

AFFECTIVE STATE RECOGNITION BASED ON EYE GAZE ANALYSIS USING TWO-STREAM CONVOLUTIONAL NETWORKS

Christina Chrysouli Nicholas Vretos Petros Daras

Information Technologies Institute
Centre for Research and Technology, Hellas, Thessaloniki, Greece
{chrysouli, vretos, daras}@iti.gr

ABSTRACT

In this paper, we propose a novel technique that combines the concept of spatially targeted optical flow with image processing, for affect state recognition, concerning a wide variety of learner types, including children with autism and mainstream children. We exploit the advantages of deep Neural Networks on image classification, by adopting a two-stream CNN approach for the recognition task, based on gaze analysis. As there is not a publicly available dataset to contain such a variety of learner types, a dataset was created in order to evaluate the performance of our algorithm. We validate our approach using this dataset, by optimising a mean-square error loss function, obtaining promising results for this challenging task.

Index Terms— Affective computing, Convolutional Neural Networks, gaze analysis

1. INTRODUCTION

Affective state characterises the experience of an emotion [1], which can be expressed through a variety of ways, including facial expressions, body posture and speech. According to Ekman [2], there are six fundamental emotions that can be expressed through facial expressions: anger, disgust, fear, happiness, sadness and surprise. As an alternative to these basic emotions, researchers have also proposed time-dependent affective states that are usually application driven [3], [4]. A detailed survey of time-dependent affective states and datasets, constructed to meet different application requirements, can be found in [5].

Affective state classification has attracted a lot of attention during the last years. Many works have been proposed that range in several aspects from input data to classification schemes. For instance, in [6], the authors propose a deep model approach that combines facial expression and body motion analysis in order to identify the affective state of a player during gameplay, which also proved robust in the absence of one of the modalities. In another recent work [7], a combination of Convolutional Neural Network and specific

image pre-processing steps (e.g. cropping, rotation correction, intensity normalisation) were proposed for facial expression classification towards the six Ekmanian emotions. Current research focuses on how emotional states are perceived from different type of learners, and not on how to identify the emotional state of a type of learner [5]. Thus, most existing datasets contain data coming from actors performing an emotion, which are later shown to different type of learners in order to examine how these subjects perceive emotions. In [8], the authors summarise some of the methods used for affective state classification, also focusing on existing datasets and benchmarks.

Convolutional Neural Networks (CNNs) have been used for a plethora of applications, including face recognition [9], action recognition [10] and affective state recognition [6], [7], achieving, thus far, remarkable results. A promising approach in CNN architectures for processing video sequences include training a two-stream CNN [11], [12], aiming at capturing the complementary information between still images and motion between frames [10].

Facial emotion recognition is not an easy task, let alone affective state recognition only from gaze, since people can vary in the way they show their expressions. Even images of the same person in the same facial expression can vary, e.g. in brightness, background and/or pose. Although a great amount of research effort has been invested on affective state recognition using the full face as an emotion indicator [7], to our knowledge, there are no papers focusing on affective state recognition based solely on the eyes. Few researchers include eye gaze analysis in multimodal approaches as auxiliary information to the facial analysis [13], [14].

In this paper, we propose a spatiotemporal CNN architecture for eye gaze analysis in order to recognise affective state from video. To do so, eye images are fed to the spatial stream, while optical flow, extracted from consecutive frames, are fed to the temporal one. Optical flow has been used for eye gaze analysis before [15], [16]. Moreover, it is well known that the gaze plays a decisive role in emotional expression, as well as in non-verbal communication. Therefore, in this paper, temporal stream is adopted so as information about the eyes

