.

As the information is retrieved during the fixations, we determine when they are produced and in which order they are performed. We call *gaze ethograms* these temporal sequences of visual fixations which are the atomic actions building up the behavioral representation. We define areas of the display which receives the user attention in order to categorize the visual fixations. The gaze ethogram may be used to recognize the subject's behavioral activity. The work in this paper is devoted to the evaluation of deep learning architectures on the task of recognizing the gaze fixation from unfiltered images of the eye region.

The rest of the paper is organized as follows. First, Section 2 provides a short view of the state-of-art in both lines of our work: gaze detection and tracking, and deep learning techniques applied to life sciences. Section 3 describes the experimental datasets and the proposed computational methods experimented with. Section 4 provides the experimental results. There, we also offer a critical discussion on the results. Finally, Section 5 we conclude with some summarizing remarks on our work and directions for future work.

2. Background

This section provides a short review of the state-of-art in related research. First, we summarize the works on gaze detection and tracking with approximate or equivalent goals. Then, we review some antecedents and current developments in the growing area of deep learning applied to life sciences.

2.1. Gaze detection and tracking

Gaze information has been used for diagnostic and active interaction purposes [10,11,18]. For example, gaze interaction has been used for communication with people suffering extreme disability [24] or in patients with Alzheimer's Disease (AD) [25]. Diagnostic applications have been widespread in many different areas such as neuroscience [26,27], influence of students' visual attention and school failure [28] or analysis of facial expression exploration in subjects with social anxiety [29].

Gaze detection has been a research challenge for a long time [17,18]. Early successful approaches [19] were based on electrooculography (EOG), which is a technique that uses a series of electrodes situated in the user's face to measure the eye motion in an electromagnetic field. Videooculography (VOG) systems [20] are optical-based systems using specific illumination systems —often infrared— that enhance the detection of eye features such as the pupil and the cornea.

There is a need for much less invasive systems that do not require the subject to wear specific intrusive technology, as is the case of EOG and VOG. Solutions based on computer vision use conventional machine learning techniques, some are based on the localization of the eyebrows [21], others use the estimation of the 3D face motion from a single camera [22]. Recent approaches based on deep learning architectures have been tested in neuroscience studies [23]. The work in this paper goes in this direction towards minimally invasive reliable gaze detection and tracking systems.

2.2. Deep learning in the life sciences

Deep Learning (DL) approaches are the protagonist of Artificial Neural Network (ANN) resurgence in the last decade [31–33]. They overcome the problem of the vanishing gradient and overfitting by various approaches. They produce a data driven hierarchy of abstract representations by stochastic gradient descent training procedures. Specifically, the convolutional neural network (CNN) [35] training produces a hierarchy of filters tuned from the data. CNNs have been extremely influential in the advance of computer vision (CV) tasks. This architecture has inspired new generations of DL networks (DLN) with diverse architectures, which are reporting superior performance on many different problems in areas such as image processing [36,37], pattern recognition and object detection [38–40], classification [41,42], tracking [48], and activity recognition from data provided by inertial sensors [49].

In the Life Sciences (LS) the number of reported DLN applications during the last five years have been growing exponentially [34]. Example applications of DLN in LS areas are the analysis of medical images in the neurosciences [23,43] and other medical areas [44,45] including early stage detection of COVID-19 in X-ray imaging [46,47]. DLNs have been also applied to facial image processing, which is a rather complex object because of many different factors like the face position and orientation, the mouth and eyes opening, and the human skin color range. There are reported DLN approaches for face contour detection [50,51], the facial components extraction [52,53], biometric facial recognition [54], and gender classification [55].

3. Materials and methods

As previously stated, our setup employs the behavioral activity recognition system to determine the activity carried out by the subject [9]. This system uses gaze ethograms to describe and identify such activities. Fig. 1(a) shows an instance of a gaze ethogram obtained from a subject reading a text on the computer display. For activity recognition purposes it is enough that the gaze tracking system identifies the gaze fixation targets corresponding to the broad areas in Fig. 1(b). The target number order has been arbitrar-

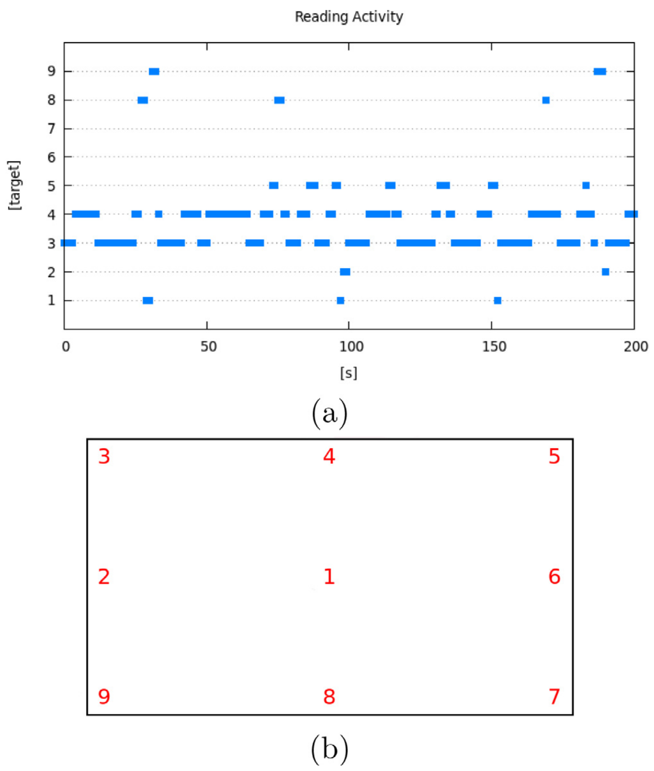


Fig. 1. (a) Gaze ethogram corresponding to the user activity "reading a text" in an experiment of duration 200 s. The targets correspond to nine different display areas in which the subject's fixations are detected. (b) Template used for calibration. The numbers denote the sequence of locations of the target areas for gaze fixations followed by the user during calibration. The same numbers are used as output categories of the DLNs.

ily defined in order to reduce the subject fatigue while performing the system calibration.

