A Driver Assistance System based on Multilayer Iconic Classifiers: Model and Assessment on Adverse Conditions

G.L. Masala* and E. Grosso, *POLCOMING Department at University of Sassari*

Abstract— **Recent work demonstrates that iconic classifiers are good candidates for the development of effective driver assistance systems, exploiting on-board micro cameras and embedded architectures.**

Following this line of research, in this paper the combined use of multilayer classifiers and iconic data reduction, based on Sanger neural networks, is investigated. It is shown that by this affordable approach it is possible to capture the essential information of the images, making worthless much more structured and time-consuming feature-based techniques. In particular, the applicability of a simplified learning stage, based on a small dictionary of poses, is considered; this peculiarity makes the system almost independent from the actual user.

A detailed model of a simple driver assistance system, based on iconic classifiers, is presented and a comparative assessment, focused on the specific task of monitoring the car driver, is performed on adverse driving conditions. Three well known classification techniques are applied, demonstrating that the iconic approach, though can be certainly improved, is characterized by robustness, accuracy and real-time response; these features prove this technology to be an ideal tool for embedded automotive applications.

I. INTRODUCTION

Available studies identify distraction and fatigue of the driver as the primary causes of car crashes [1,2]. Since late 90s, this social problem attracted the interest of the scientific community which has begun to study the development of intelligent systems, suitable to monitor the driver's state of vigilance[1]. These systems are often denoted in the literature as Advanced Driver Assistance Systems or ADAS.

Computer vision techniques have been recently considered particularly promising to detect changes in the facial features which characterize behaviors of inattentive people. Typically, persons with reduced alertness due to fatigue show longer blink duration, slow eyelid movement, small degree of eye opening, frequent nodding, yawning and drooping posture [3]. In case of visual distraction, common situations include off-road gaze direction and persistent rotation of the head.

The design of a fully automated safety system based on computer vision can benefit of a number of robust tools, coming from basic image analysis and related fields like pattern recognition and biometrics [4]. In [5] a system based on a hybrid Bayesian Network uses eyes movements, spatial and temporal measures, and some driving performance measures on a simulator. However, the application on a moving vehicle presents new challenges like changing backgrounds and sudden variations of light conditions. Moreover, a useful system should guarantee real time performance and quick adaptability to a variable set of users and to natural movements performed during driving.

Considered in a broad sense, a fully automated system should simply provide an evaluation of the input images, giving as an output a quantized or binary level of attention. Obviously, the system could benefit of a priori knowledge concerning the environment or the user (the interiors of the cars, the identity of the user), some known patterns of attention/inattention (like, for instance, closed eyes or rotated head) and classes of attention (attentive, low attentive, completely inattentive driver). However, both the external noise and the characteristics of the input transducers can heavily affect the final result. For instance, in the case of a micro video camera difficulties may arise in relation to bandwidth, resolution, sensitivity, distortion, signal to noise ratio etc. Moreover, consider that he conceptual boundary between raw input data, feature extraction and proper classification can be somewhat arbitrary. The traditional goal of the feature extractor is to characterize raw data by measurements whose values are very similar for objects that are in the same category, and very different for objects that are not. An ideal feature extractor would therefore yield a representation that would make the job of the classifier trivial; conversely, an omnipotent classifier would not need the help of a sophisticated feature extractor. Starting from this framework, in this paper the use of multilayer iconic classifiers is investigated, proving that it can be very effective in the specific task of monitoring the driver attention. In particular, the model of a simple driver assistance system is detailed, including a preliminary face detector, an image data reduction step, and a final hierarchical classifier exploiting known image patterns.

The assessment of the system is performed in real driving conditions, considering both adverse lighting and freely moving users. In particular we report the results obtained with some supervised classifiers based on Neural Networks [6] like a Feed Forward Back Propagation Network and a Probabilistic Neural Network, in comparison with a deterministic K-Nearest Neighbours classifier [7].

Following sections are organized as follows: section 2 briefly introduces the proposed model, section 3 details the materials and methods. Experimental results are presented in section 4. Finally section 5 draws some conclusions and analyses possible outcomes of this research.

