Multimed Tools Appl
DOI 10.1007/s11042-008-0240-1

Estimation of behavioral user state based on eye gaze and head pose—application in an e-learning environment

**Stylianos Asteriadis · Paraskevi Tzouveli ·
Kostas Karpouzis · Stefanos Kollias**

Abstract Most e-learning environments which utilize user feedback or profiles, collect such information based on questionnaires, resulting very often in incomplete answers, and sometimes deliberate misleading input. In this work, we present a mechanism which compiles feedback related to the behavioral state of the user (e.g. level of interest) in the context of reading an electronic document; this is achieved using a non-intrusive scheme, which uses a simple web camera to detect and track the head, eye and hand movements and provides an estimation of the level of interest and engagement with the use of a neuro-fuzzy network initialized from evidence from the idea of Theory of Mind and trained from expert-annotated data. The user does not need to interact with the proposed system, and can act as if she was not monitored at all. The proposed scheme is tested in an e-learning environment, in order to adapt the presentation of the content to the user profile and current behavioral state. Experiments show that the proposed system detects reading- and attention-related user states very effectively, in a testbed where children's reading performance is tracked.

Keywords User attention estimation · Head pose · Eye gaze ·
Facial feature detection and tracking

1 Introduction

Nowadays, it is widely accepted that putting information technology to use can enhance the learning experience: learning methods are becoming more and more portable, flexible, and

S. Asteriadis (✉) · P. Tzouveli · K. Karpouzis · S. Kollias
Image, Video and Multimedia Systems Laboratory, School of Electrical & Computer Engineering,
National Technical University of Athens, 157 80 Zographou, Athens, Greece
e-mail: stiast@image.ntua.gr

P. Tzouveli
e-mail: tpar@image.ntua.gr

K. Karpouzis
e-mail: kkarpou@image.ntua.gr

S. Kollias
e-mail: stefanos@cs.ntua.gr

adaptive, while the internet has been broadly adopted as a medium for network-enabled transfer of skills, information and knowledge, and plays a significant role in all fields of education [10]. Internet-oriented applications try to satisfy current educational needs, closing the gap between traditional educational techniques and future trends in technology-blended education. Towards this goal, various developed e-learning systems miss functionalities like educational multimedia environments, personalized capabilities and tracking of learners' input and relevance feedback [26]. Several approaches have been proposed in order to collect background and preference information about learners, providing adaptation capabilities to such systems. In this framework, non-verbal interaction can constitute a valuable extension to an e-learning system, since it presents users with the opportunity to provide feedback to the system about related preferences or dispositions without the need to answer specific questions or fill in questionnaires, usually predetermined by the system developers, or enter free text which is hard to parse semantically. From the part of the system, a non-intrusive method of receiving this kind of information is much more reliable, since it produces replicable results, and can be deployed when users actually operate the system via an interface, alleviating the need for an additional step.

In this paper, we build on information estimated from behavior-related features (attention, level of interest) of users reading documents in a computer screen. More specifically, we use the position and movement of prominent points around the eyes and the position of the irises to reconstruct vectors which illustrate the direction of gaze and head pose. These vectors provide an indication of whether the particular user looks into the screen or not and whether their eyes are fixed at a particular spot for long periods of time. Furthermore, as a complementary task, the position of the hands is also used. This information is then used to determine the behavioral state of the user towards the electronic document, i.e. the level of interest and attention, based on concept from the Theory of Mind and annotation from experts in e-education systems. Testing of this approach has been done on a database of children in front of a computer screen, by training a neuro-fuzzy system with the annotated information, and showed that our system can work as an non-intrusive means for monitoring a child's state towards what he/she is reading, without him/her to feel pushed to react in certain ways and, thus, lose his/her spontaneity. To this aim, we describe a monocular method which measures head and eyes motion, as well as motion of more prominent features.

1.1 From gaze and pose to behavioral states—the testbed application

Understanding intentions of others is an important ability that involves representing the mental states of others in one's own mind; in our case, a machine learning infrastructure. The principles and techniques that humans deploy in order to understand, predict, and manipulate the behavior of other humans is collectively referred to as a "theory of mind" (ToM; [3]). Software agents can utilize a ToM for two purposes: to anticipate the behavior of other agents [5] (e.g., preparing for certain actions that the other will perform), and to manipulate it (e.g., trying to bring the other agent in such a certain state so as to perform certain desired actions, or not perform certain unwanted actions). In a Human–Computer Interaction framework, such as the one in our case, this would translate to tracking the behavior of users, based on how they perform their reading exercise, identifying certain characteristics (e.g. staring at a fixed point for some seconds or looking away from the screen) and adapting the system to cater for the detected user state. Development of formal models for ToM has been discussed in [5] and [29]; usually, such models focus on the epistemic (e.g., beliefs) and/or motivational states (e.g., desires, intentions) of other agents, while also attempting to take into account emotions as well. This idea is in line with the claim that humans have a ToM that is not only about beliefs, desires, and intentions, but also about other mental states like emotional and attentional states [18]; in our approach, we do not attempt to identify what lies *behind* the detected user state (i.e., *why* the user appears distracted), modeling a set of beliefs, desires and intentions in the process, but utilize the state information to adapt the reading experience. More specifically, we utilize findings from ToM to relate direction of gaze towards the screen to the level of interest to the displayed document and, vice versa, staring away from the screen to distraction or lack of interest, depending on the time span. In addition to this, sudden and abrupt, as well as repeating movements are related to nervousness and frustration, which along with distraction and lack of interests are usual reading behavior states, especially in the presence of dyslexia which hampers reading performance.

The reading testbed [46] consists of an interface which displays an electronic document in PDF format and is targeted towards identifying reading problems related to dyslexia. Most of these reading problems [44] have to do with uttering words in the wrong order, missing out words or rearranging syllables within words. To alleviate this, the reading testbed highlights words in the text in the correct order, so as to help the reader with keeping the correct sequence; if the behavioral state detection module detects that the user is not looking at the screen, highlighting stops since the user is not reading anymore. In addition to this, in the case of consistent distraction or other detected reading errors, the user interface also changes the font size, to further assist the user; check [46] for a complete description of the deployed testbed.

Training of the state recognition system was based on processing test data from the deployment of this interface, annotated by electronic education experts in the framework of the Agent-Dysl FP6 Project [17]. Annotation identified video segments with particular user states of interest and related these states to detectable and measurable features. For example, annotation of a video segment labeled ‘distracted’, explained that the user is ‘distracted’ since ‘head pose is away from the screen’, in essence relating a concept described in ToM to a feature detectable with computer vision techniques. As a result, annotation of a video database obtained from children reading early school texts was put to use to train a neuro-fuzzy network, which was then used to detect behavior states from pose and gaze features. Since the aim of the project is to deploy this software to assist dyslexic children, we did not attempt to generalize this approach to adult users; in such a case, retraining and adaptation strategies [7] can be deployed to adapt the trained network to the reading patterns of a new user, without the need to train it from scratch in a supervised manner. As a general rule, the user interface takes into account the particular design requirements for this target user group [28]; in the context of the proposed work, dealing with children ([9, 42]) also means that the system should be able to account for additional reading-related antics, such as head and body movement, besides looking away from the screen and fixing on a particular spot.

1.2 Feature extraction and state recognition

There are techniques around the issue of head pose estimation which use more than one camera, or extra equipment for head pose estimation [4, 21, 32, 48, 52], techniques based on facial feature detection [19, 20], suffering from the problem of robustness, or techniques that estimate the head pose [38, 40] using face bounding boxes, requiring the detected region to be aligned with the training set. Eye gaze, that is, the direction to which the eyes are pointing in space, is a strong indicator of the focus of attention, and has been studied extensively [16]. Eye tracking systems can be grouped into head-mounted or remote, and infra-red-based or appearance-based. Infra-red-based gaze tracking systems employ the so-called red-eye effect, i.e., the difference in reflection between the cornea and the pupil. Appearance-based gaze trackers employ computer vision techniques to find the eyes in the input image and then determine the orientation of the irises. While head-mounted systems are the most accurate, they are also the most intrusive. Infrared systems are more accurate than appearance-based, but there are concerns over the safety of prolonged exposure to infra-red lights. In eye gaze estimation systems using computer vision techniques, 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 techniques [14, 15], 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 screen, without the need of any special equipment apart from a simple web camera facing the user.