We describe the overall cognitive activity recognition based on gaze ethograms elsewhere [9]. The work in this paper covers a novel proposal that utilizes deep learning for estimating the gaze fixation on the visual target areas. The system hardware configuration is a laptop computer endowed with a web camera on top of the screen upon which a user is working. The distance of the face to the camera is roughly 50 cm, and the camera view of the face is frontal, although the subject can move freely and change pose at will. We are using off the shelf web cameras that are factory installed in laptops, therefore robustness is a challenge and a limitation. The resolution of these cameras is limited and often the image quality is quite low. Additional difficulties arise from the uncontrolled illumination conditions, and the user freedom of movement in front of the camera.

3.1. Datasets

We have generated two different datasets for these experiments. Both are produced from laptop camera captured videos with resolutions of 720p, in which the subjects perform fixations in order on every target of a calibration template for 3 s. Then, those images are selected to remove examples with too blurred or very similar images and unclear target destinations and they are hand-labeled to assign the target to each one.

- The first dataset contains images from 12 subjects with different equipment and illumination conditions. The images in this dataset have been balanced in order to guarantee an equivalent number of images in each class, trying to anticipate troubles during the training stage. This dataset contains 450 images.
- The second dataset contains images from a unique subject. The videos have been recorded under different illumination conditions and varying distance to the camera. The underlying idea is to compare the performance between ANNs trained with general, multi-user data and tailored, single-user data. This dataset is composed by 700 images.

To localize the face in the images we apply a pre-trained detector based on histograms of oriented gradients (HOG) [56] as input features for classification by linear support vector machines (SVM) [57]. Once the face is localized in the image, the next step consists of determining the position of face landmarks. This problem is known as face alignment. We use a previously trained ensemble of regression trees to estimate the face landmark position directly from a sparse subset of pixels intensities [60]. The method returns 68 2D points in the image that can be used to localize the eyes, eyebrows, nose, mouth, and jawline. This approach allows almost real-time response, although we have found trouble when the user is wearing some kind of glasses during the data capture. Finally, we select the eye area in the original images to add them to the validation datasets. Fig. 2 shows some examples of those images and the corresponding label. In this case we use the target identifier method for labeling [61].

3.2. Deep transfer learning

We have retrained six models of DLN. We use a transfer learning approach, where each DLN has been previously trained over data from the *ImageNet* challenge. We keep the weights of the intermediate layers, retraining the final layers that produce the actual classification output. In deep transfer learning [30] the already trained DLN hidden layers are assumed as a general feature extraction procedure, defining a manifold that can be used to map the input data for classification or regression tasks that are different from the original one. Task specific information is provided when training the output layers of the DLN.

We use the same output fully connected layer for all the nets. It contains 10 neurons, each one identifies one of the display target areas. The tenth neuron is used to detect cases in which the subject has the eyes closed. We have chose a softmax function for computing the classes due to it is usually recommended for likelihood computation in multi-class domains. Unless stated otherwise, we have used the Adam optimizer with a learning rate of 10^{-4}. We have validated the retrained models by cross-validation in all the experiments reported in the next section. We have repeated 30 times a 80% hold out validation, where we by randomly select a 80% of the datasets for training and use the remaining 20% for test. We report the average accuracy of the test results.

3.2.1. VGG19

The first DLN evaluated was the classical VGG19 [62]. It is a classical CNN which has 19 convolutional layers followed by max-pooling layers to reduce the image size. In order to adapt the model to our data we have removed the last layers of the pre-trained network, and added and trained two fully connected layers with 50 and 25 neurons, respectively, with its owns dropout layers to reduce the overfitting.

3.2.2. Inception-v3

The Inception neural network [63] has several versions, the fourth is the most recent. We have used the previous version because of its availability. Its structure is composed of a pattern of layers that is replicated along the net. There are modules with multiple convolutional layers in parallel that extract different image features, which are concatenated at the end of the module. We have added and trained an additional fully connected layer to fit it to our datasets.

3.2.3. Xception

The Xception neural network [64] uses the same modular composition idea of Inception architectures but there is a modification in the patterns: it changes the parallel convolutional layers by separable convolutional layers. These new layers allow to reduce the computations, being the time required to train much more images, considerably shorter.

The Xception structure presents three different stages. The initial stage applies a filter to the image for reducing the image size while it keeps the convolutional layers. The middle stage are repeated modules, which are duplicated up to eight times. The final stage has been modified and retrained to adapt it to our datasets. Here we have used two fully connected layers with 50 and 20 neurons followed by a single dropout layer to avoid overfitting.

3.2.4. ResNet50

ResNet50 [65] is a residual network. This kind of DLN architecture tries to model the residual of the prediction at previous layers. It has direct propagation of the input along the layers of the network in order to compute this residual. This structural feature alleviates the vanishing gradient problem and provides interesting computational properties, such that the computation at a given layer is independent from previous layers. Residual networks may have a very large number of layers, the one that we retrain on our datasets has 50 layers that are grouped into several blocks. At the beginning of each block, the computed residual is stored and it is used at the end of the block with the computed weights. In this case, we are using the SGD optimizer due to its superior performance against other optimizers for this kind of networks. The learning rate is 10^{-5} and the decay rate is 10^{-6} for each iteration.