II. PROPOSED MODEL

The proposed attention model is based on a two-layer classifier whose final goal is to associate single frames with the attention state of the driver. For sake of simplicity only two states of attention ("attentive" or "inattentive") are considered.

* G.L. Masala is with Department of Political Science, Communication, Engineering and Information Technologies, Computer Vision Laboratory, Viale Mancini, 5, 07100, Sassari, ITALY (corresponding author to provide phone: +39(79)229486; fax: +39(79)229482; e-mail: gilmasala@uniss.it).

E. Grosso is with Department of Political Science, Communication, Engineering and Information Technologies, Computer Vision Laboratory, Viale Mancini, 5, 07100, Sassari, ITALY (e-mail: grosso@uniss.it).
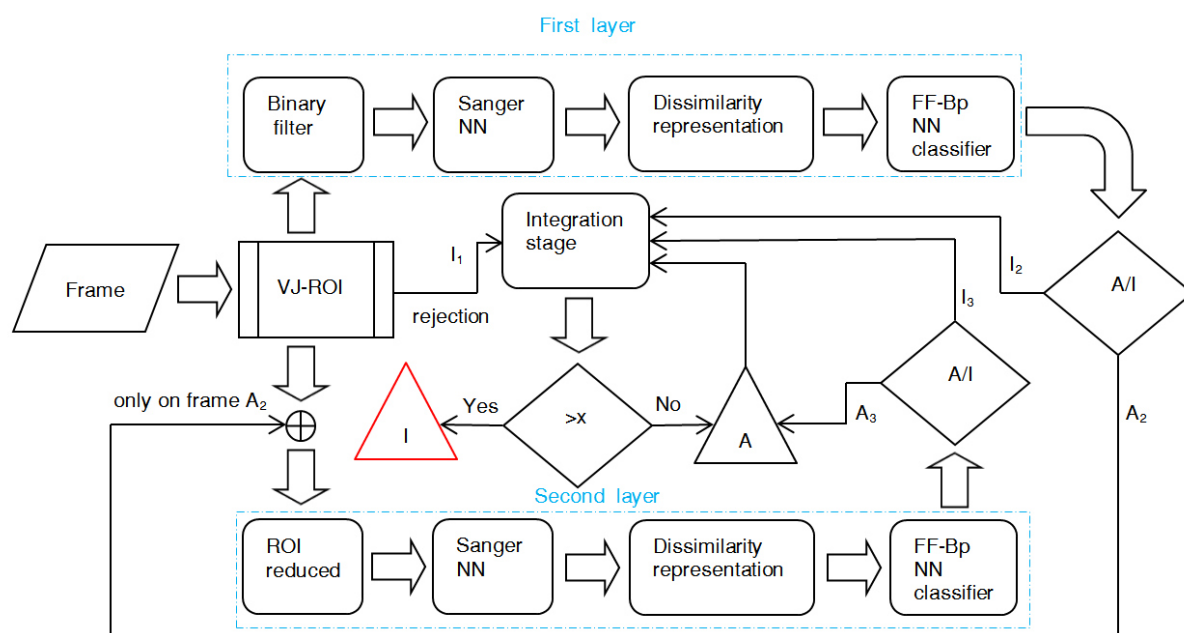
Fig.1 The input frame is first processed by the Viola-Jones algorithm; frames rejected from VJ are considered as inattentive I_1 while the resulting region of interest (ROI) is passed in a binary filter and coded through an Sanger neural network. A dissimilarity representation of the features using the distances from model dictionary is made. Finally, a trainable classifier decides about the attentive (A_2) or inattentive (I_2) poses of the driver. The second layer works only on frame A_2 ; the resulting ROI is reduced in a cell size around the eyes and coded through an Sanger neural network. A dissimilarity representation, using minimum distance from the model dictionary is made. Finally, a trainable classifier decides about the attentive (A_3) or inattentive (I_3) eyes status of the driver. An integration stage sum sequences of the inattentive states I and provides alarm only if we have a tunable sequence x of I.

Processing performed by the first layer is devoted to the detection of wrong head pose (then including drowsiness due to fatigue and visual distraction) while the second layer distinguishes between open/closed eyes, a measure strictly related to fatigue. A block diagram of the complete system is shown in figure 1.

