A Neuro-fuzzy Approach to User Attention Recognition

Stylianos Asteriadis, Kostas Karpouzis, and Stefanos Kollias

National Technical University of Athens,
School of Electrical and Computer Engineering,
Image, Video and Multimedia Systems Laboratory,
GR-157 80 Zographou, Greece
stiast@image.ece.ntua.gr, {kkarpou,stefanos}@cs.ntua.gr

Abstract. User attention recognition in front of a monitor or a specific task is a crucial issue in many applications, ranging from e-learning to driving. Visual input is very important when extracting information regarding a user's attention when recorded with a camera. However, intrusive equipment (special helmets, glasses equipped with cameras recording the eye movements, etc.) impose constraints on users spontaneity, especially when the target group consists of under aged users. In this paper, we propose a system for inferring user attention (state) in front of a computer monitor, only with the usage of a simple camera. The system can be used for real time applications and does not need calibration in terms of camera parameters. It can function under normal lighting conditions and needs no adaptation for each user.

Keywords: Head Pose, Eye Gaze, Facial Feature Detection, User Attention Estimation.

1 Introduction

For the estimation of a user's attention (state) in front of a computer monitor, two key issues are the movements (rotational and translational) of his head, as well as the directionality of his eye gaze. Not a lot of work has been published offering a combination of the two criteria, in order to evaluate the attentiveness or non-attentiveness of a person in front of a camera, especially without the need of specially designed hardware. In the current work, a top to down approach is presented, from head detection to fusion of biometrics using a specially trained neuro-fuzzy system, for evaluating the attentive state of the user.

There are techniques around the issue of head pose estimation which use more than one camera, or extra equipment ([1],[2],[3],[4],[5]), techniques based on facial feature detection ([6],[7]), suffering from the problem of robustness, or techniques that estimate the head pose ([8],[9]) using face bounding boxes, requiring the detected region to be aligned with a training set. In eye gaze estimation systems, the eye detection accuracy plays a significant role depending on the application (higher for security, medical and control systems). Some of the existing eye gaze

V. Kůrková et al. (Eds.): ICANN 2008, Part I, LNCS 5163, pp. 927–936, 2008.

techniques ([10],[11]), estimate the iris contours (circles or ellipses) on the image plane and, using edge operators, detect the iris outer boundaries. While not a lot of work has been done towards combining eye gaze and head pose information, in a non-intrusive environment, the proposed method uses a combination of the two inputs, together with other biometrics to infer the user's attention in front of a computer monitor, without the need of any special equipment apart from a simple web camera facing the user.

The structure of the paper is the following: Details on features used for User Attention Recognition are discussed in Section 2. Firstly, we discuss our facial feature detection method. Then, head pose and eye gaze estimation are analytically described. In Section 3, the training and testing of a Neuro-Fuzzy system is presented for characterizing the state of a user at frequent time intervals. Conclusions and future work are discussed in Section 4.

2 Feature Extraction for User Attention Recognition

For estimating a user's attention to the monitor of a computer, estimating his head pose, his eye gaze and his distance from the monitor are essential elements. To this aim, a neuro-fuzzy inference system was built, using as input the above metrics, as will be discussed hereafter. The system first detects the face of the user, when he is facing the camera frontally. Facial features are subsequently detected and tracked. Based on the facial feature movements, decisions regarding the user's head position and eye gaze are taken. The user's distance from the monitor is also estimated.

2.1 Facial Feature Detection

For face detection, the Boosted Cascade method described in [12] is employed, as it gives robust and real time results. The method output provides with the face region in which, usually, there exists background which can be asymmetric with regards to the right and left part of the region. An accurate estimate of the face region is important, since, for facial feature detection, blocks of predefined size will be used and, for this reason, accurate face detection is necessary. A postprocess method used in [13] was employed here for refining the face position in the frame.

For eye center localization, an approach based on [13] was used. For the detection of the eye corners (left, right, upper and lower) a technique similar to that described in [14] is used: Having found the eye center, a small area around it is used for the rest of the points to be detected.

In the current work, the point between the nostrils has also been used, as will be seen later. To this aim, nostrils have been detected at the frontal position of the user. Starting from the middle point between the eyes, an area of length equal to the inter-ocular distance is considered and the two darkest points are found to be the nostrils.