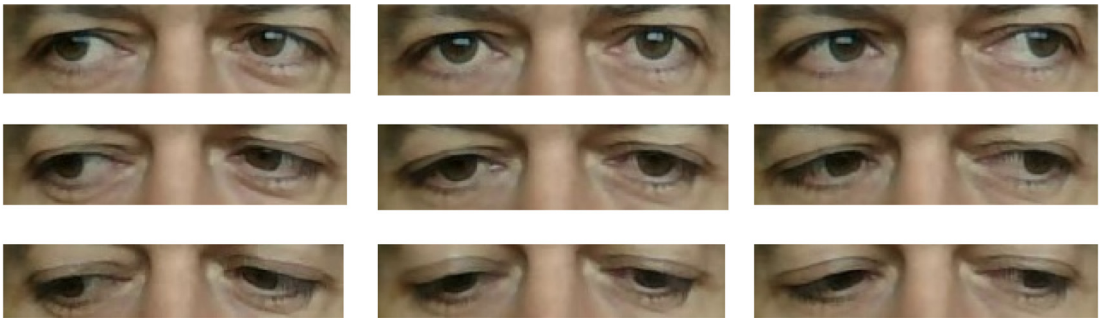
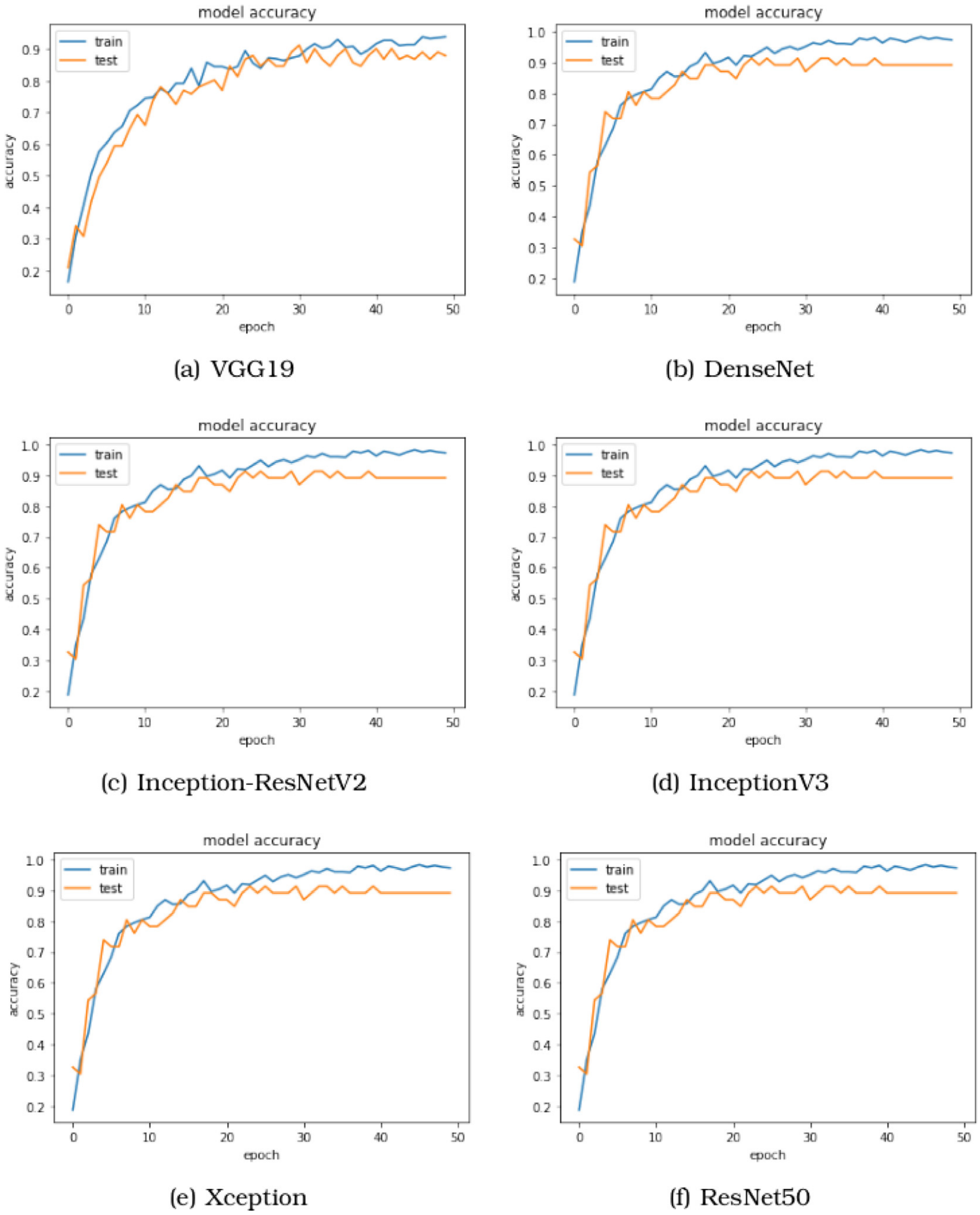


Fig. 2. Examples of eye region images captured while subjects are performing fixations in each target that compose the datasets. These images are the input to the DLNs providing the gaze fixation identification.