Note that the a preliminary processing step is applied to each gray-scale frame in order to extract a small region of interest (ROI) containing face-candidates. To this purpose, the Viola Jones (VJ) face detector is used [8]. The Viola Jones detector is a state of the art powerful tool, which relies on a simple Haar-like image features representation, and off-line AdaBoost learning. The detector is very fast and produces fairly regular results on various light conditions; however it fails when the face of the driver is partly or totally out of the field view. It also fails in case of partial occlusion of the face and in case of manifest rotation of the head; all these cases conservatively bring to the immediate association of the frame to the "inattentive state".

Both the following layers work on extracted ROIs: these ROIs are first scaled to a fixed dimension (280x280 pixels), then are processed giving rise to the final classification. Note that the system knows about the origin of the classification; therefore it can distinguish between "inattentive" frames due to absence of face candidates (I_1), "inattentive" frames due to inappropriate head pose (I_2) and "inattentive" frames due to closed eyes (I_3). This information is used by the final integration block, deciding conveniently about the alarm state of the system.

III. MATERIALS AND METHODS

One of the key issues related to the proposed approach concerns the adoption of two Sanger neural networks (one for each layer) in order to reduce the dimensionality of the images corresponding to face candidates. A Sanger neural network is a simple three-layer feed-forward unsupervised network (with linear transfer function in the hidden neurons) which develops an internal representation corresponding to the principal components analysis of the full input data set. The input and output layers have the same dimension of the input patterns while the dimension of the hidden layer, corresponding to the number of the principal components, is determined during the training phase on model. Each network is trained as an auto-encoder [7,9-11], in such a way to reproduce at the output the input data. Starting from a typical number of principal components (12) used in eigen-faces detection [12] and using a small number of training frames (120 frames from the adopted dictionary of poses) we found the best configuration for 16 principal components. Only these values, representing the optimal reduction of the iconic data, are passed to the subsequent classifiers.

Note that the use of a dictionary of poses to train the Sanger networks has some interesting consequences. First of all each Sanger network is trained once; this means that processing can be executed off-line and without any reference to effective users. Secondly, once fixed the weights of the Sanger networks, data reduction can be easily obtained by projecting each ROI in the final feature space (i.e. by product of the Sanger weight vector for the row data frames). This operation is very fast, giving as a result a very

compact representation of the iconic image content both for the first and the second classification layer.

A. Representing dissimilarity

Representation based on dissimilarity is a well-known concept in the pattern recognition literature [13-15] and it is a very good alternative to the traditional feature-based description whenever relations between objects must be captured [13].

To construct a decision rule based on dissimilarity, a model reference set R with r elements is commonly used: R consists of prototypes which are representatives of all involved classes. In the learning process, a training set T of t elements is then adopted to build the $t \times r$ dissimilarity matrix $D(T,R)$ relating all training objects to all prototypes. The information on a set S of s new objects is provided in terms of their distances to R, i.e. as an $s \times r$ matrix $D(S,R)$.

In the above approach, a key factor is the discriminative power of the adopted measure of difference, but intrinsic properties of the adopted metric must be also considered. In fact, many traditional optimization methods are not appropriate when dissimilarities entail non-metric properties, then violating the triangle inequality axiom.

A final remark concerns the dimension of the feature space where measures are performed. In order to guarantee a good representation of the real data distribution, the number of samples must be much higher than the dimension n of the space; a reduction of the spatial dimensionality is therefore important to maintain a compact model reference set, and besides, to contain computational burden.