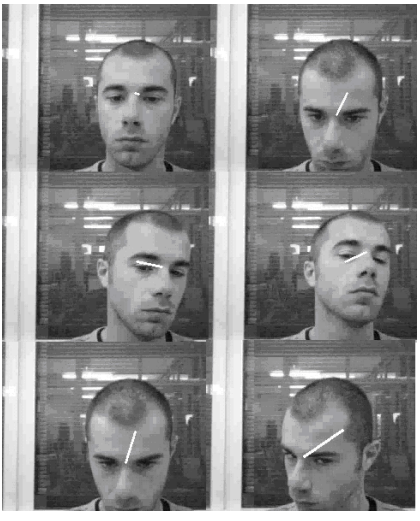

Fig. 1. Various head poses captured in front of a computer monitor using a simple web-camera. The white line (head pose vector) corresponds to the displacement of the point in the middle of the eyes with regards to the frontal pose.

2.2 Feature Tracking and Head Pose Estimation

Eye center tracking in a video sequence is done using a three-level Lucas-Kanade algorithm and, if only rotational movements of the face are considered, the movement of the middle point between the tracked eye centers can give a good insight of the face pose. This is achieved by subtracting the 2D image coordinates of its position at the frontal view of the face from those corresponding to a head rotation. The result can further be normalized by the inter-ocular distance in order the results to be scale independent. The resulting vector, from now on will be referred to as head pose vector (see Fig. 1).

To handle real application problems, where the user moves parallel and vertical to the camera plane, a series of rules has been extracted. If the user moves parallel to the image plane, the fraction between the inter-ocular distance and the vertical distance between the eyes and the nostrils remains almost constant. In this case, no rotation of the face is considered and, thus, the frontal pose is determined. Also, rapid rotations, apart from occluding some of the features (and, consequently, tracking is lost) make it difficult for the visible features to be tracked. In such cases, when the user comes back to his frontal position, the vector corresponding to pose estimation reduces in length and stays fixed for as long as the user is looking at the monitor. In these cases, the algorithm can re-initialize by re-detecting the face and the facial features.

2.3 Eye Gaze Estimation

For gaze detection, the areas defined by the four points around each eye are used. Prototype eye areas depicting right, left, upper and lower gaze directionality (see Fig. 2) are used to calculate mean greyscale images corresponding to each gaze direction. The areas defined by the four detected points around the eyes, are then correlated to these images.

Fig. 2. Prototype eye patches of different eye directionality

The normalized differences of the correlation values of the eye area with the left and right, as well as upper and lower mean gaze images are calculated:

$$H_r = \frac{(R_{r,l} - R_{r,r})}{max(R_{r,l}, R_{r,r}, R_{r,u}, R_{r,d})} \tag{1}$$

$$V_r = \frac{(R_{r,u} - R_{r,d})}{max(R_{r,l}, R_{r,r}, R_{r,u}, R_{r,d})} \tag{2}$$

$$H_l = \frac{(R_{l,l} - R_{l,r})}{max(R_{r,l}, R_{r,r}, R_{r,u}, R_{r,d})} \tag{3}$$

$$V_l = \frac{(R_{l,u} - R_{l,d})}{max(R_{r,l}, R_{r,r}, R_{r,u}, R_{r,d})} \tag{4}$$

Where $R_{i,j}$ is the correlation of the i (i=left,right) eye with the j (j=left, right, upper, lower) mean grayscale image. The normalized value of the horizontal and vertical gaze directionalities (conventionally, angles) are then the weighted mean:

$$H = ((2 - l) \cdot H_r + l \cdot H_l)/2 \tag{5}$$

$$V = ((2 - l) \cdot V_r + l \cdot V_l)/2 \tag{6}$$

where l is the fraction of the mean intensity in the left and right areas. This fraction is used to weight the gaze directionality values so that eye areas of greater luminance are favored in cases of shadowed faces. This pair of values (H,V) constitutes the gaze vector, as it will be called from now on in this paper. Typical cases of a person whose head is rotated frontally in relation to the monitor but his eyes are moving are shown in Fig. 3. The black line corresponds to the gaze vector whose magnitude declares the degree of the gaze away from the center of the field of view of the user.