(a) VGG19

(b) DenseNet

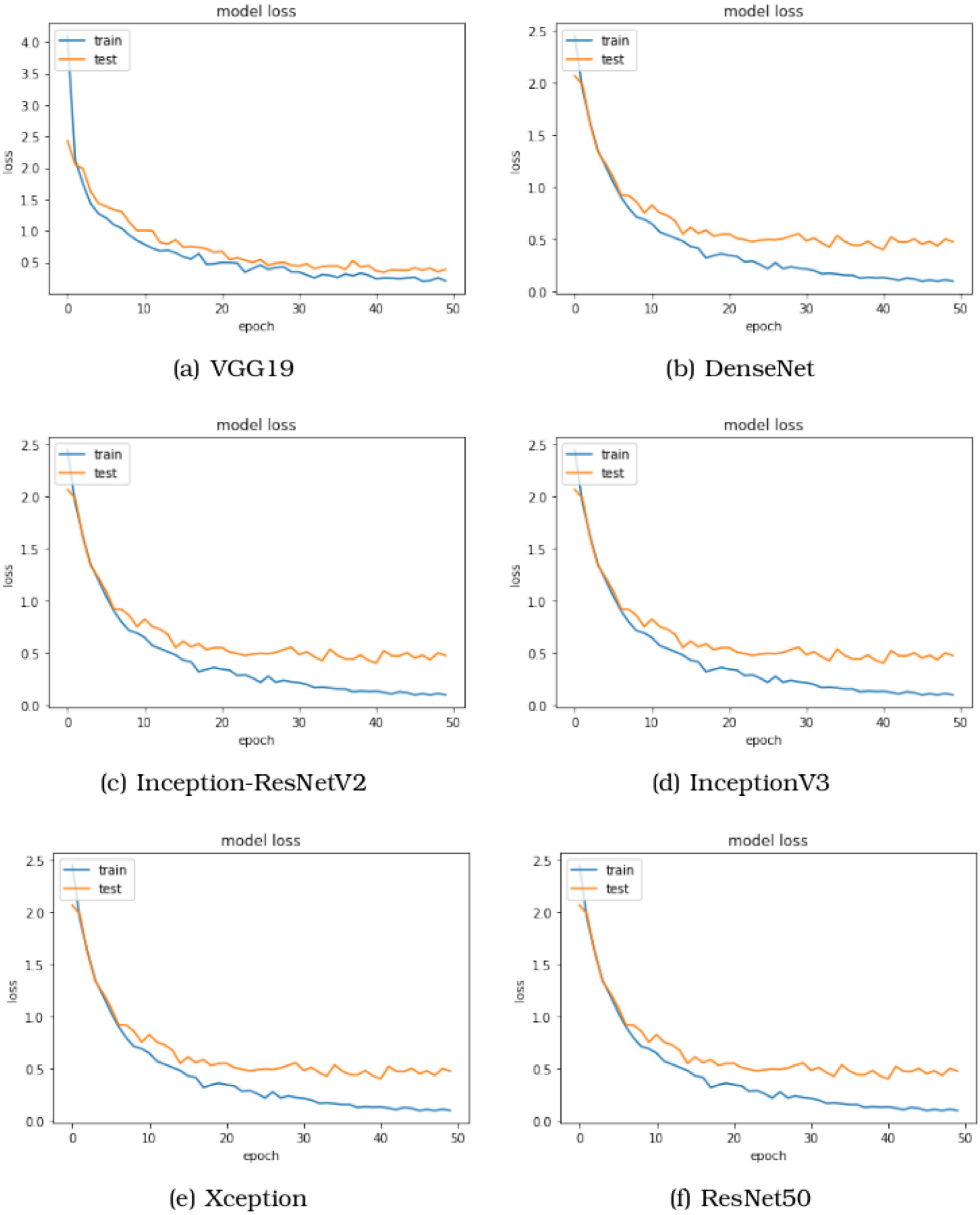
(c) Inception-ResNetV2

(d) InceptionV3

(e) Xception

(f) ResNet50

Fig. 3. Accuracy curves in training and test with the multi-user dataset.



(a) VGG19

(b) DenseNet

(c) Inception-ResNetV2

(d) InceptionV3

(e) Xception

(f) ResNet50

Fig. 4. Categorical cross-entropy loss curves in training and test with the multi-user dataset.

Table 1
Nets accuracy and error with the multi-user dataset.

Net	Best Accuracy	Lowest Error
VGG19	89.01%	0.3452
Inception-v3	86.96%	0.5529
Xception	82.61%	0.6023
ResNet50	86.96%	0.6531
Inception-ResNet-v2	84.78%	0.6248
DenseNet	91.30%	0.4281

3.2.5. Inception-ResNet-v2

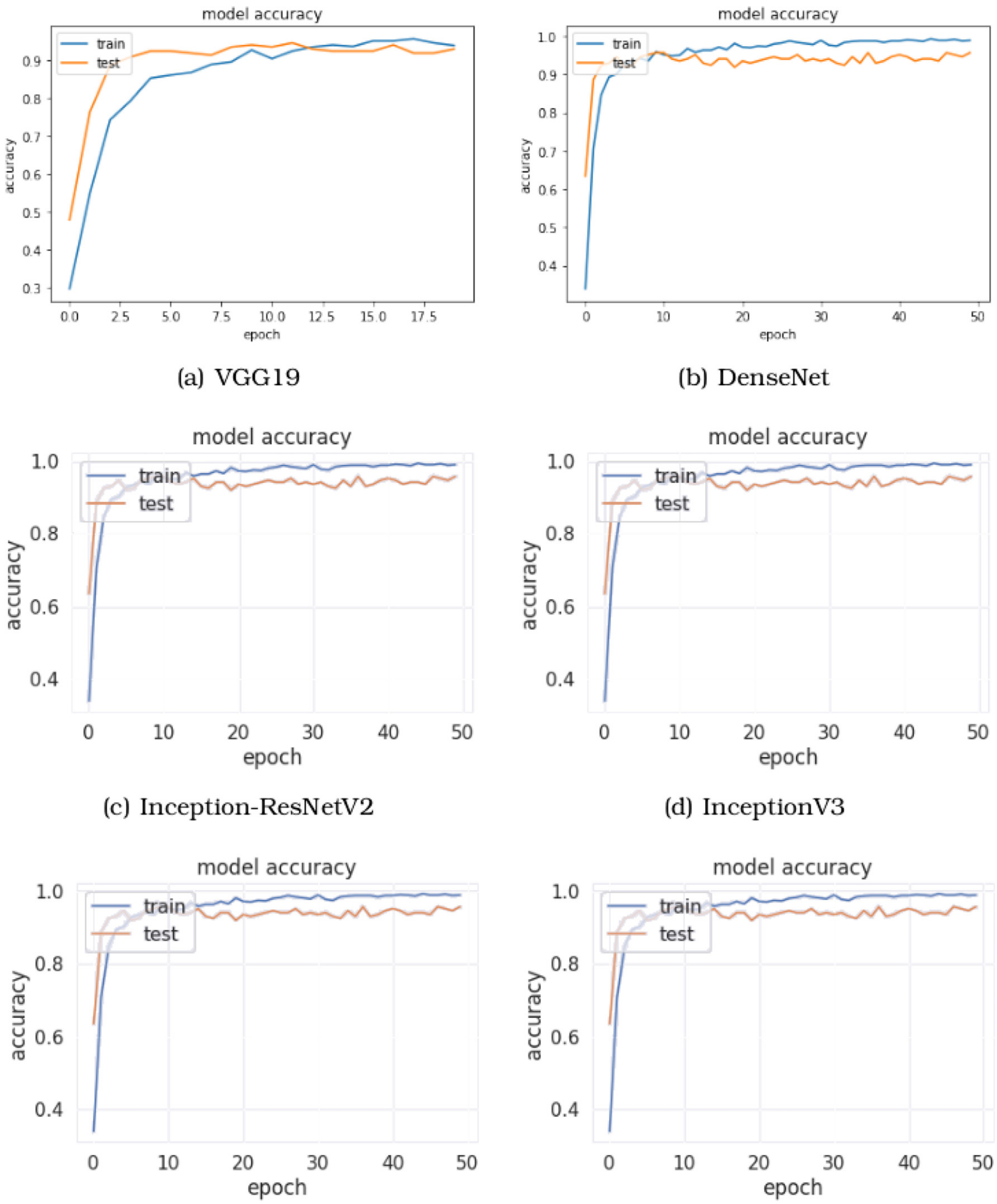
The Inception-Resnet net [66,67] combine both ResNet and Inception approaches to create a model with the advantages provided by them. The structure is composed by several blocks besides the parallel convolutional layers used to concatenate blocks. Inside the blocks there are repeated modules of the Inception flavor. There is also connections from the beginning of blocks to the end similar to ResNet ones.