Also, in the research literature, not a lot of work has been published in the area of human–computer interaction regarding the issue of user attention recognition, especially if unconstrained and non-intrusive environments are to be considered. The biggest part of relative research is towards estimating the field of attention of a person in meeting conditions. For example, in [41], a method is presented for estimating the focus of attention in a meeting environment. In this approach, the attention of a person at a meeting table is estimated based on acoustics (who is speaking) and visual information. An omni-directional camera is used to detect faces and neural networks are employed for estimating the head pose of each participant. Eye gaze information is not taken into account in this work, but a Bayesian approach is followed in order the focus of attention for each participant to be detected based on his head pose estimates. In [2], the focus of attention of meeting participants is also estimated. In this approach, also head pose is used, and environmental targets (including other participants) are marked as possible points of attention. Maximum a Posteriori probabilities, or Hidden Markov Models (HMM) are used for the evaluation of the point of attention of a participant. In [31], the attention of a user is estimated in an office environment, and in particular at a workstation including a PC and other objects. A pair of cameras is used for capturing the user's face and 3D models of face and eyes are used to determine his head pose and eye gaze. In [35], the authors investigate the structure of a conversation in four-people meetings, in a closed environment. In this work, head motions and utterance presence/absence are used for inferring the roles of participants in conversational models (speaker, addressees, side-participants). The motion of the head is calculated using magnetic sensors applied on the head of each user and probabilistic models are built. In [49], a study on the effectiveness of a facial movement tracking system in front of a computer, in order to evaluate the relation between facial and physiological responses of users in front of a computer screen, based on events produced by the computer, is conducted. Facial movement differences in front of different computer-based events were also analysed. The experiments used 15 participants and a quiz provoking, either amusement or surprise. Strong stimuli proved to trigger significant facial movement, and weaker ones evoked facial reactions, but only to a limited extent. Furthermore, physiological and facial responses did not always concur. However, it is proved in this work that a non-intrusive facial feature tracking system can be an important source of information in Affective Computing. Similar to user attention estimation, a lot of work has appeared in literature around the issue of facial expression recognition. For example, in

[27], a non-intrusive way to analyze facial data for Facial Expression Classification and Affect Interpretation is used. The authors employ an infrared camera and run experiments on a dataset consisting of 21 participants adopting neutral and four facial expressions: Happiness, disgust, sadness and fear. After being tested for normality, the thermal intensity values of certain facial features undergo Principal Component Analysis and LDA is further used for the classification of intentional facial expressions.

Much work has also been done in the field of driver's attention recognition. Typical works are reported [39] and [13]. In [39], a top-to-down approach for estimating the driver's attention is presented, using a monocular system, without the need of special set-up in terms of hardware, or calibration. In this system, the authors detect and track the eyes and the lips of the driver, using color predicates. The geometrical layout of these features gives a good input for estimating the head rotation, as well as the eye gaze direction. Furthermore, the eyes closure is also used. The above metrics are used to evaluate the attention of a driver by using three finite state automata: One using eyes closure and the other two using up/down and right/left head rotations respectively. In [13], the authors detect driver vigilance by detecting the eyes and discriminating between eyes open or closed. For this purpose, experiments were carried out in real conditions, with a camera placed in the central part of the car dashboard and a Gaussian model is built to describe the driver's normal conditions. Eye closures for certain amount of time denote inattention or fatigue.

1.3 System architecture

In the proposed method, the face and facial features of the user in front of a computer monitor are first detected. All is needed is a simple webcam, and no calibration or learning is required for each different user. Based on the head movements and the eye areas, head pose and eye gaze are estimated. The distance changes of the user from the computer screen are also monitored. More characteristics (mouth opening, eyebrows frowning, hand movements) can also be detected and play their role in determining the user's attention to what he/she is reading. Experiments were carried out using part of the detected biometrics to estimate the state of the user and results show that our system can be used for real-time situations, as for example in an e-learning system. In this direction, we have adopted the extended version [45] of IEEE Reference Model (WG) of the Learning Technology Standards Committee in which a profiling procedure is included. The detected biometrics of the user are used to determine the user profile and adapt the provided multimedia presentation to the current user needs. A diagram that illustrates the steps employed in the method for estimating user states is shown in Fig. 1. More precisely, face detection is done first. Starting from the face area, hands are detected and tracked, and their movements feed the User Attention Estimation module. Furthermore, within the face area, facial features are detected and tracked in subsequent frames. Based on facial feature tracking, (see "Section 3.2", Table 1), user states are inferred, directly (by tracking e.g. the eyebrow or mouth movements) and indirectly (by calculating the eye gaze, the head pose and the movements of the user back and forth) through a Neuro Fuzzy inference system (see "Section 5").

This paper is organized as follows: "Section 2" describes the lower-level processes used to locate and track the prominent facial feature points, the eye gaze, the head pose and the hands positions. "Section 3" correlates this information with particular user states, providing actual results from system deployment. "Section 4" presents the e-learning framework including the non-verbal feedback and "Section 5" presents experimental results. "Section 6" concludes the paper.

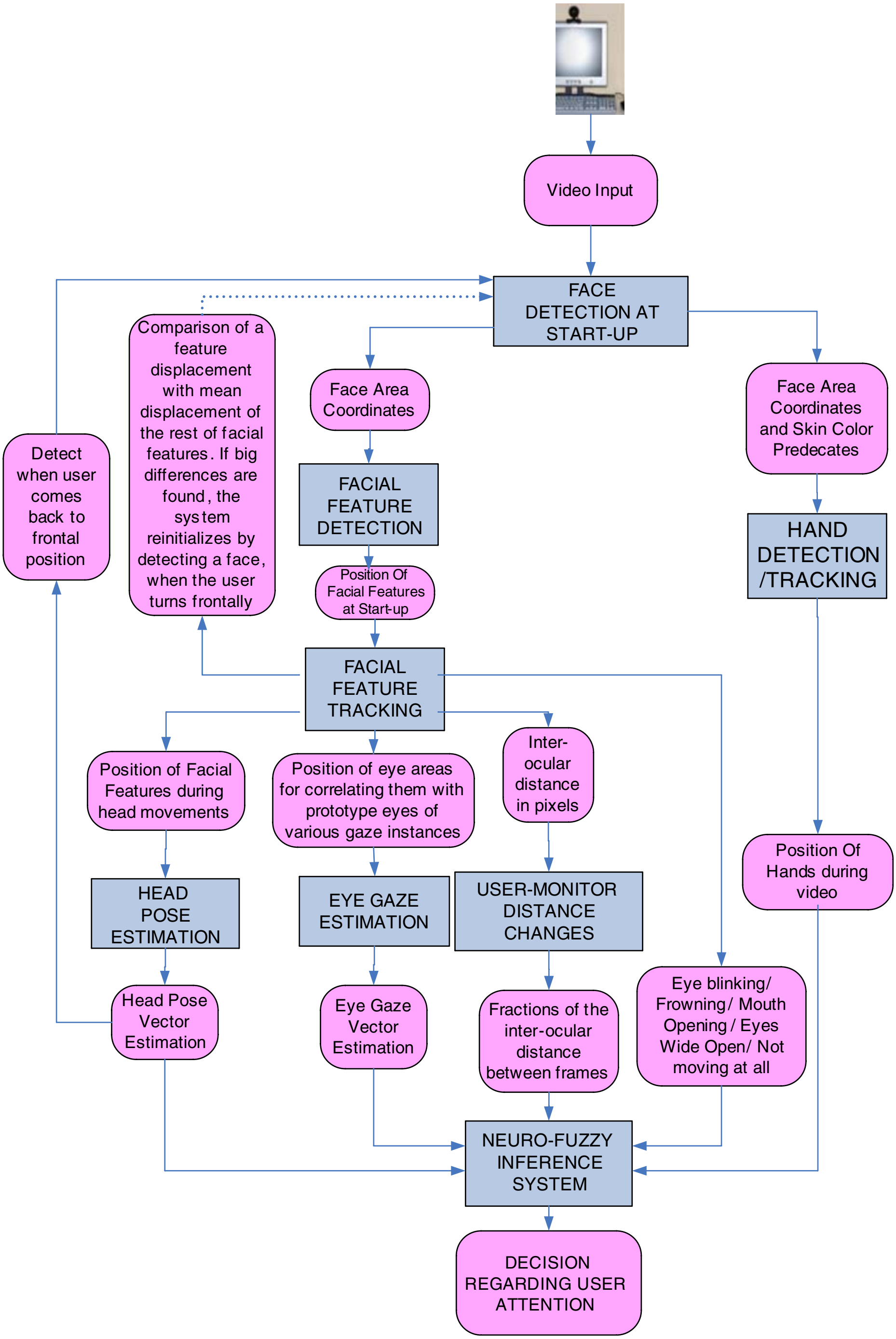

Video Input
FACE DETECTION AT START-UP
Face Area Coordinates
Face Area Coordinates and Skin Color Predecates
FACIAL FEATURE DETECTION
HAND DETECTION /TRACKING
Position Of Facial Features at Start-up
Detect when user comes back to frontal position
Comparison of a feature displacement with mean displacement of the rest of facial features. If big differences are found, the system reinitializes by detecting a face, when the user turns frontally
FACIAL FEATURE TRACKING
Position of Facial Features during head movements
Position of eye areas for correlating them with prototype eyes of various gaze instances
Inter-ocular distance in pixels
Position Of Hands during video
HEAD POSE ESTIMATION
EYE GAZE ESTIMATION
USER-MONITOR DISTANCE CHANGES
Head Pose Vector Estimation
Eye Gaze Vector Estimation
Fractions of the inter-ocular distance between frames
Eye blinking/ Frowning/ Mouth Opening / Eyes Wide Open/ Not moving at all
NEURO-FUZZY INFERENCE SYSTEM
DECISION REGARDING USER ATTENTION