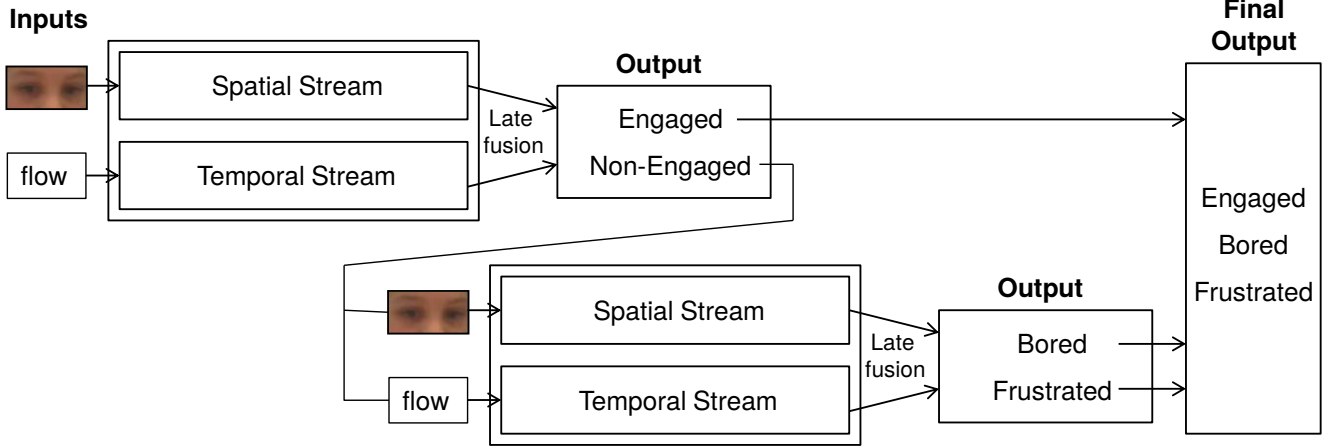


Fig. 1: Proposed pipeline, using 2 two-stream CNNs. Note that inputs are the same in both networks (Section 2.3).

movement evolution in time, is infused to the model. Our overall goal is to perceive the correlative information between the appearance from still images and motion among frames.

The novel contribution of this paper is threefold: 1) We make use of a network and a technique, primarily used for action recognition, in order to calculate the, way more subtle, movement of the eyes. 2) We have adopted a two-step recognition process, that contributes to the final structure of the network, and also leads to the improvement of the results. 3) A dataset, that contains a variety of learner types involved in use cases of the MaTHiSiS project, is constructed.

The remainder of this paper is organised as follows. In Section 2, the problem at hand is declared, along with a detailed description of the proposed network architecture and the developed method. In Section 3, we present the constructed dataset, while, in Section 4 the experimental results of the proposed algorithm are presented and described. Finally, in Section 5, conclusions are drawn and future work is discussed.

2. PROPOSED METHOD

Video can be interpreted as a continue sequence of still images or as an image evolving in time, therefore, it can be decomposed into spatial and temporal components. The proposed method follows this, rather straightforward, division into two streams, one spatial that handles eye images extracted from video frames, and one temporal that handles motion across frames, by means of optical flow. Each stream is implemented using a distinct CNN, based on the same concept proposed in [10], which are therefore combined using a late fusion layer. Although we borrow a technique from action recognition, eye movement is way more subtle motion and needs different handling. The aim of the proposed method involves classifying eye images from each learner

type, into three time-dependent affective states, namely engagement, boredom and frustration.

2.1. CNN for affective state recognition

A sequential CNN usually consists of an input layer, followed by multiple hidden layers and an output layer. The hidden layers typically consist of convolution, pooling, fully connected and normalisation layers. Depending on the problem at hand, an appropriate network architecture is constructed. The spatial CNN of the proposed method operates on still images, by directly taking individual video frames as network inputs, followed by several convolution layers, pooling layers and fully connected layers, as presented in Table 1. Finally, in order to obtain the class prediction, a sigmoid function is used in the last layer:

$$S(z) = \frac{1}{1 + e^{-z}}, \quad (1)$$

where z is the input value, and $S(z)$ is the obtained output in the range of $[0, 1]$, which is considered as the predicted class probability. The spatial network processes one frame at time and, substantially, carries visual information about objects depicted in the video.