3.2.6. DenseNet

DenseNet [68] follows a design idea similar to ResNet although now authors add the residual to each block globally and not only in the end of each block. Thus, it appears several connections from the inputs the convolutional layers in each block to the outputs of other blocks. Thanks to this modification the net is more compact and requires lesser layers to extract information from the image because of each layer can receive information from previous layers.

Table 2
Normalized confusion matrix from DenseNet and the multi-user dataset.

		Predicted Target									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Actual Target	**0**	1	0	0	0	0	0	0	0	0	0
	1	0	1	0	0	0	0	0	0	0	0
	2	0	0	.833	0	0	0	0	0	0	.167
	3	0	0	.143	.857	0	0	0	0	0	0
	4	0	0	0	0	1	0	0	0	0	0
	5	0	0	0	0	0	.667	0	0	0	.333
	6	0	0	0	0	0	0	.800	.200	0	0
	7	0	0	0	0	0	0	0	1	0	0
	8	0	0	0	0	0	0	0	0	1	0
	9	0	0	0	0	0	0	0	0	0	1



(a) VGG19

(b) DenseNet

(c) Inception-ResNetV2

(d) InceptionV3

Fig. 5. Average accuracy curves in training and test over the single-user dataset.

4. Experimental results and discussion

Now we discuss the learning results and general performance obtained with the DLNs described above. First, we show the multi-user dataset results. Then, we summarize and compare the results obtained on the single-user dataset.

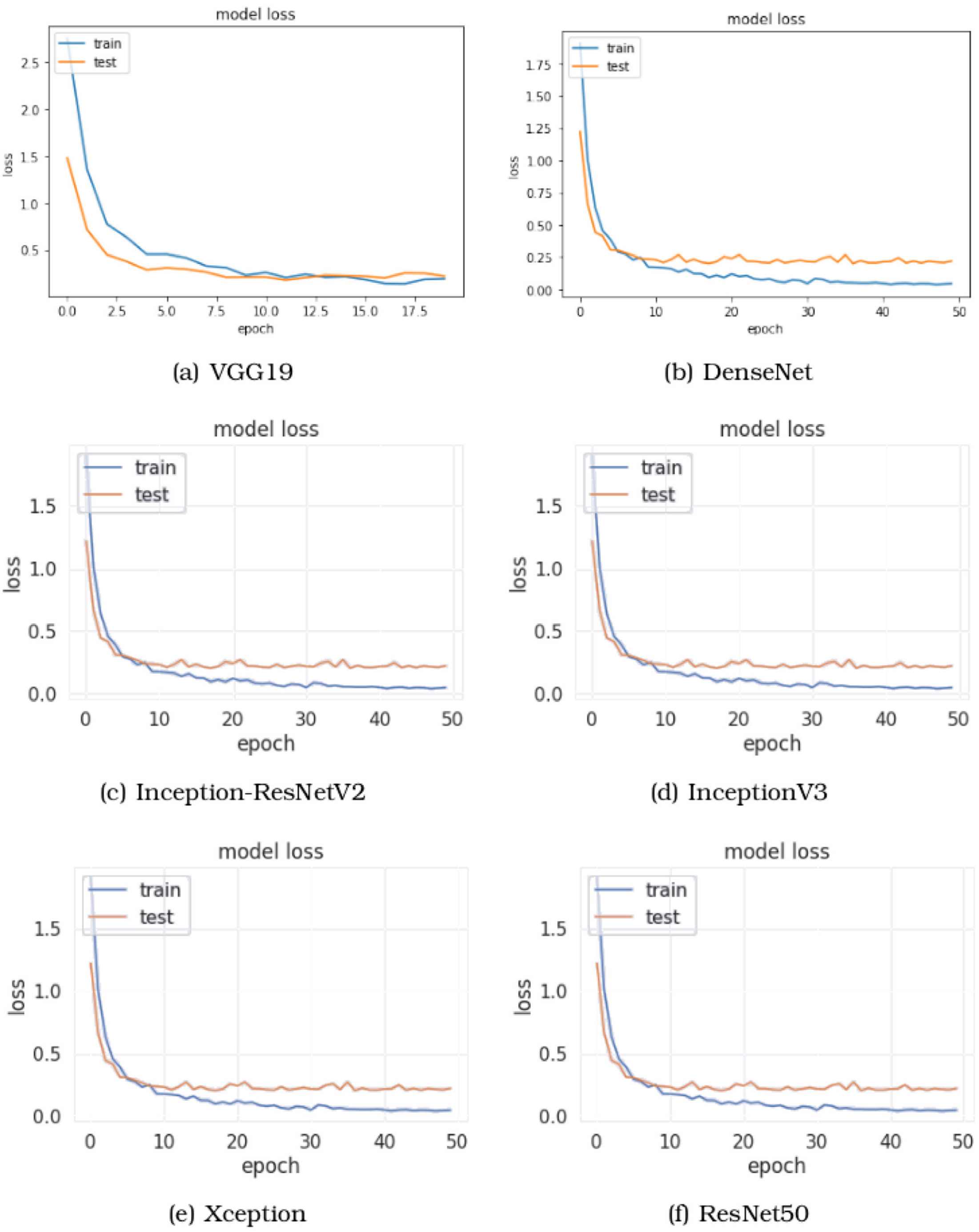
4.1. Multi-user dataset

Fig. 3 depicts the average accuracy curves obtained in training and test phases with the multi-user dataset. All the nets achieve a high accuracy after a low number of epochs. Moreover, the tendency in both training and test are almost parallel in every case.

Table 3
Best accuracy and error achieved by DLN models over the single-user dataset.

Net	Best Accuracy	Lowest Error
VGG19	94.62%	0.1788
Inception-v3	93.55%	0.3151
Xception	93.55%	0.2475
ResNet50	91.40%	0.2486
Inception-ResNet-v2	91.24%	0.2982
DenseNet	95.70%	0.2195

The features extracted by the pretrained hidden layers of the DLNs appear to provide a good baseline for this problem and our dataset. With the exception of VGG19, the DLNs stall after roughly 30



(a) VGG19

(b) DenseNet

(c) Inception-ResNetV2

(d) InceptionV3

(e) Xception

(f) ResNet50

Fig. 6. Average categorical cross-entropy loss curves during training and test over the single-user dataset.