Fig. 1 Steps employed for estimating user attention. *Rounded rectangles* are input/output from one component to the other, while *non-rounded rectangles* are the various algorithmic parts (components) of the approach

Table 1 Possible state classification

Visual evidence	Behavioral evidence	State
Eyes is not looking at the screen Head is moving (direction/speed) Eyes blinking Severity "Frozen" lips "Frozen" face Mouth open	Head motion Blink of the eyes Frown	Frustrated/struggling to read
Eyes is not looking at the screen Eyes wide open Head is moving(direction/speed)	Somebody talks to learner/a noise is heard	Distracted
Eyes looking at the screen Eyes wide open Mouth open Hand(s) covering eyes Hand covering mouth	Yawns/tries to stop reading by speaking to someone	Tired/sleepy
Eyes looking at the screen Eyes wide open Head is moving(direction/speed) Hand(s) covering eyes Hand covering mouth	Not look at the screen/speaks to someone/stops reading	Not paying attention
Eyes looking at the screen Eyes wide open Head is not moving (direction/speed)	Stares at the screen/ makes a remarkable effort to interact, facing great difficulties	Attentive
Severity Eyes looking at the screen Eyes wide open Mouth open Head is moving(direction/speed)	Stares at the screen/find something remarkable	Full of interest

2 Detection and tracking of facial features—gaze and pose estimation

2.1 Facial feature detection and tracking

Facial feature extraction is a crucial step to numerous applications such us face recognition, human–computer interaction, facial expression recognition, surveillance and gaze/pose detection. In their vast majority, the approaches in bibliography use face detection as a preprocessing step. This is usually necessary in order to tackle with scale problems as, localizing a face in an image is more scale-independent than starting with the localization of special facial features. When only facial features are detected (starting from the whole image and not from the face region of interest), the size and the position of the face in the image have to be pre-determined and, thus, such algorithms are devoted to special cases, such as driver's attention recognition [39] where the user's position with regards to a camera is almost stable. In such techniques, color [39] predicates, shape of facial features and their geometrical relations [12] are used as criteria for the extraction of facial characteristics.

On the other side, facial feature detection is more scale-independent when the face is detected as a preprocessing step. In this case, the face region can be normalized to certain dimensions, making the task of facial feature detection more robust. For example, in [11] a multi-stage approach is used to locate features on a face. First, the face is detected using the boosted cascaded classifier algorithm by Viola and Jones [47]. The same classifier is trained using facial feature patches to detect facial features. A novel shape constraint, the Pairwise Reinforcement of Feature Responses (PRFR) is used to improve the localization accuracy of the detected features. In [25], a three stage technique is used for eye center localization. The Hausdorff distance between edges of the image and an edge model of the face is used to detect the face area. At the second stage, the Hausdorff distance between the image edges and a more refined model of the area around the eyes is used for more accurate localization of the upper area of the head. Finally, a Multi-Layer Perception (MLP) is used for finding the exact pupil locations. In [23], an SVM-based approach is used for face detection. Following, eye-areas are located using a feed-forward neural network and the face is brought to a horizontal position based on the eye positions. Starting from these points, edge information and luminance values are used for eyebrow and nostrils detection. Further masks are created to refine the eye positions, based on edge, luminance and morphological operations. Similar approaches are followed for the detection of mouth points.

In this work, prior to eye and mouth region detection, face detection is applied on the face images. The face is detected using the Boosted Cascade method, described in [47]. The output of this method is usually the face region with some background. Furthermore, the position of the face is often not centered in the detected sub-image. Since the detection of the eyes and mouth will be done in blocks of a predefined size, it is very important to have an accurate face detection algorithm. Consequently, a technique to post-process the results of the face detector is used by applying an ellipse fitting algorithm [1].

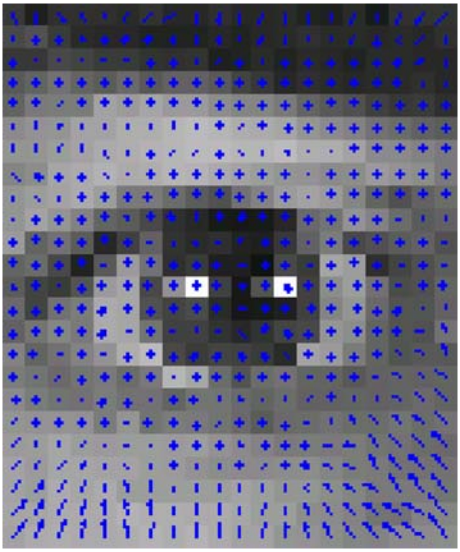
A template matching technique follows for the facial feature areas detection step [1]: The face region found by the face detection step is brought to certain dimensions and the corresponding edge map is extracted. Subsequently, for each pixel on the edge map, a vector pointing to the closest edge is calculated and its x, y coordinates are stored. The final result is a vector field encoding the geometry of the face. Prototype eye patches were used for the calculation of their corresponding vector fields (Fig. 2) and the mean vector field was used as prototype for searching similar vector fields on areas of specified dimensions on the face vector field.

For the eye center detection, the area of the eye, found at the previous step, is brought back to its initial dimensions on the image and a light reflection removal step is employed. The grayscale image of the eye area is converted into a binary image and small white connected components are removed. The areas that correspond to such components on the original image are substituted by the average of their surrounding area. The final result is an eye area having reflections removed. Subsequently, horizontal and vertical derivative maps are extracted from the resulting image and they are projected on the vertical and horizontal axis respectively. The mean of a set of the largest projections is used for an estimate of the eye center. Following, a small window around the detected point is used for the darkest patch to be detected, and its center is considered as the refined position of the eye center.

For the detection of the eye corners and eyelids (left, right, upper and lower eye points) a technique similar to that described in [53] is used: Having found the eye center, a small area around it is used for the rest of the points to be detected and Generalized Projection Functions (GPFs) are calculated. Local maxima of the above functions are used to declare the positions of the eye boundaries.

For the mouth area localization, a similar approach to that of the eye area localization is used: The vector field of the face is used and template images are used for the extraction of a prototype vector field of the mouth area. Subsequently, similar vector fields are searched for on the lower part of the normalized face image. However, as the mouth has, many times,

Fig. 2 The vector field of an eye area

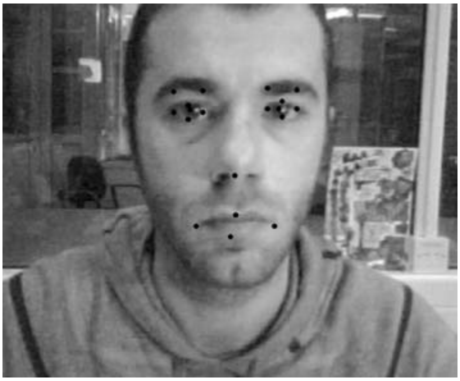


similar luminance values with its surrounding skin, an extra factor is also taken into account. That is, at every search area, the mean value of the hue component is calculated and added to the distance from the mean vector fields of the mouth. Minimum values declare mouth existence.

For the extraction of the mouth points of interest (mouth corners and mouth upper/lower points), the hue component is also used. Based on the hue values of the mouth, the detected mouth area is binarized and small connected components whose value is close to 0° are discarded similar to the light reflection removal technique employed for the eyes. The remainder is the largest connected component which is considered as the mouth area. The leftmost and rightmost points of this area are considered as the mouth corners, while the upper and lower points of this area correspond to the upper and lower points of the lips.

In our approach, eyebrows are also detected. The eyebrows are features of high intensity contrast and detail. Starting from each eye-corner, a line segment is expanded upwards, with length set equal to the eye length. The darkest point of this segment gives a very good estimate of a point on the eyebrow. An example of detected feature points is shown in Fig. 3.

Fig. 3 Detected facial features



Once the positions of the facial feature points of interest are known on a frontal face, tracking can follow. In this way, gaze detection and pose estimation can be determined, not only on a single frame, but on a series of frames. Also, calculating changes of the inter-ocular distance in a series of frames, it is easy to determine the distance changes of a user from the camera. Furthermore, tracking saves computational time, since detecting the characteristics at every frame is more time demanding, and tracking can achieve better results in cases of large displacement of the face from its frontal position. In our case, tracking was done using an iterative, three-pyramid Lucas–Kanade tracker [6]. An example of a person's movement with relation to the camera is shown in Fig. 4.