In the proposed approach the dissimilarity measure is performed by traditional Euclidean metric; distances are calculated on the Sanger components coding the input images, as shown in fig.1 The model reference set **R** is composed of 120 images (r=120) for the first layer and 68 images (r=68) for the second layer while the training set **T** is composed of several thousand of images, depending on the layer and on the considered subject. We denote the set **R** as "dictionary of poses" because the set is composed of images of a real user during the driving. Images are taken during three different sessions, with different conditions of light and slightly different distance from the camera. The same user appears with glasses and without glasses; different wrong poses of the head are also simulated by asking the user to look at eight fixed markers around the car. Open/close condition of the eyes is finally simulated asking the user to close the eyes both for correct and wrong poses of the head and simulating nodding. Some examples of dictionary of poses are shown in figure 2.
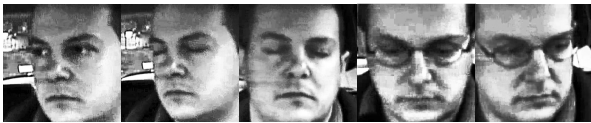

Fig. 2. Some frames from dictionary of poses

The adoption of a dictionary of poses it's very important in order to condensate into the model all the problems related to light and zoom. More in detail, the algorithm for the dissimilarity representation needs the computation of a measure of distance for each record i of the training set T with respect to each record k of the representation set R:

$$d_{ik} = \sqrt{\sum_m (t_m - r_m)^2} \qquad (1)$$

Note that

$$T_i = (t_1, t_2, .., t_{16}) \qquad i = 1,..,t \qquad (2)$$

$$R_k = (r_1, r_2, .. r_{16}) \qquad k = 1,..,r \qquad (3)$$

Each record of T and R is a vector with m=16 elements (Sanger components) while t defines the number of images of the set T and r the number of images of the set R (dictionary of poses).

A single row of the dissimilarity matrix will express all the distances of the generic training element t_i with respect to the reference set. In preliminary experiment we tried other notions of distances or data reduction of representative distances using mean, median and minimum, in according with [13,14], but the setting of dissimilarity presented in this paragraph provides best results.

As the final classes of the reference set are a-priori known, these distances can be obviously grouped in a number of subsets equal to the total number of classes and used to feed the training stage of the classifiers.

B. Classification

For the final classification step we compared three different types of techniques: a Feed Forward Back Propagation Neural Network (FF-Bp), a Bayesian approach available through Probabilistic Neural Networks (PNN) and a deterministic method as the K-Nearest Neighbours (K-NN).

All the classifiers are trained on a train set and tested on a validation set to determine the optimal configuration. The test set used for the experimental session is completely independent from the two previous sets.

Feed Forward Back Propagation Neural Networks [6,7] provide a not algorithmic, but very efficient, approach. Back propagation is used for learning: for a supervised system, the network is trained by using samples of known classes whilst for unsupervised systems the training is based on the minimization of a generic function of the data and the network's output.

Differently from FF-Bp, probabilistic Neural Networks [6] provides a general solution to pattern classification problems by following a Bayesian approach. The probabilistic neural network uses a supervised train set to develop distribution functions within a pattern layer. These

functions, in the recall mode, are used to estimate the likelihood of an input feature vector being part of a learned category, or class. The learned patterns can also be combined or weighted with the a priori probability, also called the relative frequency, of each category to determine the most likely class for a given input vector.

The last classifier compared is the well-known K-Nearest Neighbors[7]. For this type of deterministic classifier it is necessary to have a good train set, not too small, and a good discriminating distance. K-NN performs well in multi-class simultaneous problem solving. There exists an optimal choice for the value of the parameter K which brings to the best performance of the classifier. This value of K is often approximately close to $N^{1/2}$ where N is the number of the training samples.

C. Database

Even though several important databases are available for testing face and head pose recognition techniques (i.e IDIAP Head Pose Database [16], Feret [17] and others) video sequences of persons driving a car, captured by on-board cameras, are very few in number and hardly available. The acquisition of a small database has been therefore considered an essential requirement in order to validate the proposed approach.

The experimental setup has been conceived having in mind the need of collecting images during real driving; for this reason a wifi pinhole camera has been installed on the windshield of a car in a suitable position not to annoy the driver during a short trip. The camera allows the recording of several minutes of video during typical driving situations.

Data from two acquisition sessions of each driver, in different moments of the day and various conditions of ambient light, were collected. The users were driving both wearing glasses or not, without caring about the position of the seat and of the camera. Each session consists of 3 minutes of video recording, manually classified in attentive and inattentive states by poses and eyes [9] . For our experiment 5 representative people were chosen for a total of 10 sessions: a woman and four men (with glasses, without glasses, beard, smiling). Some ROIs extracted from the various sessions are shown in figure 3.