On the other hand, the temporal CNN encodes motion information across the frames of a video. The proposed architecture follows the same model architecture as the spatial one, only instead of feeding images as input, it operates on optical flow displacement fields. Optical flow is calculated in pairs of overlapping consecutive frames, and the horizontal and vertical components produced are used to train the temporal stream. The hidden layers as well as the output prediction of the model are constructed according to the spatial stream.

The CNN models (spatial and temporal) utilise the rectified linear unit (ReLU) as an activation function between all layers, except when fully connected layers are applied (two

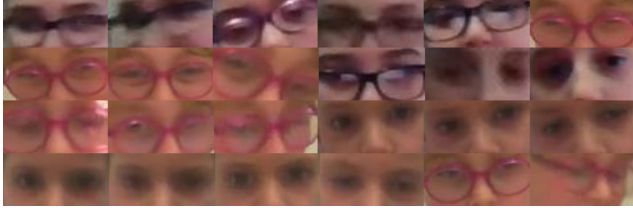


Fig. 2: MaTHiSiS dataset: eye images from Boredom class (left) and Frustration class (right) ¹

Table 1: Proposed network architecture.

Network stage	Output size	Layer type
Conv 1	20x41	Conv.[7x7, 64]+MaxPool 2x2
Conv 2	10x20	Conv.[5x5, 128]+MaxPool 2x2
Conv 3	10x20	Conv.[3x3, 128]
Conv 4	10x20	Conv.[3x3, 128]
Conv 5	5x10	Conv.[3x3, 128]+MaxPool 2x2
Classifier		FC[256]+FC[128]+Sigmoid, 1

last layers and output). A suitable loss function, for the dual classification task, is employed: binary cross entropy. The Adam optimisation algorithm was used with the following configuration parameters: $learning_rate = 0.001$, $b_1 = 0.9$ and $b_2 = 0.999$, where b_1 , b_2 are the exponential decay rates for the first and second moment estimates, respectively.

2.2. Optical flow

Optical flow is calculated between a pair of images, usually consecutive, for all frames f in a video. Horn–Schunck’s Method [17] is a quite popular algorithm for optical flow calculation, due to its simplicity in implementation, as well as its acceptable results in a short time frame. The algorithm is based on a differential technique, which assumes smoothness in the flow over the whole image, and the estimated velocity field is computed by using a gradient constraint (i.e. brightness constancy).

The desired optical flow vector field (u, v) can be formulated as a minimisation problem of global energy functional [17]:

$$E_{HS}(u, v) = \iint \left[(I_x u + I_y v + I_t)^2 + \alpha (|\nabla u|^2 + |\nabla v|^2) \right] dx dy, \quad (2)$$

where I_x , I_y , I_t are the spatiotemporal image brightness derivatives, u , v are the horizontal and vertical displacements

¹The authors have chosen to use these images since person identification is practically impossible. These images have been collected within the context of the MaTHiSiS project after having obtained the informed consent of the legal representatives of the subjects.

respectively, ∇ is the gradient operator and α is a smoothness factor that weights the regularization. The larger the value of α , the stronger the penalisation of large flow gradients, leading to smoother flow fields.

We have computed optical flow based on the Horn–Schunck method [17], using parameters: expected smoothness of optical flow Smoothness=1, and MaxIteration=100. The larger the MaxIteration is, the better the estimation of optical flow of objects with low velocity.

Although more advanced techniques for calculating optical flow exist [18], we have chosen Horn–Schunck’s method as its results captured the movement of the eyes well, in significant less time. Horn–Schunck’s Method produced velocity fields with values close to zero for areas with little or no movement, thus, spatially located on the area of interest (i.e. the eyes). On the other hand, Brox’s approach [18] returned not so spatially located velocity fields, resulting in similar velocity fields for a quite large number of video frames, making it unsuitable choice for our classification task. This problem may arise from the nature of our problem, as the images used only contain part of the face, around a subject’s eyes (i.e. no background).