2.2 Gaze detection and pose estimation

In the current work, features are detected and tracked, allowing for a relative freedom of the user. Under these circumstances, the gaze directionality can be approximately determined, which is enough for attention recognition purposes, as well as for general decisions regarding one's gaze. For gaze detection, the area defined by the four points around the eye is used. Eye areas depicting right, left, upper and lower gaze directionalities are used to calculate mean grayscale images corresponding to each gaze direction. The areas defined by the four detected points around the eyes, are then correlated to these images. 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 with the following equations:

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

$$H_l = \frac{\left(R_{l,l} - R_{l,r}\right)}{\max\left(R_{l,l}, R_{l,r}, R_{l,u}, R_{l,d}\right)}, V_l = \frac{R_{l,u} - R_{l,d}}{\max\left(R_{l,l}, R_{l,r}, R_{l,u}, R_{l,d}\right)},$$

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

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

$$V = ((2 - l) \cdot V_r + l \cdot V_l)/2$$

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

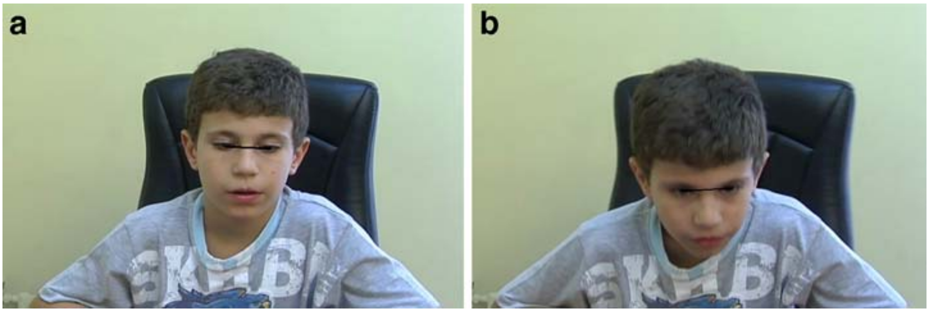


Fig. 4 Example of determining a person's movement towards the camera

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 (see Fig. 5).

If only rotational movements of the face are considered, the movement of the middle point between the eyes can give a good insight of the face pose. This is achieved if we have a priori knowledge of the frontal pose of the face, and keep the movements of this point with regards to its frontal position. These movements are towards the head rotation and give a good insight of the face pose, thus forming the head pose vector. The head pose vector can further be normalized by the eye distance in order the results to be scale independent. Face pose estimation results taken in office lighting conditions with a webcam are shown in Fig. 6. To handle real application problems, where the user moves parallel and vertical to the camera plane, a series of rules has been extracted to recover tracking failures. To this aim, nostrils have been detected at the frontal view of the user. Starting from the middle point between the eyes, an orthogonal area of length equal to the inter-ocular distance is considered and the two darkest points are found to be the nostrils.

If the user moves parallel to the image plane, the inter-ocular distance and the vertical distance between the eyes and the nostrils remain 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 (where, consequently, tracking is lost) make it difficult for the visible features to be tracked. In such cases, when the user comes back to a frontal position, the vector corresponding to pose estimation reduces in length and stays fixed for as long as the user is looking at the screen. In these cases, the algorithm can reinitialize by re-detecting the face, the facial features and tracking. Further constraints can be imposed by considering the mean displacement of the facial characteristics.

At each frame, the mean displacement of the facial features coordinates is calculated. If the displacement of a feature is below or above certain fractions of the mean displacement, the mean displacement of the rest of the features is recalculated and the new position of the outlier is acquired by considering its position at the previous frame shifted by the mean displacement of the rest of the features.

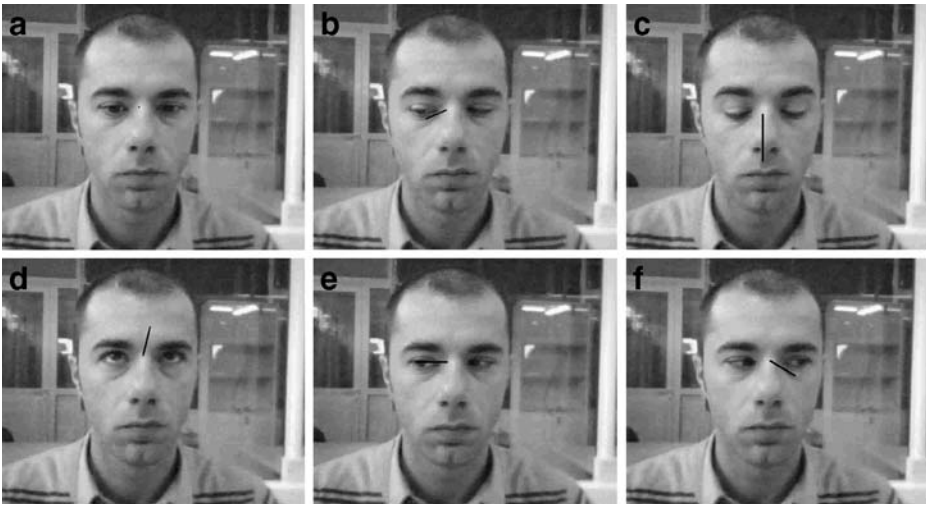


Fig. 5 **a**–**f** Various gaze instances captured in front of a computer screen using a simple web-cam. The *black line* (gaze vector) shows the magnitude of the gaze vector, as well as its directionality

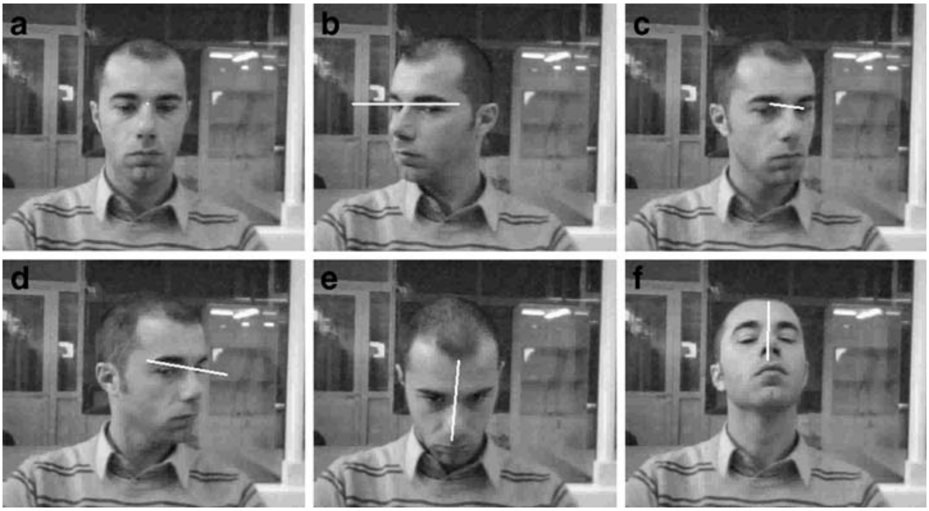


Fig. 6 **a**–**f** Various head pose instances captured in front of a computer screen using a simple web-cam. The *white line* (head pose vector) shows the magnitude of the head pose vector, as well as its directionality

2.3 Hand detection and tracking

Regarding gesture analysis, several approaches have been reviewed for the head–hand tracking module all of them mentioned both in [50] and in [34]. From these only video based methods were considered since motion capture or other intrusive techniques would interfere with the person's emotional state. The major factors taken under consideration are computational cost and robustness, resulting in an accurate near real-time skin detection and tracking module. The general process involves the creation of *moving skin* masks, namely skin color areas that are tracked between subsequent frames. By tracking the centroid of those masks, we produce an estimate of the user's movements. A priori knowledge concerning the human body and the circumstances when filming the gestures was incorporated into the module indicating the different body parts (head, right hand, left hand).

For each frame a skin color probability matrix is computed by calculating the joint probability of the Cr/Cb image values. The *skin color* mask is then obtained from the skin probability matrix using thresholding. Possible moving areas are found by thresholding the difference between the current frame and the next, resulting in the *possible-motion* mask. This mask does not contain information about the direction or the magnitude of the movement, but is only indicative of the motion and is used to accelerate the algorithm by concentrating tracking only in moving image areas. Both color and motion masks contain a large number of small objects due to the presence of noise and objects with color similar to the skin. To overcome this, morphological filtering is employed on both masks to remove small objects.

All described morphological operations are carried out with a disk-structuring element with a radius of 1% of the image width. The distance transform of the color mask is first calculated and only objects above the desired size are retained. These objects are used as markers for the morphological reconstruction of the initial color mask. The color mask is then closed to provide better centroid calculation.

For the next frame, a new moving skin mask is created, and a one-to-one object correspondence is performed. Object correspondence between two frames is performed on the color mask and is based on object centroid distance for objects of similar (at least 50%) area. In the case of hand object merging and splitting, e.g., in the case of clapping, we establish a new matching of the left-most candidate object to the user's right hand and the right-most object to the left hand. The Sagittal plane information of the gesture was ignored since it would require depth information from the video stream and it would make the performance of the proposed algorithm very poor or would require a side camera and parallel processing of the two streams. The object correspondence heuristic makes it possible to individually track the hand segments correctly, at least during usual meaningful gesture sequences (see Fig. 7). In addition, the fusion of color and motion information eliminates any background noise or artifacts, thus reinforcing the robustness of the proposed approach. More details on the specified method can be found in [30].