Table 4
Normalized confusion matrix from DenseNet over the single-user dataset.

		Predicted Target									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Actual Target	**0**	1	0	0	0	0	0	0	0	0	0
	1	0	1	0	0	0	0	0	0	0	0
	2	0	0	1	0	0	0	0	0	0	0
	3	0	0	0	1	0	0	0	0	0	0
	4	0	0	0	0	.937	0	0	0	.063	0
	5	0	.040	0	0	0	.800	.120	0	0	.040
	6	0	0	0	0	0	0	1	0	0	0
	7	.133	0	0	0	0	0	0	.867	0	0
	8	0	0	0	0	0	0	0	0	1	0
	9	0	0	0	0	0	0	0	0	0	1

epochs. It usually comes from the fact that DLNs start to overfit to the training data. Therefore, probably they would need more data to keep improving the training.

Training minimizes the categorical cross-entropy loss in order to compare the real distribution with the predicted one. The lower the output of this function, the greater the degree of similarity of both distributions, and the greater the expected accuracy of classification. Fig. 4 show the average evolution of this error measure in both training and test datasets. These curves are highly negatively correlated to the ones in Fig. 3: the higher the accuracy, the lower the error. Here, we also observe that the best training evolution results are obtained with the VGG19 net. Other nets tend to overfit after the first epochs.

Table 1 shows the highest accuracy and the lowest error achieved by every DLN architecture. The results indicate that we have been able to get at least an accuracy of 80% with the new data. DenseNet achieves the best result with more than 90%.

Table 2 shows a typical test confusion matrix obtained by the retrained DenseNet on the multi-user dataset. Usually the DLNs tend to output erroneous targets when subjects look at the target areas located at the bottom of the display. The reason is that the eyes are often closed in those images so that it is hard to determine the right target under these conditions.

4.2. Single-user dataset

The proposed transfer learning architectures achieve better results when they are evaluated on the second dataset, which is composed of images from a single user. The learning problem appears easier than in the multi-user case, because we remove the data variability due to the user. Moreover, the dataset is larger than in the multi-user case. We can observe in Fig. 5 that all DLN models achieve an accuracy greater than 80% in just a few epochs. However, overfitting appears in the initial epochs so that retraining the DLNs do not improve their performance anymore. The hypothesized reason is the high similarity between all the images in the dataset. It might be partially solved by removing redundant data or by applying additional regularization methods apart from the dropout layers deployed at design time. Fig. 6 shows the evolution of the loss function on learning and test datasets. The error rate falls rapidly but it remains stable after the first epochs.

Table 3 summarizes the highest accuracy and the lowest error achieved by every DLN architecture after transfer learning. The results are better than the obtained with the multi-user dataset. All the DLN architectures achieve accuracies over 90%. DenseNet provides the best results. Note that here the nets are learning to classify the gaze corresponding to just one subject. This gives an idea about how important can be to tailor the classifiers to a final user.

Table 4 shows a typical confusion matrix generated from DenseNet and the single-user dataset. The confusion error in predicted targets follows a very similar pattern to the multi-user case.

4.3. Discussion

The models achieve competitive results with both datasets. The test accuracy achieved over the single-user dataset is greater but these results must be taken with care. The multi-user case could offer the better solution for a global system or *default mode*, while the single-user case has to be retrained for each particular user.

We expect that the results with the multi-user dataset should be improved if more images from new users are added to the dataset because the current number of images is not particularly high and DLN methods usually require larger datasets for effective training.

Also we have used the same structure and layers in nets for both datasets. Probably we could modify some layers in order to manage the overfitting problems found with the single-user dataset, as previously commented.

5. Conclusions and further work

We have presented a method for gaze fixation detection based on deep transfer learning in the context of behavioral activity recognition systems. This is usually an important part of such systems. In our case we must achieve the best performance of the gaze tracking systems because the goal of our system is to determine activities that a user carries out in front a computer and the inputs come from the camera on top of the screen.

In spite of the reduced datasets used in the experiments the use of available public pre-trained networks for domain transfer learning allows to achieve good performance with affordable computational cost. The best results according the recognition accuracy have been reported by the DenseNet model. Other models require lower training time or are easier to implement, so it should be considered as just one item to consider.

Future works will check innovative recent DL. Specifically, the recommendation of the reviewers concerning the 3D-ResNet35 architecture [69,70] that promises enhanced results due to its ability to process 3D data. Another alternative for future work is to create a new architecture from scratch. We should extend our dataset for this endeavor, because a basic requirement of DLN training are large datasets.

CRediT authorship contribution statement

Javier de Lope: Conceptualization, Methodology, Software, Writing – review & editing. **Manuel Graña:** Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been supported by FEDER funds through MINECO project TIN2017-85827-P. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 777720. XinZhe Jin contributed some early computational experiences.

References

[1] D.J. Anderson, P. Perona, Toward a science of Computational Ethology, Neuron 84 (2014) 18–31.

[2] M. Vrigkas, C. Nikou, I. Kakadiaris, A review of human activity recognition methods, Front. Robot. Artif. Intell. 2 (2015) 11.

[3] S.-R. Ke, H. Le Uyen Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, K.-H. Choi, A review on video-based human activity recognition, Computers 2(2) (2013) 88–131..

[4] J.Y. Yang, J.S. Wang, Y.P. Chen, Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers, Pattern Recogn. Lett. 29 (16) (2008) 2213–2220.

[5] M. Graña, M. Aguilar-Moreno, J. De Lope, I. Baglietto, X. Garmendia, Improved activity recognition combining inertial motion sensors and electroencephalogram signals, Int. J. Neural Syst. 30 (10) (2020) 2050053.

[6] A. Lentzas, D. Vrakas, Non-intrusive human activity recognition and abnormal behavior detection on elderly people: A review, Artif. Intell. Rev. 53 (2020) 1975–2021.

[7] N. Tasnim, M. Islam, J.-H. Baek, Deep learning-based action recognition using 3D skeleton joints information, Inventions 5 (2020) 49.

[8] J.H.F. Abeelen, Mouse mutants studied by means of ethological methods, Genetica 34 (1964) 79–94.