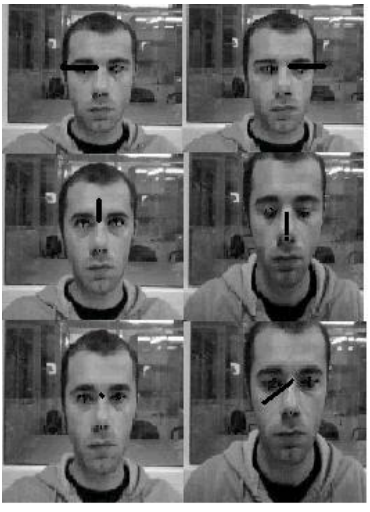

Fig. 3. Various gaze instances captured in front of a computer monitor using a simple web-cam. The black line (gaze vector) shows the magnitude of the gaze vector, as well as its directionality.

3 Neuro-fuzzy System Design and Evaluation

The target group of this research involves 20 learners at a certain level of education - they are 9 years old +/- 18 months. The selection of these children is based on the fact that they have not experienced problem of comprehension and use of language but they face barrier to learning caused by learning problems. One of the challenges that we had to face was the age of the participants of the experiment. As they are children, they are more active in the duration of the video recording. In addition, it was a unique opportunity to test the facial feature detection method that we use, as well as the eye-gaze and head-pose estimation algorithms to this difficult target group. The experimental tests took place in a familiar learning environment for the children: their school. In this framework, we collected videos of children from a Greek and a Danish school. We installed a web camera in the computer of the school and we asked each of the children to read an electronic document which was displayed in a 17' monitor.

As mentioned above, every time specific geometric criteria are fulfilled, the algorithm re-initializes. The first frame has to be a face looking at the camera frontally and, at this point, the pose vector has length equal to zero. As the person changes his head pose, this vector increases in length (see Fig. 4). Also, as a person's eyes look at different directions, the gaze vector changes in magnitude (see Fig. 5) [1].

[1] For privacy reasons, the videos used were from adults in our laboratory.

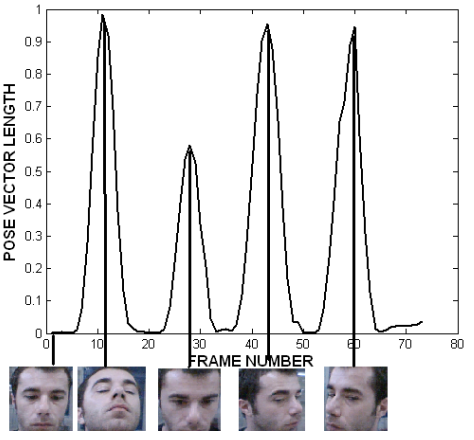

Fig. 4. Pose changes during a video of a person in front of a monitor

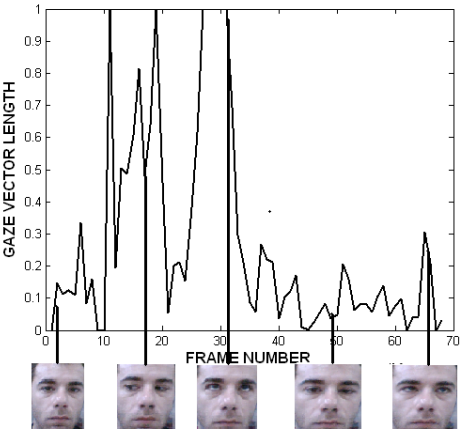

Fig. 5. Gaze changes during a video of a person not moving his head in front of a monitor

In order to estimate a user's attention, based on children's video files, a Sugeno-type fuzzy inference system was built. The metrics used in our case were the head pose vector, the gaze vector and the inter-ocular distance fractions between consecutive frames. We used 10 video segments, between 800 and 1200 frames each, and examined the means of the above metrics within video samples of 50 frames. Thus, for our experiments we had a total of 200 samples. The lengths of the videos were chosen so that, in each shot, all states were equally allocated. For training and testing, a leave-one-out approach was followed.