3 User states

3.1 Prerequisites for the detection of user states

The human face is the site for major sensory inputs and major communicative outputs. It houses the majority of our sensory apparatus, as well as our speech production apparatus. It is used to identify other members of our species, to gather information about age, gender, attractiveness, and personality, and to regulate conversation by gazing or nodding. Moreover, the human face is our pre-eminent means of communicating and understanding somebody's affective, cognitive, and other mental states and intentions on the basis of the shown facial expression. Hence, automatic analysis of the human face is indispensable in the context of natural HCI [36]. Furthermore, as most existing face detection algorithms are

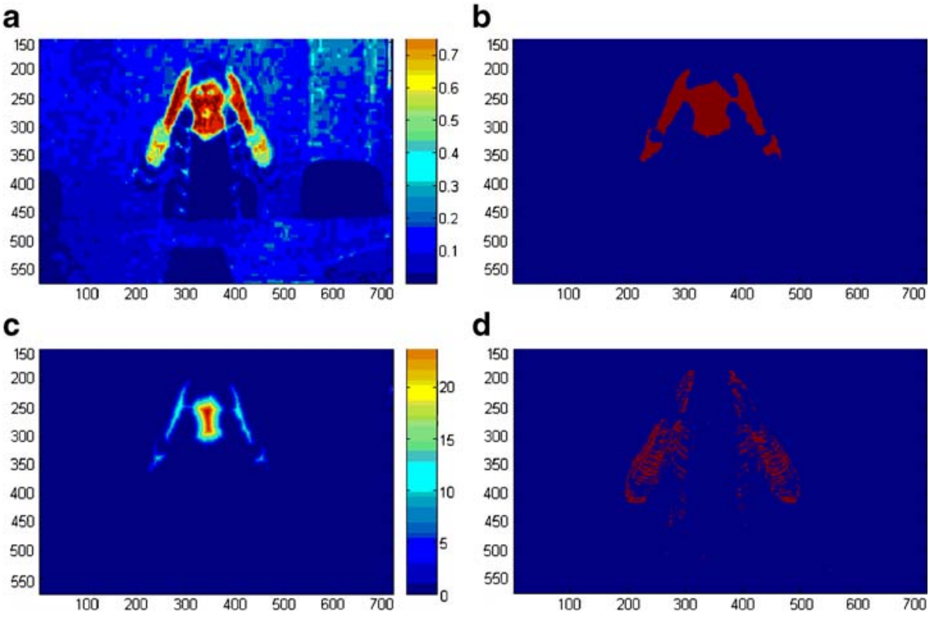


Fig. 7 Key steps in hand detection and tracking (**a**) Skin probability (**b**) Thresholding and morphology operators (**c**) Distance transformation (**d**) Frame Difference

scale and environment independent, this step is the ideal starting point for deploying an unconstrained technique, able to function under various lighting conditions, not considering any constraints regarding the distance of the user from the camera. Also, face detection is very important for further characteristics to be detected, such as facial features and hand positions, if skin color models are to be updated.

Numerous techniques have been developed for face detection in still images [51]. However, most of them can detect only upright faces in frontal or near-frontal views [23]. A method that can handle out-of-plane head motions is the statistical method for 3D object detection proposed by [37]. Other methods, e.g. [22], emphasize statistical learning techniques and use appearance features, including the real-time face detection scheme proposed by Viola and Jones [47], which is arguably the most commonly employed face detector in automatic facial expression analysis.

3.2 Classification of user states

A very important part of an educational system should be both recording and analysis of learner's state. The system becomes more efficient when it can be adjusted to every user. All learners do not concentrate for the same period of time or lose their interest when the rate of learning material is not very fast. In this section, we analyze the way that the learner state is captured, intending to use this information in order to enrich the learners' profile providing to the learner the proper learning material and in a proper learning rate.

The proposed system has expressive power to capture learner's behaviors and provide well modeled into six general learners' states (Frustrated/Struggling to read, Distracted, Tired/Sleepy, Not paying attention, Attentive, Full of interest). Analytically, the expressions of the face and the body of the learner are extracted using face and hands gestures analysis together with posture and motion. Our approach pulls out the position and shape of mouth, eyes, eyebrows and hands as well as features related to them (visual evidences, see Tables 1 and 2). These expressions which are captured, while the learner interacts with the computer, provide a set of behavioral evidences. Fusion of these behavioral evidences specifies the current learner's state.

Table 1 summarizes the initial mapping between positions of the detected feature points, distances between them and gaze and pose direction to the different states. Each of the state described, is also associated with some visual evidence, which is used to describe them. The visual evidences are relative to specific attributes, each of them has specific values appropriated to characterize a state (see Section 5: Experimental results).

4 Learners' states and e-Learning

One of the technical novelties introduced in the proposed e-learning system is the handling of its learners in a personalized manner, by building different profiles according to their behavior. In [45], an extension of the IEEE Reference Model (WG) of the Learning Technology Standards Committee (LTSA) has been presented (Fig. 8). In this approach, the user behavior is defined analyzing his (her) answers to an e-questionnaire. In order to enhance this approach, we have enriched the Evaluation module of the e-learning system with the aforementioned modules: the detection and tracking of facial features as well as the gaze and pose estimation and with the user states. The current section presents the design and implementation of a profile-based framework, which matches content to the context of interaction, so that the latter can be adapted to each learner's needs and capabilities. The

Table 2 Visual evidences

Visual evidence(s)	Possible values	Related feature or feature point(s)	Attributes
Severity	Mild/moderate/ severe	Eyebrows	Movement of eyebrows towards the eyes
Eyes looking at the screen	Yes/no	Gaze	Eye gaze vector length
Eyes wide open	Strong/above normal/ normal/below normal/ reduced	Eye area	Distance between points around the eye
Head is moving	Yes/no	Eye centers	Head pose vector length
Head is moving (direction)	None/forward/backward/ up/down/left/right	Eye centers	Inter-ocular distance increasing, head pose vector length
Head is moving (speed)	None/fast/normal/ slow	Eye centers	Head pose vector length first derivative
Frown	Yes/no	Eyebrows	Movement of eyebrows towards the eyes
Eyes blinking	Yes/no	Gaze	Direction and fluctuations of eye gaze vector
“Frozen” lips	Yes/no	Mouth area	Mouth feature points not moving
“Frozen” face	Yes/no	Facial area	Facial feature points not moving
Mouth open	Yes/no	Mouth area	Distance between uppermost and lowermost point on the mouth
Hand(s) covering eyes	Yes/no	Eye area–hands	Occluded by another skin patch
Hand covering mouth	Yes/no	Mouth area–hands	Occluded by another skin patch

overall system is able to continuously adapt to each learners states, preferences, usage history and specific needs as they evolve during its usage.

In this section, the architecture of the proposed e-learning system is described. Generally, the proposed e-learning system contains three types of entities: processes (oval),

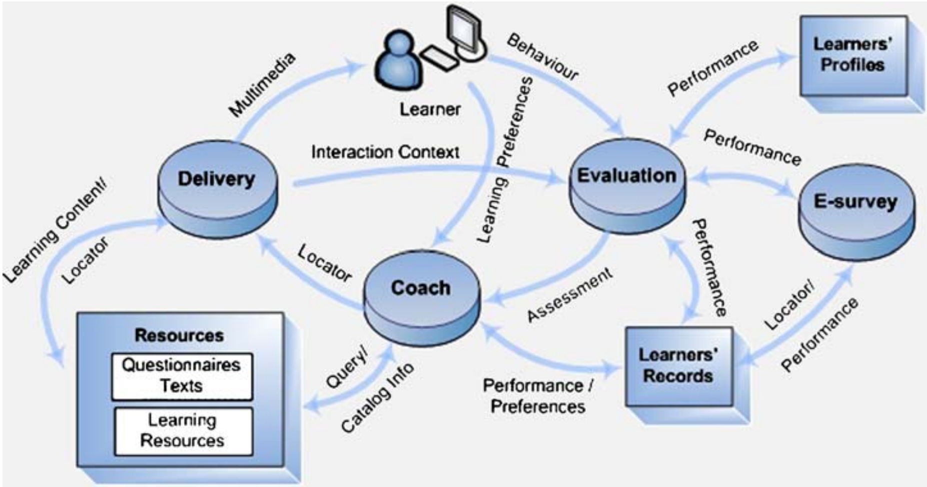


Fig. 8 Extension of LTSA IEEE learning system

stores (rectangles) and flows (vectors), each of those forming an active system component that transforms its inputs into outputs. Particularly, the first entity processes the information received from the stores entities via flows. The stores entities implement an inactive system component used as an information repository and the flows direct from one entity to another. In [45], the user behavior is defined analyzing his (her) answers to an e-questionnaire. In this paper, the behavior information includes the recording of user reaction by a web camera which is placed in front of him. The learner entity's observable behavior is given as input to the evaluation process which creates performance information flow stored in the learner records. The evaluation process has been enriched by the non-verbal feedback mechanism providing information about the user states. Performance information can come from both the evaluation process (e.g. video recordings, grades on lessons, learners' profile) and the coach process (e.g. certifications). The learner records store hold information about the past (e.g. historical learner records), the present (e.g. current assessments for suspending and resuming sessions, current behavior) and the future (e.g. presentation rate, pedagogy, learner).