[9] J. De Lope, M. Graña, Behavioral activity recognition based on gaze ethograms, Int. J. Neural Syst. 30 (7) (2020) 2050025.

[10] A. George, Image based eye gaze tracking and its applications. arXiv 2019, 1907.04325..

[11] R. Hof, How do you Google? New eye tracking study reveals huge changes, Forbes Online, 2015.

[12] B. Cassin, S. Solomon, Dictionary of Eye Terminology, Triad Publising Company, Gainesville, Florida, 1990.

[13] J.D. Enderle, D.A. Sierra, A new linear muscle fiber model for neural control of saccades, Int. J. Neural Syst. 73 (2013) 1350002..

[14] R.G. Hussain, M.A. Ghazanfar, M.A. Azam, U. Naeem, S.U. Rehman, A performance comparison of machine learning classification approaches for robust activity of daily living recognition, Artif. Intell. Rev. 52(1) (2019) 357–379..

[15] G. Andrienko, N. Andrienko, G. Budziak, J. Dykes, G. Fuchs, T. von Landesberger, H. Weber, Visual analysis of pressure in football, Data Min. Knowl. Disc. 31 (6) (2017) 1793–1839.

[16] E.E. Stone, M. Skubic, Unobtrusive, continuous, in-home gait measurement using the Microsoft Kinect, IEEE Trans. Biomed. Eng. 60 (10) (2013) 2925–2932.

[17] A.T. Duchowski, Eye Tracking Methodology — Theory and Practice, Springer, Cham, 2017.

[18] A.T. Duchowski, Gaze-based interaction: A 30 year retrospective, Comput. Graph. 73 (2018) 59–69.

[19] L.R. Young, D. Sheena, Survey of eye movement recording methods, Behav. Res. Methods Instrum. 7 (5) (1975) 397–439.

[20] B.W. Blakley, L. Chan, Methods considerations for nystagmography, J. Otolaryngol. Head Neck Surg. 44 (2015) 25.

[21] L. Florea, C. Florea, C. Vertan, Recognition of the gaze direction: Anchoring with eyebrows, J. Vis. Commun. Image Rep. 35 (2016) 67–77.

[22] K.R. Park, J.J. Lee, J. Kim, Gaze position detection by computing the three dimensional facial positions and motions, Pattern Recogn. 35 (11) (2002) 2559–2569.

[23] Y.-H. Yiu, M. Aboulatta, T. Raiser, L. Ophey, V.L. Flanagin, P. Zu Eulenburg, S.-A. Ahmadi, Deep-vog: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning, J. Neurosci. Methods 324 (2019) 108307.

[24] N. Barbara, T.A. Camilleri, K.P. Camilleri, EOG-based eye movement detection and gaze estimation for an asynchronous virtual keyboard, Biomed. Signal Process. Control 47 (2019) 159–167.

[25] P.M. Insch, G. Slessor, J. Warrington, L.H. Phillips, Gaze detection and gaze cuing in Alzheimer's Disease, Brain Cogn. 116 (2017) 47–53.

[26] O. Grynszpan, J. Bouteiller, S. Grynszpan, F. Le Barillier, J.C. Martin, J. Nadel, Altered sense of gaze leading in autism, Res. Autism Spect. Disord. 67 (2019) 101441.

[27] J. Kim, J. Seo, T.H. Laine, Detecting boredom from eye gaze and EEG, Biomed. Signal Process. Control 46 (2018) 302–313.

[28] M.-J. Tsai, H.-T. Hou, M.-L. Lai, W.-Y. Liu, F.-Y. Yang, Visual attention for solving multiple-choice science problem: An eye-tracking analysis, Comput. Educ. 58 (2012) 375–385.

[29] A. Gutierrez-Garcia, A. Fernandez-Martin, M. Del Libano, M.G. Calvo, Selective gaze direction and interpretation of facial expressions in social anxiety, Pers. Individ. Differ. 147 (2019) 297–305.

[30] M. Talo, U.B. Baloglu, O. Yildirim, U.R. Acharya, Application of deep transfer learning for automated brain abnormality classification using MR images, Cogn. Syst. Res. 54 (2019) 176–188.

[31] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, U. Montreal, Greedy layer-wise training of deep networks (2007) 19.

[32] G.E. Hinton, D. Osindero, Y-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18(7) (2006) 1527–1554..

[33] M.A. Ranzato, Y.-L. Boureau, Y. LeCun, Sparse feature learning for deep belief networks, Conf. Neural Inf. Proc. Syst. (2007) 1185–1192.

[34] D. Bacciu, P. Lisboa, J. Martin-Guerrero, R. Stoean, A. Vellido, Bioinformatics and medicine in the era of deep learning, 2018, arXiv:1802.09791..

[35] Y. Lecun, P. Haffner, L. Bottou, Y. Bengio, Object recognition with gradient-based learning, in: Shape, Contour and Grouping in Computer Vision. Lecture Notes in Computer Science, vol 1681. Springer, Berlin, Heidelberg. doi: 10.1007/3-540-46805-6_19..

[36] L.A. Gatys, A.S. Ecker, M. Bethge, Image style transfer using convolutional neural networks, IEEE Conf. on Computer Vision and Pattern Recognition (2016) 2414–2423.

[37] G. Antipov, M. Baccouche, J. Dugelay, Face aging with conditional generative adversarial networks, IEEE Int. Conf. in Image Processing (2017) 2089–2093.

[38] A. Ucar, Y. Demir, C. Guzelis, Object recognition and detection with deep learning for autonomous driving applications, Simulation 93 (2017).

[39] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: IEEE Int. Conf. in Image Processing, 2016.

[40] K. Potdar, C. Pai, S. Akolkar, A convolutional neural network based live object recognition system as blind aid, 2018..

[41] B. Ma, X. Li, Y. Xia, Y. Zhang, Autonomous deep learning: A genetic DCNN designer for image classification, Neurocomputing 379 (2020) 152–161.

[42] Y. Zhang, Y. Wang, X.-Y. Liu, S. Mi, M.-L. Zhang, Large-scale multi-label classification using unknown streaming images, Pattern Recogn. 99 (2020) 107100.

[43] M. Talo, U.B. Baloglu, O. Yildirim, U.R. Acharya, Application of deep transfer learning for automated brain abnormality classification using MR images, Cogn. Syst. Res. 54 (2019) 176–188.