Fig. 3. Some frames of the persons used in our experiment; it is possible to see bad frames affected by interference or adverse lighting.

Note that the quality of the images is generally low, and that lighting and noise effects make really hard the classification task. In our perspective, however, these data well reflect the real operating environment of a driver assistance system.

All the images of the second session compose the test set (blind set) which is therefore used only to measure the performance of the system. In tables 1 and 2 datasets are shown.

TABLE 1 DATASETS FOR FIRST LAYER.

Persons	Sets pose detection in N° of Frames		
	Characteristics	*Train and validation*	*Test (blind)*
1	Woman	3238	3004
2	Smiling man	3750	2914
3	Man with glasses	2640	3229
4	Man	3127	3649
5	Man with beard	2752	2564

TABLE 2 DATASET FOR THE SECOND LAYER

Persons	Sets eyes detection in N° of Frames		
	Characteristics	*Train and validation*	*Test (blind)*
1	Woman	2779	2846
2	Smiling man	3213	2627
3	Man with glasses	2393	2894
4	Man	2758	3020
5	Man with beard	2399	2376

We used a Self Organizing Map (SOM) [19,23,24] to perform a random sampling over the first session of the available datasets. This SOM sorts out all samples into homogeneous groups from which we extracted a small amount of images and composed the training and validation sets.

IV. RESULTS

In a binary classification problem four possible outcomes must be considered: results on single frames classification are then given in term of true/false positive prediction and true/false negative prediction. In our model a positive condition corresponds to the presence of some inattentive status of the driver.

Table 3 shows the results for the classifiers on the blind test as average of the five representative people considered for the poses detection.

TABLE 3 CLASSIFIER FOR THE FIRST LAYER

Classifier	Results on Test set for poses detection (Mean of 5 persons)		
	Inattentive%	*Attentive %*	*ACC.%±dev.st*
FF-Bp	81.2	76.4	76±1.9
KNN	45.8	96.6	70±16
PNN	70.8	84.3	78±7.3

Table 4 shows the results for the classifiers on the blind test as average of the five representative people considered for the eyes detection.

TABLE 4 CLASSIFIER FOR THE SECOND LAYER

Classifier	Results on Test set for eyes detection (Mean of 5 persons)		
	Inattentive%	Attentive %	ACC.%±dev.st
FF-Bp	66.1	78.9	73±10.9
KNN	78.7	69.9	76±15.6
PNN	65.6	81.4	73±14.5

From tables 3 and 4 we can see how the FF-BP neural network provides the best results compared to other classifiers. In particular, the standard deviations are always lower than others presented, providing more stability of response with respect to varying people.

In tables 5 the detailed performances of the various layers of the system are presented for the FF-BP classifiers. It is worth noting that for the adopted setup configuration the VJ stage has a very high accuracy, reaching almost 100% of correct classification and rejecting 28.2% of the processed frames. This result is not surprising considering that the blind test sets include regular driving but also simulated inattentive behaviors. The last row of table 5 shows values related to the overall system; in this case accuracy, sensitivity (inattentive status) and specificity (attentive status) are computed just considering inattentive and attentive frames, without caring about the rejection stage.

TABLE 5 PERFORMANCES OF THE SYSTEM (FF-BP CLASSIFIERS)

Inattentive states	Results on single frame (Mean 5 persons)			
	Accuracy %	Sensitivity %	Specificity%	Frames %
VJ rejected (I_1)	98.8	99.5	98.7	28.2
Pose detection (I_2)	76.0	74.1	79.2	24.8
Eyes detection(I_3)	73.3	66.1	78.9	26.1
Mean weighted for total I respect to the number of frames	83.2	80.5	86.1	

V. DISCUSSION AND CONCLUSION

A thorough comparison of the proposed method with analogous results published in the literature is quite difficult due to real lack of common database protocols. Moreover most of the available results focus on the detection of specific temporal features like PERCLOS, blink and nodding frequency.