2.3. Fusion

The framework that was followed, in predicting the affective state of a subject, consists of a two-step recognition process. The ultimate goal is to assign every video frame to one of the three affective states: engagement, boredom and frustration.

The first step is to decide if a subject in a video frame is engaged or not. The second step is to decide if the subjects, classified as non-engaged, classify as bored or frustrated. Thus, a two-stream CNN is employed for the first step, and another one for the second. Figure 1 illustrates the proposed framework architecture.

The network that classifies a subject as engaged or non-engaged is trained using data from all affective states; subjects that are labelled as bored and frustrated constitute the non-engagement class, while subjects that are labelled as engaged constitute engagement class. On the contrary, the network that classifies a subject as frustrated or bored is trained using data only from these two affective states. The classifying procedure is as follows: First, a subject’s eye-image and its optical flow, computed as in (2), passes through the first two-

stream CNN, which decides whether this subject is engaged or not. If the subject is classified as engaged, then the procedure stops, with Engagement class as outcome. If the subject is classified as non-engaged, then the same image needs to pass through the second two-stream CNN in order to decide between Boredom and Frustration class.

By training two different two-stream CNNs, we aim at a better overall distinction between the three classes, as shown in the experimental results Section (Section 4). Lastly, each network processes one frame at a time and the predictions on single frames are fused by averaging probabilities of each network.

3. MATHISIS DATASET

The necessity of creating a dataset derived from the fact that there was no available dataset that contained all learner types involved in use cases of the MaTHiSiS project. The full dataset is a collection of data (i.e. video, sound, skeleton and mobile) and metadata (i.e. ground truth annotations) over them. The dataset that we have used in our experiments is, in fact, a subset of the MaTHiSiS dataset. For brevity, this subset will be referred as “MaTHiSiS dataset” or dataset for the rest of this paper.

The MaTHiSiS dataset contains videos of people interacting with a computer (both children and adults), capturing their different facial expressions. In more detail, a Kinect camera (640×480 resolution) was used for the recordings, which was set quite close to the users’ face, resulting in detailed images of the subjects’ face and eyes, as illustrated in Figure 2. The dataset has a good variety of users, as it involves data from multiple learner types: Mainstream Education case (MEC), Industrial Training case (ITC), Career Guidance & Distance Learning case (CGDLC), Profound and Multiple Learning Difficulties case (PMLDC) and Autism Spectrum Condition case (ASC). The cases are presented in Table 2, along with the age group that they belong.

No instructions were given to the subjects on how to react to the different use cases and, as a result, the dataset contains a wide range of head poses and facial expressions. In Figure 2, a small fraction of the great variety within each class is represented, which, combined with the similarity between the classes, the variation in the illumination conditions, as well as the facial props, result in a highly diverse and, therefore, challenging dataset.

The process of attaching labels to the data collected was executed by the teacher responsible for each subject, or by psychologists or pedagogical experts, who were present at the process of data collection, based on their on-set consultation with the teacher. The labelling procedure resulted in associating video segments of the learning session to *only one* of the three labels, namely engagement, boredom and frustration, also attaching a level of certainty to each decision.

The teachers’ role to labelling procedure was really important especially to the AS and PMLD cases, as the subjects’ expressions and behaviour were more likely to be ambiguous. Thus, it was crucial that the teacher would take part in the process of labelling, as he/she would be the one to know and understand the emotional state of the subject.