The pose and gaze vector magnitudes (see Fig. 4, 5) are values between zero and one, while the inter-ocular distance fraction was calculated as a fraction of

Table 1. Neuro-Fuzzy system decision accuracy

State	Average Error	%success
attentive	0.04	100
non-attentive	0.24	72
TOTAL	0.117	87.7

the inter-ocular distance measured every time the algorithm was re-initialized. Thus, the values of the inter-ocular distance between consecutive frames follows the distribution shown in Fig. 6. It can be noticed that the mean is shifted on the left. This is due to the fact that, as a face rotates right and left, the projected inter-ocular distance is reduced. In fact, these values were used, as will be seen later, together with pose and gaze vectors, for inferring a person's attention. The output's values, based on the annotation of the database, were either 1, declaring those time segments when the child was attentive, and 0 for those periods when the child was not paying attention to the monitor.

Prior to training, our data were clustered using the sub-cluster algorithm [15]. This algorithm, instead of using a grid partition of the data, clusters them and, thus, leads to fuzzy systems deprived of the curse of dimensionality. For clustering, many radius values for the cluster centers were tried and the one that gave the best trade-off between complexity and accuracy was 0.1 for all inputs. The number of clusters created by the algorithm determines the optimum number of the fuzzy rules. After defining the fuzzy inference system architecture, its parameters (membership function centers and widths), were acquired by applying a least squares and back-propagation gradient descent method [16].

After training, the output surface using the pose vector length and gaze vector length are shown in Fig. 7, and the respective surface using the pose vector length

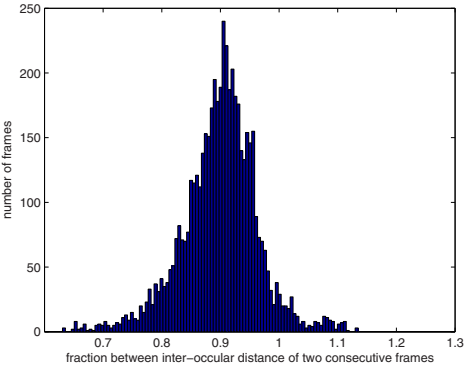

Fig. 6. Distribution of the fraction of the inter-ocular distance between consecutive frames

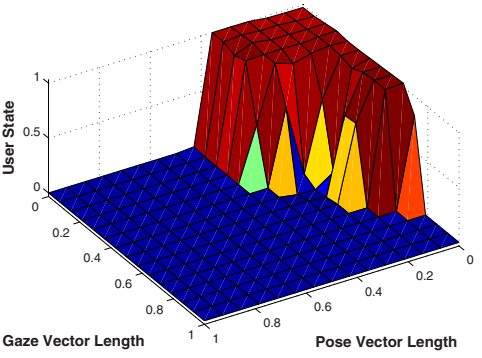

Fig. 7. Output surface using the pose vector length and gaze vector length

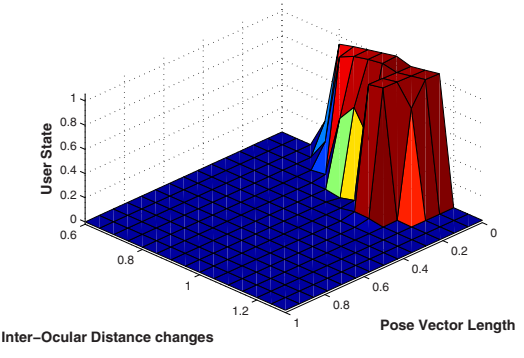

Fig. 8. Output surface using the pose vector length and inter-ocular distance fraction between successive frames

and inter-ocular distance fraction as inputs is shown in Fig. 8. It can be seen from the above figures that, for small values of the pose vector length and the gaze vector length, the output takes values close to one (attentive), while, as gaze or pose vector lengths increase in magnitude, the output's value goes to zero. Similarly, it can be seen that , for small values of the pose vector length, and for small changes of the inter-ocular distance (the fraction takes values close to 1), the user can be considered as attentive (the output is close to one). Large values of the pose vector length mean that the user is non-attentive (the output is close to zero), while sudden, large changes of the inter-ocular distance mean that the user is non-attentive. Usually, these sudden changes occur when the user rotates rapidly right and left. The following table summarizes the results of our approach. In the column denoted as average error, the average absolute error between the output and the annotation is reported, while in the second column, a crisp decision is considered, meaning that a state is inferred to be attentive if the output value is above 0.5 and non-attentive if it is below 0.5.