[44] U. Raghavendra, H. Fujita, S.V. Bhandary, A. Gudigar, J.H. Tan, U.R. Acharya, Deep convolutional neural network for accurate diagnosis of glaucoma using digital fundus images, Inf. Sci. 441 (2018) 41–49.

[45] O. Yildirim, M. Talo, B. Ay, U.B. Baloglu, G. Aydin, U.R. Acharya, Automated detection of diabetic subject using pre-trained 2D-CNN models with frequency spectrum images extracted from heart rate signals, Comput. Biol. Med. 113 (2019) 103387.

[46] Wang, L.; Wong, A. COVID-NET: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. 2020..

[47] F. Shan, Y. Gao, J. Wang, W. Shi, N. Shi, M. Han, Z. Xue, D. Shen, Y. Shi, Lung infection quantification of COVID-19 in CT images with deep learning, 2020..

[48] W. Ouyang, X. Wang, Joint deep learning for pedestrian detection, IEEE Int. Conf. in Computer Vision, 2013.

[49] R. Grzeszick, J.M. Lenk, F.M. Rueda, G.A. Fink, S. Feldhort, M. ten Hompel, Deep neural network based human activity recognition for the order picking process, iWOAR 2017..

[50] H. Jiang, E. Learned-Miller, Face detection with Faster RCNN, IEEE Int. Conf. Automatic Face Gesture Recognition (2017) 650–657.

[51] X. Sun, P. Wu, S.C. Hoi, Face detection using deep learning: An improved Faster RCNN approach, Neurocomputing 299 (2018) 42–50.

[52] R. Ranjan, V.M. Patel, R. Chellappa, A deep pyramid deformable part model for face detection. CoRR 2015, abs/1508.04389..

[53] S. Yang, P. Luo, C.C. Loy, X. Tang, Faceness-net: Face detection through deep facial part responses, IEEE Trans. Pattern Anal. Mach. Intell. (2017).

[54] W. Wang, J. Yang, J. Xiao, S. Li, D. Zhou, Face recognition based on deep learning, in: Q. Zu, B. Hu, N. Gu, S. Seng (Eds.), Human Centered Computing, Springer, Cham, 2015, pp. 812–820.

[55] M. Islam, N. Tasnim, J.-H. Baek, Human gender classification using transfer learning via Pareto frontier CNN networks, Inventions 5 (2020) 16.

[56] N. Dalal, B. Triggs, Histogram of oriented gradients for human detection, IEEE Conf. Comp. Vision and Pattern Recognition (2005) 886–893.

[57] C. Cortes, V.N. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

[60] V. Kazemi, J. Sullivan, One millisecond face alignment with an ensemble of regression trees, IEEE Conf. Computer Vision and Pattern Recognition (2014) 1867–1874.

[61] J. De Lope, M. Graña, Comparison of labeling methods for behavioral activity classification based on gaze ethograms, in: E.A. De la Cal, J.R. Villar Flecha, H. Quintian, E. Corchado (Eds.), Hybrid Artificial Intelligent Systems, Springer, Cham, 2020, pp. 132–144.

[62] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv 2014, 1409.1556..

[63] S. Giri, B. Joshi, Transfer learning based image visualization using CNN, Int. J. Artif. Intell. Appl. 10 (4) (2019) 47–55.

[64] F. Chollet, Xception: Deep learning with depthwise separable convolutions. CoRR 2016, abs/1610.02357..

[65] A. Mahmood, A. Giraldo, M. Bennamoun, S. An, F. Sohel, F. Boussaid, R. Hovey, R. Fisher, G. Kendrick, Automatic hierarchical classification of kelps using deep residual features, Sensors 20 (2020) 447.

[66] A.Alemi, Improving Inception and image classification in TensorFlow, GoogleBlog 2016. URL: https://ai.googleblog.com/2016/08/improving-inception-and-image.html..

[67] C. Szegedy, S. Ioffe, V. Vanhoucke, Inception-v4, Inception-ResNet and the impact of residual connections on learning. CoRR 2016, abs/1602.07261..

[68] G. Huang, Z. Liu, K.Q. Weinberger, Densely connected convolutional networks, CoRR 2016, abs/1608.06993..

[69] M.A. Bhimra, U. Nazir, M. Taj, Using 3D Residual Network for Spatio-temporal Analysis of Remote Sensing Data, ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 1403–1407. doi: 10.1109/ICASSP.2019.8682286..

[70] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatio-temporal features with 3D convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.

Further reading

[58] A. Mohan, C. Papageorgiou, T. Poggio, Example-based object detection in images by components, IEEE Trans. Pattern Anal. Mach. Intell. 23 (4) (2001) 349–361.

[59] P. Viola, M.J. Jones, D. Snow, Detecting pedestrians using patterns of motion and appearance, IEEE Int. Conf. Computer Vision 2 (2003) 734–741.

Manuel Graña Romay received the M.Sc. and Ph.D. degrees in Computer Science from Universidad del Pais Vasco (UPV/EHU), Donostia, Spain, in 1982 and 1989, respectively. His current position is a Full Professor (Catedrático de Universidad) with the Computer Science and Artificial Intelligence Department of the Universidad del Pais Vasco (UPV/EHU) since 1998, where he acted as head of department in the period 2005–2007. He is the head of the Computational Intelligence Group (Grupo de Inteligencia Computational), which has been recognized as excellent research group by the Basque Government with continuous specific funding since 2005, last grant for the period 2019–2021. The research group has carried out over 30 national funded research projects, three European Commission funded projects, and some private company research contracts. The research works in the group spread over a great variety of topics, including applications of artificial intelligence and computational intelligence to linked multicomponent robotic systems, reinforcement learning, medical image in the neurosciences, multimodal human computer interaction, remote sensing image processing, content based image retrieval, lattice computing, semantic modeling, data processing, classification, and data mining. He has been advisor for over 35 PhD Thesis, editor of more than 20 books of proceedings and collections of works on monographic topics, editor of more than 15 special issues in journals, and co-author of more than 200 journal papers (ISI indexed journals). He is associated editor of Neurocomputing, Information Fusion, Computational Intelligence and Neurosciences, Frontiers in Big Data, Journal of Mathematical Imaging and Vision.