In a general case, learner profiles are initially defined by the experts and are stored in the learners' profiles store. Each one of them describes characteristics of learners, learning needs and preferences. The information flow that the evaluation process provides, selects the existing learner profile from the learner profiles store that best matches the current learner's states. Thus, once a learner is assigned to a learner profile, the coach uses information in order to locate the learning material from the learning resources store that best match learner's profile.

In order to extract learner profiles, the evaluation process takes into consideration the current learner behavior and the learners' profile store. After that, the learner's attention is computed, classifying the learner's profile, which is then stored in the learner records. In addition, the learner profiles store contains the current profiles of the system. Thus, once a learner is assigned to a learner profile, the coach uses information in order to locate the learning material from the learning resources store that best matches learner's profile and displays it to him in an appropriate way according to his learning ability. Change of learners' profiles can be performed during their training, updating the learner records store.

The entity coach can receive performance information from the learner records at any time. Performance information, such as assessment information, certifications and preferences are stored in the learner records by the coach process. Based on this information, the coach process generates queries and forwards them to the learning resources store, in order to request learning materials that are appropriate for each learner. Finally the delivery process transforms these materials via learning content store into a presentation for the learner.

The learning resources store is a database that represents knowledge, information, and other resources used in the learning experiences. The learning resources store replies to the coach with catalog info, which may be used by the delivery process to retrieve learning contents from the corresponding store. Finally, the delivery process transforms information obtained via learning content store into a presentation, transferred to the learner entity via a multimedia flow. Finally, the entities: learning resources, learners' records and learners' profiles are stored in the database of the server.

In this framework, the proposed system can be considered as a personalized reading environment that is set according to the learner profile. In order to extract learner profiles, the evaluation process takes into consideration the current learner behavior and the learners' profile store. Firstly, the learner starts to read and his/her behavior is recorded using a web camera. The system, on-line, analyzes the input of the camera and computes the learner's attention (learner's states and states duration) classifying the learner's profile, which is then stored in the learner records.

In addition, the learner profiles store contains the current profiles of the system. Thus, once a learner is assigned to a learner profile, the coach uses information in order to locate the learning material from the learning resources store that best matches learner's profile and display it to him in a appropriate way (font size, font color, line spacing, stop/start highlighting) according to his learning ability. For example, if a learner is frustrated, the text highlighting rate in the presentation will be reduced and the font size will be enlarged. In case the learner is distracted or tired, a sound will be heard, trying to turn his/her attention back to the screen and the calibration of font formatting will be presented. If the state of the learner is attentive or full of interest the presentation format will not be changed.

Change of learners' profiles can be performed during their training, updating the learner records store. Moreover, the system allows an electronic survey to be conducted based on learners' states records. The main role of this component is the presentation of statistical analysis results, conducted upon the already stored learners' states records. A new learner profile can be created or existing ones can be adapted based on statistical analysis. New learner profiles or adapted versions of them are stored in the learner profiles store.

5 Experimental results

5.1 Recording and annotating the test data set

In order to train, test and evaluate the feature extraction and state detection capabilities of the proposed architecture, we set up a recording experiment in the framework of the Agent-Dysl project. A group of 20 students aged 8–10 (the target group of the project) participated in the experiment, which consisted of local special education experts selecting children with difficulties in decoding written text and presenting them with texts with varying selections of common pitfalls (e.g. single words or combinations of words often pronounced wrong, etc.). The selected students had participated in a nation-wide assessment of their ability to decode written text (or on an equivalent means of assessment) and scored among the lower quartile.

The experiment was set in inclusive classes within the mainstream primary schools in Greece. Each of the students attends the support program for no more than 8 h/week. During this time, students work together with their teacher in the fields of language or mathematics using the available equipment that this special class provides. Each inclusive class has one or two PCs (with equipment such as camera, microphone, and loudspeakers) for educational purposes. In these PCs, the proposed system was installed in order to be accessed by these students. These students are called to read a specific chapter of the electronic version of their course book. In our approach, we used 20 video segments in our experiments, between 800 and 1,200 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 100 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 protocol was followed.

Following clip selection, dyslexia experts came in to annotate the video clips, indicating segments which contained user states related to reading and dyslexia. This annotation was used in the following as a 'ground truth': distances and transformations of facial features for each state were used to train the neurofuzzy network, which was then used for testing purposes. However, the mapping between detectable features and user states was not documented from the experts in this process: for example, an expert annotated a particular video clip as ‘frustration’ but did not go into detail to also mention that the particular label was chosen because of the frown in the student’s eyebrows. For the vast majority of detectable states, this was a straightforward task for the design team, working mainly on past research on emotions and facial and hand gesture expressivity [23], [30]. However, some states were difficult to document since the features that participate in their expression were not always the same or could not be measured using the proposed processes (e.g. ‘frozen lips’, which *is* noticeable for an expert, but cannot be represented as a deformation of the detectable and tractable feature points). The following subsections elaborate on the actual features and states detectable by our approach, i.e. around the eyes, eyebrows, mouth and/or related to hand location and movement.

5.2 Detection of features—recognition of behavioral states

In order to estimate a user’s attention based on the children’s video files, a Sugeno-type fuzzy inference system [43] was built. Due to their good approximation and generalization capabilities, Sugeno-type systems have found numerous applications from simple neuro-fuzzy models [24] to multilayered classifiers [33]. The metrics used in our case were the head pose vector, the gaze vector and the inter-ocular distance.

The pose and gaze vector magnitudes were values between zero and one, while the inter-ocular distance was calculated as a fraction of the inter-ocular distance with respect to its value at the moment the algorithm initializes. Thus, the values of the inter-ocular distance between consecutive frames follow the distribution shown in Fig. 9.

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.

Prior to training, our data were clustered using the sub-cluster algorithm in [8]. 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

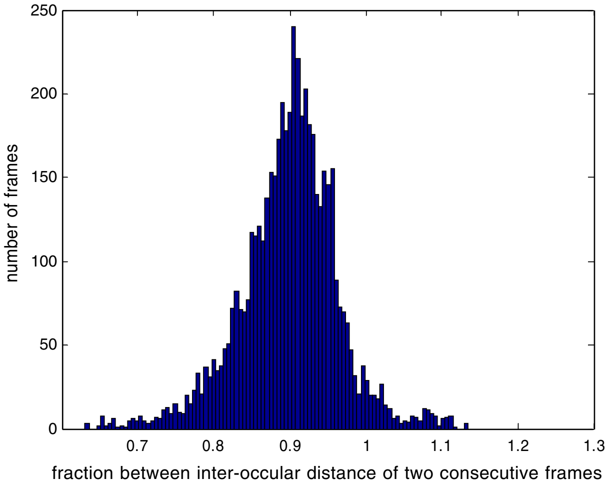


Fig. 9 Distribution of the fraction of the inter-ocular distance with respect to its value at the initialization of the algorithm

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 backpropagation gradient descent method [24].

After training, the output surface using the pose vector length and gaze vector length are shown in Fig. 10a, and the respective surface using the pose vector length and inter-ocular distance fractions as inputs is shown in Fig. 10b.