3.1. Data pre-processing

The video data, collected as described in Section 3, underwent a pre-processing procedure in order to be able to use them in the CNN. First, a face detector was adopted so as to decide if a face exists in each video frame. For this task, IntraFace software for facial image analysis was employed [19]. The software detects a face by searching for facial landmarks, 49 landmarks in total, 6 around each eye. If no facial landmarks are found, it assumes that no faces are present in this particular image. Only frames that IntraFace has returned landmarks are kept, while the rest are discarded. Then, a bounding box is calculated, which includes only the facial landmarks around the eyes, with a small margin around, and a new image is created by cropping the initial frame. This newly created image contains only subject’s eyes.

Since the eye images, that we end up with, are not necessarily of the same size, we scale the images towards a mean bounding box, so as to be used in CNN training procedure. In more detail, the mean width and mean height of all the eye images are calculated (only the images of the training set is taken into account) and all the images are scaled up or down towards these new dimensions. The distortions by scaling are negligible in MaTHiSiS dataset, as illustrated in Figure 2, in which the eye images are of size 41×83 . The characteristics of the data obtained, after pre-processing procedure, are presented in the first four columns of Table 2.

4. EXPERIMENTS

In order to evaluate the performance of the proposed method, we have conducted several experiments using the MaTHiSiS dataset. Each learner type in the dataset was treated as a different case, while two more cases were created of the union of some cases. In more detail, ITC and CGDLC were merged so as to create a new class of “mainstream adults”, while ASC and PMLDC created the “children with learning difficulties” class.

Generally, in most cases, the classes were severely unbalanced, favourably to engagement affective state class. To overcome this issue, a weight has been assigned to each class, aiming at increasing the importance of the under-represented classes. Moreover, the optical flow was pre-calculated, using (2), and stored for time efficiency reasons.

We perform cross validation for each one of the cases presented in Table 2, in order to inspect how the model performs when we feed it with unknown data, in terms of accuracy of

Table 2: Results on MaTHiSiS dataset

Learner type	age group	#images	#subjects	#folds	one-step	two-step	
					E/B/F acc	E/NE acc	E/B/F acc
ASC	children	40414	11	5	54.64	64.41	62.87
CGDLC	adults	35502	5	4	84.84	86.82	86.82
ITC	adults	6788	5	5	40.58	75.33	60.91
MEC	children	249441	20	5	76.30	76.29	76.30
PMLDC	children	5670	2	5	82.75*	95.35*	92.76*
ASC+PMLDC	children	46084	13	5	65.64	63.97	60.45
CGDLC+ITC	adults	42290	10	5	63.30	71.41	62.87

its predictions. Depending on the number of subjects in each case, 4 or 5 folds have been used. Precautions have been taken in order to assure that eye images from the same subject do not appear in both training and test set, and that the features learnt by the CNN are not subject-dependent. This was applied to all created folds of the cross validation process, except for PMLDC. PMLDC is a special case, as there were only 2 subjects, thus, training and test sets are not completely independent (results marked with a star).

The number of epochs was set to 80, but an early stopping criterion was also employed. The criterion stops the training process if the metric under evaluation (i.e. accuracy) does not improve for *patience* = 5 consecutive epochs. Usually, the training converged around the 10th epoch.

The results obtained are summarized in Table 2 and concern the mean accuracy (denoted as “acc”) of all the folds trained. In Table 2, E/NE stands for the two-class Engagement/ Non Engagement case and E/B/F stands for the three-class Engagement/ Boredom/ Frustration case. The column for one-step case concerns the recognition directly for 3 classes, and the configuration of the experiments are the same as the two-step recognition process, for comparison reasons. We notice that the two-step recognition process, with a fusion layer, generally performs better than if the decision was based directly on the three classes. Moreover, despite the noisy nature of the data, we notice that the accuracy of the 3-class prediction is, in most cases, equally or better, proportionally, than the two-class prediction.

5. CONCLUSION

We have presented a novel algorithm that attempts to incorporate two different representations of an image, in order to achieve good classification results. Using a two-stream CNN, we recognize the affective state of a learner from a video sequence, only with the aid of eye images and their corresponding optical flow. To our knowledge, this is the first method to tackle affective state recognition problem based solely on gaze analysis. The experiments that conducted

showed promising results.