4 Conclusions and Future Work

In this work, the steps of a method for user attentiveness evaluation in front of a computer monitor was presented. The advantages of the method are that it does not require any special setup in terms of hardware, it runs real-time, it is user independent and can work efficiently under uncontrolled lighting conditions. A neuro-fuzzy system was developed for the evaluation of the state of the user. The results of using a neuro-fuzzy system showed that, discriminating between attentiveness and non-attentiveness does not have to be a crisp decision but takes into account the magnitude of the biometrics it uses in order to characterize a user with a certain degree of certainty. Future work shall include more biometric measurements (mouth, hands movements, frowning of the eyebrows) in order to develop a system that can discriminate among more states (tiredness, frustration, distraction).

Acknowledgments

This work was partially funded by European Commission IST Project 'AGENT-DYSL' (under contract FP6 IST-2005-2.5.11 e-Inclusion-034549) and by IST Project 'FEELIX',(under contract FP6 IST-045169).

References

1. Voit, M., Nickel, K., Stiefelhagen, R.: Multi-view head pose estimation using neural networks. In: Second Canadian Conference on Computer and Robot Vision (CRV), Victoria, BC, Canada, pp. 347–352. IEEE Computer Society, Los Alamitos (2005)
2. Mao, Y., Suen, C.Y., Sun, C., Feng, C.: Pose estimation based on two images from different views. In: Eighth IEEE Workshop on Applications of Computer Vision (WACV), Washington, DC, USA, p. 9. IEEE Computer Society, Los Alamitos (2007)
3. Beymer, D., Flickner, M.: Eye gaze tracking using an active stereo head. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Madison, WI, USA, vol. 2, pp. 451–458. IEEE Computer Society, Los Alamitos (2003)
4. Meyer, A., Böhme, M., Martinetz, T., Barth, E.: A single-camera remote eye tracker. LNCS (LNAI), pp. 208–211. Springer, Heidelberg (2006)
5. Hennessey, C., Noureddin, B., Lawrence, P.D.: A single camera eye-gaze tracking system with free head motion. In: Proceedings of the Eye Tracking Research & Application Symposium (ETRA), San Diego, California, USA, pp. 87–94. ACM, New York (2006)
6. Gee, A., Cipolla, R.: Non-intrusive gaze tracking for human-computer interaction. In: Int. Conference on Mechatronics and Machine Vision in Pract., Toowoomba, Australia, pp. 112–117 (1994)
7. Gourier, N., Hall, D., Crowley, J.: Estimating face orientation from robust detection of salient facial features. In: International Workshop on Visual Observation of Deictic Gestures (ICPR), Cambridge, UK (2004)

8. Seo, K., Cohen, I., You, S., Neumann, U.: Face pose estimation system by combining hybrid ica-svm learning and re-registration. In: 5th Asian Conference on Computer Vision, Jeju, Korea (2004)
9. Stiefelhagen, R.: Estimating Head Pose with Neural Networks - Results on the Pointing 2004 ICPR Workshop Evaluation Data. In: Pointing 2004 Workshop (ICPR), Cambridge, UK (August 2004)
10. Daugman, J.: High confidence visual recognition of persons by a test of statistical independence. IEEE Trans. Pattern Anal. Mach. Intell. 15(11), 1148–1161 (1993)
11. Deng, J.Y., Lai, F.: Region-based template deformation and masking for eye-feature extraction and description. Pattern Recognition 30(3), 403–419 (1997)
12. Viola, P.A., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: Conference on Computer Vision and Pattern Recognition (CVPR), Kauai, HI, vol. 1, pp. 511–518 (December 2001)
13. Asteriadis, S., Nikolaidis, N., Pitas, I., Pardàs, M.: Detection of facial characteristics based on edge information. In: Second International Conference on Computer Vision Theory and Applications(VISAPP), Barcelona, Spain
14. Zhou, Z.H., Geng, X.: Projection functions for eye detection. Pattern Recognition 37(5), 1049–1056 (2004)
15. Chiu, S.L.: Fuzzy Model Identification Based on Cluster Estimation. Journal of Intelligent and Fuzzy Systems 2(3) (1994)
16. Jang, J.S.R.: ANFIS: Adaptive-Network-Based Fuzzy Inference System. IEEE Transactions on Systems, Man, and Cybernetics 23, 665–684 (1993)