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 vector and pose vector lengths increase in magnitude, the output's value goes to zero. Similarly, in Fig. 10b it can be seen that for small values of the pose vector length, and for small changes of the inter-ocular distance, 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. Table 3 summarizes the results of our approach. In the

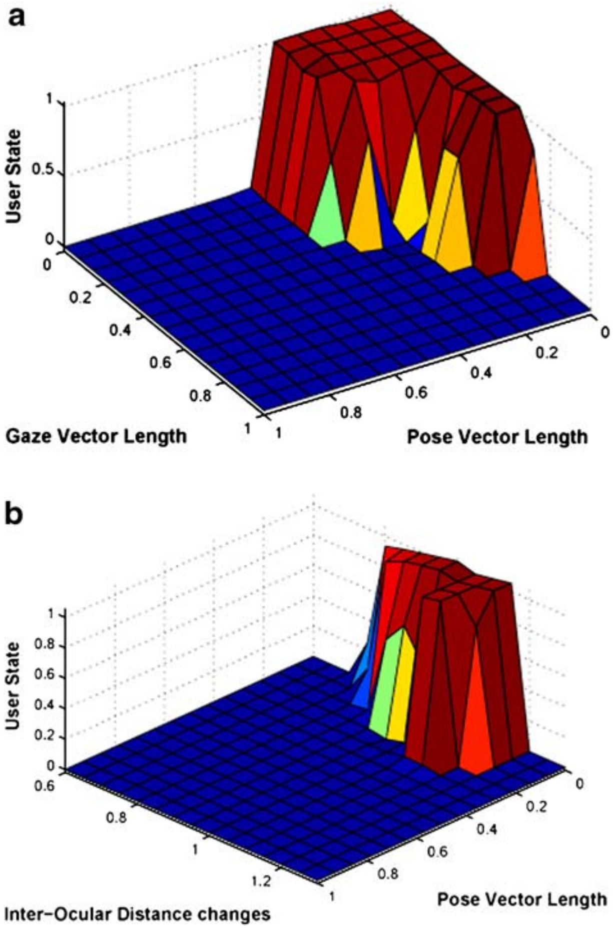


Fig. 10 **a** Output surface using the pose vector length and gaze vector length. **b** Output surface using the pose vector length and inter-ocular distance function

Table 3 Performance of the neuro-fuzzy classifier

State	Average error	% success
Attentive	0.04	100
Non-attentive	0.24	72
Total	0.117	87.7

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 (Table 3).

5.3 Relation of behavior states to the testbed application

The proposed architecture and reading user interface have been put to use in the context of the Agent-Dysl project, and their joint performance will be evaluated by both learners and teachers/instructors. The evaluation results will include findings of an in-depth survey of state of the art teaching practices regarding learners with reading difficulties (dyslexia), based on teaching and learning scenarios selected by expert teachers participating in the project. Moreover, the results will encompass information regarding the overall acceptability of the system performance; field tests will start in September in Greece, Denmark and Spain, coinciding with the start of the school season.

During the design process, measurable indicators describing an effort, performance and/or outcome of particular activities have been set for evaluating and making decisions regarding the typical read-only scenario. In this scenario, the learners specify the document to be loaded. In the following, they read the document based on their individual profile (which contains preferences for reading speed, font size, highlighting of successive words to be uttered in the correct order, etc.). While reading, this profile may be adjusted based on the actual performance. These indicators are:

- *Learners*' *reading pace and precision*

This indicator is based on comparison of pace and precision while the learner reads the same text in two environments (with or without Agent-Dysl software).

- *Learners*' *intrinsic motivation*

This indicator is based on comparison of the learner's motivation related to two reading scenarios (with or without Agent-Dysl software). The learner motivation is observed and coded while reading proper texts.

- *Learner's self-esteem*

This indicator is based on comparison of the learner's self-esteem related to self-directed learning tasks, i.e. learning efforts where the learner interprets some written learning material The learner motivation is observed and coded while reading proper texts in two scenarios.

As mentioned in the previous section, a group of students and their teachers were presented with the reading testbed in order to collect data for the training and testing of the user state detection process. During this initial feedback, teachers mentioned that the ability of a software to support reading, by identifying when the reader faces away from the screen is very valuable to the text highlighting process (so that it stops highlighting and resumes when the student starts facing the screen again), as well as to modeling the reading traits of a particular child. In the same context, differentiating lack of interest from a momentary lapse of concentration is also very useful so as not to make false assumptions on the reading performance; the proposed approach caters for this, by deploying concepts of fuzziness, making the decision system more robust to small changes and individual differences.

Teachers also pinpointed a number of complications associated with the internal parts of the architecture and with the small training set. Most of them had to do with the reduced ability of the neurofuzzy network to generalize state detection to a large number of children when trained with a small number of samples. This issue will be tackled by generating an individual profile for each set of students in the countries where the application will be deployed (Greece, Spain and Denmark); this profile will be updated with additional information as the application is used by the students and, consequently, training will be improved. Another issue was related to a selection of states not detectable by this testbed, such as ‘frozen lips’ mentioned before; however, the proposed approach and the developed testbed were at no point meant to replace the teacher in the assessment process, but to support in identifying *selected* issues related to user states and assist the students while reading.

6 Conclusions

An important possibility of non-verbal feedback concerning the interest of a person towards a web page, multimedia presentation, video clip or any other form of electronic document is the degree of engagement or interest towards the computer screen it is shown on. Head pose and movement, direction of gaze, as well as measurements of hand gesture expressivity are a vital part of this kind of feedback. We present a system used in the context of HCI to extract the degree of interest and engagement of students reading documents on a computer screen. This system will be used to correlate student performance and individual reading habits in the presence of dyslexia and to provide measurable feedback on their progress. The advantage of our system is that it is non-intrusive and real-time, thus supporting applications where spontaneity is of high importance and environments of various conditions regarding lighting and natural behavior. In addition to this, estimation of user states can be applied to a large number of applications, since it provides non-verbal feedback via a non-intrusive manner. In the framework of entertainment, which is the focus of the Callas IP project, user engagement is associated with the ‘pleasure’ and ‘arousal’ components of the PAD (Pleasure-Arousal-Dominance) emotion representation model; as a result, the application has a way of ‘knowing’ whether the viewer likes or is, at least, interested in a particular form of interaction and continues in the same manner or may choose to present different content in case boredom or lack of interest are detected. As a general rule, eye gaze can also be used as an indicator of selection, e.g. of a particular exhibit in a museum, or a dress when looking at a shop window, and may assist or even replace conventional mouse and keyboard interfaces in the presence of severe handicaps. All these possible applications benefit from the fact that the deployed sensors (in our case, a simple camera) are inexpensive and, more importantly, non-intrusive and that the user is already meant to look at a specific point (the screen, in the reading testbed, or the shop window, in the example mentioned earlier); as a result, the frontal view which is required to initialize the feature extraction process is easy to obtain and no manual or explicit initialization is required.

Acknowledgement This work has been funded by the FP6 IP Callas (Conveying Affectiveness in Leading-edge Living Adaptive Systems), Contract Number IST-34800 and the FP6 STREP Agent-Dysl (Accommodative Intelligent Educational Environments for Dyslexic learners) Contract Number IST-034549.

References

1. Asteriadis S, Nikolaidis N, Pitas I, Pardas M (2007) Detection of facial characteristics based on edge information, In: Proceedings of the Second International Conference on Computer Vision Theory and Applications (VISAPP), Barcelona, Spain, vol. 2, pp 247–252
2. Ba SO, Odobez JM (2006) A study on visual focus of attention recognition from head pose in a meeting room. In: Third Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI06), Washington, USA, pp 1–3
3. Baron-Cohen S (1995) Mindblindness. MIT, Cambridge
4. Beymer D, Flickner M (2003) Eye gaze tracking using an active stereo head. In: Proc. Of IEEE CVPR, Madison, WI, vol. 2, pp 451–458
5. Bosse T, Memon ZA, Treur J (2007) A two-level BDI-agent model for theory of mind and its use in social manipulation. In: Proceedings of the AISB 2007 Workshop on Mindful Environments, pp 335–342
6. Bouguet JY (2000) Pyramidal implementation of the Lucas Kanade tracker. OpenCV Documentation
7. Caridakis G, Karpouzis K, Kollias S (2008) User and context adaptive neural networks for emotion recognition. Neurocomputing 71:2553–2562 available online 9 May 2008
8. Chiu S (1994) Fuzzy model identification based on cluster estimation. J Intell Fuzzy Syst 2(3):267–278
9. Christie J, Johnsen E (1983) The role of play in social–intellectual development. R Educ Res 53(1):93–115
10. Commission of European Communities (2000) Communication from the Commission: e-learning—designing tomorrow's education. Commission of European Communities, Brussels
11. Cristinacce D, Cootes T, Scott I (2004) A multi-stage approach to facial feature detection. In: Proceedings of the 15th British Machine Vision Conference, London, UK, pp 277–286
12. D' Orazio T, Leo M, Cicirelli G, Distante A (2004) An algorithm for real time eye detection in face images. Pattern Recogn 3:278–281
13. D' Orazio T, Leo M, Guaragnella C, Distante A (2007) A visual approach for driver inattention detection. Pattern Recogn 40(8):2341–2355
14. Daugman JG (1993) High confidence visual recognition of persons by a test of statistical independence. IEEE Trans Pattern Anal Mach Intell 15:1148–1161
15. Deng JY, Lai F (1997) Region-based template deformation and masking for eye-feature extraction and description. Pattern Recogn 30(3):403–419
16. Duchowski AT (2002) A breadth-first survey of eye tracking applications. Behav Res Meth Instrum Comput 34(4):455–470
17. FP6 STREP (2007) Agent Dysl project. http://www.agent-dysl.eu. Accessed 10 August 2008
18. Gärdenfors P (2001) Slicing the theory of mind. In: Collin F (ed) Danish yearbook for philosophy. vol. 36. Museum Tusculanum Press, Copenhagen, pp 7–34
19. Gee AH, Cipolla R (1994) Non-intrusive gaze tracking for human–computer interaction. In: Proceedings of the International Conference on Mechatronics and Machine Vision in Practice Proceedings, Toowoomba, Australia, pp 112–117
20. Gourier N, Hall D, Crowley J (2004) Estimating face orientation using robust detection of salient facial features. In: Proceedings of Pointing, ICPR, International Workshop on Visual Observation of Deictic Gestures, Cambridge, UK
21. Hennessey C, Noureddin B, Lawrence P (2006) A single camera eye-gaze tracking system with free head motion. In: Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '05), San Diego, CA, USA, pp 87–94
22. Huang KS, Trivedi MM (2004) Robust real-time detection, tracking, and pose estimation of faces in video. In: Proceedings of the International Conference on Pattern Recognition (ICPR), Cambridge, UK, vol. 3, pp 965–968
23. Ioannou S, Caridakis G, Karpouzis K, Kollias S (2007) Robust feature detection for facial expression recognition. Int J Image Video Process 29081
24. Jang J-SR (1993) ANFIS: adaptive-network-based fuzzy inference systems. IEEE Trans Syst Man Cybernetics 23(3):665–685