In the future, we aim to enrich the MaTHiSiS dataset with more subjects. As the quantity of input images and, specifically, the variation of the subjects in the dataset increases, the results will also improve. We shall also focus our efforts on experimenting with different ways to handle the imbalanced classes, for example by developing a technique to exclude part of the data from training. Moreover, we could experiment with alternative network architectures, including different ways of fusion, or even with different optical flow algorithms, with scholastic fine tuning.

6. ACKNOWLEDGEMENTS

The work presented in this document is a result of MaTHiSiS project. This research has received funding from the European Union’s Horizon 2020 Programme (H2020-ICT-2015) under Grant Agreement No. 687772.

7. REFERENCES

- [1] Rosalind W Picard and Roalind Picard, *Affective computing*, vol. 252, MIT press Cambridge, 1997.
- [2] Paul Ekman and Wallace V Friesen, “Facial action coding system: A technique for the measurement of facial movement,” 1978.
- [3] Ryan SJd Baker, Sidney K D’Mello, Ma Mercedes T Rodrigo, and Arthur C Graesser, “Better to be frustrated than bored: The incidence, persistence, and impact of learners cognitive-affective states during interactions with three different computer-based learning environments,” *International Journal of Human-Computer Studies*, vol. 68, no. 4, pp. 223–241, 2010.
- [4] Hongying Meng and Nadia Bianchi-Berthouze, “Affective state level recognition in naturalistic facial and vocal expressions,” *IEEE Transactions on Cybernetics*, vol. 44, no. 3, pp. 315–328, 2014.

- [5] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2009.
- [6] Athanasios Psaltis, Kyriaki Kaza, Kiriakos Stefanidis, Spyridon Thermos, Konstantinos C Apostolakis, Kosmas Dimitropoulos, and Petros Daras, "Multimodal affective state recognition in serious games applications," in *Imaging Systems and Techniques (IST), 2016 IEEE International Conference on*. IEEE, 2016, pp. 435–439.
- [7] André Teixeira Lopes, Edilson de Aguiar, Alberto F De Souza, and Thiago Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610–628, 2017.
- [8] Ciprian Adrian Corneanu, Marc Oliu Simon, Jeffrey F Cohn, and Sergio Escalera Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.
- [9] Bo-Kyeong Kim, Jihyeon Roh, Suh-Yeon Dong, and Soo-Young Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 173–189, 2016.
- [10] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," *CoRR*, vol. abs/1406.2199, 2014.
- [11] Hao Ye, Zuxuan Wu, Rui-Wei Zhao, Xi Wang, Yu-Gang Jiang, and Xiangyang Xue, "Evaluating two-stream cnn for video classification," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015, pp. 435–442.
- [12] Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander G Hauptmann, "Hidden two-stream convolutional networks for action recognition," *arXiv preprint arXiv:1704.00389*, 2017.
- [13] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.
- [14] Jiajia Li, Grace Ngai, Hong Va Leong, and Stephen CF Chan, "Multimodal human attention detection for reading from facial expression, eye gaze, and mouse dynamics," *ACM SIGAPP Applied Computing Review*, vol. 16, no. 3, pp. 37–49, 2016.
- [15] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato, "Gaze estimation from eye appearance: a head pose-free method via eye image synthesis," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3680–3693, 2015.
- [16] George Leifman, Dmitry Rudoy, Tristan Swedish, Eduardo Bayro-Corrochano, and Ramesh Raskar, "Learning gaze transitions from depth to improve video saliency estimation," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [17] Berthold KP Horn and Brian G Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.
- [18] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert, "High accuracy optical flow estimation based on a theory for warping," *Computer Vision-ECCV 2004*, pp. 25–36, 2004.
- [19] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. F. Cohn, "Intraface," in *Automatic Face and Gesture Recognition*, 2015.