25. Jesorsky O, Kirchberg KJ, Frischholz RW (2001) Robust face detection using the Hausdorff distance. In: Proceedings of the Third International Conference on Audio and Video-based Biometric Person Authentication (AVBPA), pp 90–95
26. Karagiannidis C, Sampson DG, Cardinali F (2002) An architecture for web-based e-learning promoting reusable adaptive educational e-content. Educ Technol Soc 5(4):27–37
27. Khan MM, Ward RD, Ingleby M (2006) Automated facial expression classification and affect interpretation using infrared measurement of facial skin temperature. ACM Trans Auton Adaptive Syst 1(1):1–113
28. Lillard A (1993) Pretend play skills and the child's theory of mind. Child Dev 64(2):348–371
29. Marsella SC, Pynadath DV, Read SJ (2004) PsychSim: agent-based modeling of social interaction and influence. In: Lovett M, et al. (eds) Proceedings of ICCM'04. Pittsburg, Pennsylvania, USA, pp 243–248
30. Martin J-C, Caridakis G, Devillers L, Karpouzis L, Abrilian S (2007) Manual annotation and automatic image processing of multimodal emotional behaviors: validating the annotation of TV interviews. Personal and ubiquitous computing (Special issue on Emerging Multimodal Interfaces). Springer, Heidelberg
31. Matumoto Y, Ogasawara T, Zelinsky A (2002) Behavior recognition based on head pose and gaze direction measurement. In: Proceedings of 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems, vol. 3, pp 2127–2132
32. Meyer A, Böhme M, Martinetz T, Barth E (2006) A single-camera remote eye tracker. In: Andre E (ed) Perception and interactive technologies (Lecture notes in artificial intelligence). vol. 4021. Springer, Heidelberg, pp 208–211
33. Mitrakis N, Theocharis J, Petridis V (2008) A multilayered neuro-fuzzy classifier with self-organizing properties, fuzzy sets and systems. doi:10.1016/j.fss.2008.01.032
34. Ong S, Ranganath S (2005) Automatic sign language analysis: a survey and the future beyond lexical meaning. IEEE Trans Pattern Anal Mach Intell 27(6):873–891
35. Otsuka K, Takemae Y, Yamato J, Murase H (2005) A probabilistic inference of multiparty-conversation structure based on Markov switching models of gaze patterns, head direction and utterance. In: Proceedings of International Conf. On Multi-modal and Interfaces, Trento
36. Pantic M, Patras I (2006) Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. IEEE Trans Syst Man Cybern B 36(2):433–449
37. Schneiderman H, Kanade T (2000) A statistical model for 3D object detection applied to faces and cars. IEEE Comput Soc Conf Vis Pattern Recogn 1:746–751
38. Seo K, Cohen I, You S, Neumann U (2004) Face pose estimation system by combining hybrid ICA-SVM learning and re-registration, In: Proceedings of the 5th Asian Conference on Computer Vision, Jeju, Korea
39. Smith P, Shah M, da Vitoria Lobo N (2003) Determining driver visual attention with one camera. IEEE Trans Intell Transportation Syst 4(4):205–218
40. Stiefelhagen R (2004) Estimating head pose with neural networks—results on the pointing. In: 04 ICPR Workshop Evaluation Data, Proceedings of Pointing, ICPR, International Workshop on Visual Observation of Deictic Gestures, Cambridge, UK
41. Stiefelhagen R, Yang J, Waibel A (2001) Estimating focus of attention based on gaze and sound. In: Proceedings of the Workshop on Perceptive User Interfaces, Orlando, Florida
42. Sylva K, Runer JS, Genova P (1976) The role of play in the problem-solving of children 3–5 years old. In: Bruner J, Jolly A, Sylva K (eds) PlayPIts role in development and evolution. Basic Books, New York
43. Takagi T, Sugeno M (1985) Fuzzy identification of systems and its applications to modelling and control. IEEE Trans Syst Man Cybern 15(1):116–132
44. Tzouveli P, Mitropoulou E, Ntalianis K, Kollias S, Symvonis A (2007) Design of an accommodative intelligent educational environments for dyslexic learners. In: Proceedings of the 11th Conference on Learning Difficulties in the Framework of School Education, Athens, Greece
45. Tzouveli P, Mylonas P, Kollias S (2008) An intelligent e-learning system based on learner profiling and learning resources adaptation. Comput Educ 51(1):224–238
46. Tzouveli P, Schmidt A, Schneider M, Symvonis A, Kollias S (2008) Adaptive reading assistance for the inclusion of students with dyslexia: the AGENT-DYSL approach. In: Proceedings of the 8th IEEE International Conference on Advanced Learning Technologies (ICALT 2008), Santander, Cantabria, Spain
47. Viola P, Jones M (2004) Robust real-time face detection. Comput Vis 57(2):137–154
48. Voit M, Nickel K, Stiefelhagen R (2005) Multi-view head pose estimation using neural networks. In: Proc of the Computer and Robot Vision (CRV'05), Victoria, BC, Canada,347–352
49. Ward RD (2004) An analysis of facial movement tracking in ordinary human-computer interaction. Interacting with Computers 16(5):879–896
50. Wu Y, Huang T (2001) Hand modeling, analysis, and recognition for vision-based human computer interaction. IEEE Signal Proc 18:51–60

51. Yang MH, Kriegman DJ, Ahuja N (2002) Detecting faces in images: a survey. IEEE Trans Pattern Anal Mach Intell 24(1):34–58
52. Yuxing M, Ching Y, Suen CS, Chunhua F (2007) Pose estimation based on two images from different views. In: Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV'07), Austin, Texas, USA, 9–16
53. Zhou ZH, Geng X (2004) Projection functions for eye detection. Pattern Recogn 37(5):1049–1056

Stylianos Asteriadis graduated from the School of Electrical and Computer Engineering of Aristotle University of Thessaloniki, Greece, in 2004. In 2006, he received his M.Sc. degree on Digital Media from the department of Computer Science of the same University. In December 2006, he joined the Image, Video and Multimedia Systems Laboratory of the School of Electrical and Computer Engineering of National Technical University of Athens, where he is pursuing his Ph.D. His research interests include Image and Video Analysis, Stereovision, Pattern analysis and Human–Computer interaction.

Paraskevi Tzouveli graduated from the School of Electrical and Computer Engineering of National Technical University of Athens in 2001 and she is currently pursuing her Ph.D. degree at the Image, Video, and Multimedia Systems Laboratory at the same University. Her current research interests lie in the areas of image and video analysis, information retrieval, knowledge manipulation and e-learning systems.

Dr Kostas Karpouzis graduated from the School of Electrical and Computer Engineering of the National Technical University of Athens in 1998 and received his Ph.D. degree in 2001 from the same University. His current research interests lie in the areas of human computer interaction, image and video processing, sign language synthesis and virtual reality. Dr. Karpouzis has published more than 70 papers in international journals and proceedings of international conferences. He is a member of the technical committee of the International Conference on Image Processing (ICIP) and a reviewer in many international journals. Dr. Karpouzis is an associate researcher at the Institute of Communication and Computer Systems (ICCS), a core researcher of the Humane FP6Network of Excellence and holds an adjunct lecturer position at the University of Piraeus, teaching Medical Informatics and Image Processing. He is also a national representative in IFIP Working Groups 12.5 'Artificial Intelligence Applications' and 3.2 'Informatics and ICT in Higher Education'.

Prof. Stefanos Kollias is a professor in School of Electrical and Computer Engineering of National and Technical University of Athens and the director of the Image, Video and Multimedia Laboratory. His research interest include image and video processing, analysis, coding, storage, retrieval, multimedia systems, computer graphics and virtual reality, artificial intelligence, neural networks, human computer interaction, medical imaging.