In [18], authors report comparable results for ten sequences records in real drive situations (10 users involved). The performance detecting inattentive states like nodding and wrong face pose is 72.5% and 87.5%, respectively, while fusing a complex set of different measures (nodding, face pose, gaze, eye closure duration and blinking frequency) the detection of the driver inattentiveness level reaches 97%. For PERCLOS, a performance around 90% is reported. In [19] tests on six video sequences, collected using a driving simulator, are presented. Accuracies in the classification of the PERCLOS range from 89.5 % to 98.2%, giving an average of 93.8% for the whole dataset. In [5] authors obtain the accuracy of 88% ±8 through a system based on a hybrid Bayesian Network which uses eyes movements, spatial and temporal measures, and some driving performance measures such as the standard deviation of steering wheel position, the mean steering error and the standard deviation of lane position. These results, again, refer to a simulator-based experiment.

The overall performance of the system proposed in this paper reaches an average accuracy of 83% for real sequences captured on-board and in uncontrolled/adverse conditions. This result states that the proposed technique, though extremely simple with respect to structured feature-based approaches, performs comparably well in different environmental situations. To this respect, it is worth noting that current results are strictly related to the classification of single frames, therefore admitting a significant level of improvement related to a more convenient use of additional information pertaining the temporal sequence of events and the specific type of inattentive states (the layer originating the actual rejection). These features will be certainly taken into account in the future design of appropriate alarm strategies.

Summarizing, the main contribution of this paper is the proposal of an novel method, based on binary iconic classifiers and achieving good levels of accuracy and real time performance, therefore particularly suitable for effective automotive applications. The paper explains how the adoption of complex cues or specific facial features can be efficiently replaced by adopting a generalized model of the inattentive drive, coming from a small dictionary of poses and totally independent from the actual user. With respect to previous work in the field [9] several major improvements can be noted: first of all the extension of the database to multiple sessions / multiple users and to real on-board sequences allowed a thorough validation of the approach; secondly the adoption of a dictionary of poses in order to train the Sanger network makes the image-reduction task totally independent from the actual user. The robust performances of a feed forward neural network is demonstrated with respect to other classifiers, using adverse lighting images. Moreover, the proposed method allows for a simple generalization of additional inattention states: yawning or drooping postures can be easily introduced by adding a limited number of new training samples in the dictionary.

Concerning weak points, it is worth noting that for both the classification stages an initial training of the system is yet required for each new user; this procedure requires less than one minute of training, which is an acceptable duration, but

also requires an active cooperation of the new user, who must simulate both attentive and inattentive states.

Current research is devoted to the simplification of this remaining training phase, deriving from the dictionary of poses a generic model of attention, totally independent from the single user, and devising a minimal "user adaptation" procedure, of about 5 seconds, during which the model is adjusted to the iconic appearance of the current user. As an alternative to this approach, also a "continuous user adaptation" could be considered, modulated by the actual speed of the vehicle.

First results in this sense are encouraging. In particular, it is now clear that an iconic generalization of attention states can be efficiently applied to a small population of users. However, the extension of this approach to very large sets of users requires further investigation.

REFERENCES

[1] A. Kircher, M. Uddman, J. Sandin, "Vehicle control and drowsiness" Technical Report VTI-922A, Swedish National Road and Transport Research Institute , 2002.

[2] The role of driver fatigue *in* commercial road transport crashes. European Transport Safety Council, Brussels 2001.

[3] L. M. Bergasa, J. Nuevo, M. A. Sotelo, R. Barea, and E. Lopez. Visual Monitoring of Driver Inattention. *In Comput. Intel. In Automotive Applications, SCI, pp 25-51, Springer,* 2008.

[4] W. Zhao, R. Chellappa, A. Rosenfeld, P.J. Phillips, Face Recognition: A Literature Survey, ACM Computing Surveys, pp. 399-458, 2003.

[5] Liang, Y., Lee, J.D. "A hybrid Bayesian Network approach to detect driver cognitive distraction" Transportation Research Part C: Emerging Technologies, 38, pp. 146-155, 2014.

[6] S. Haykin "Neural Networks – A comprehensive foundation", second edition, Prentice Hall, 1999.

[7] O. Duda, P. E. Hart, D. G. Stark, "Pattern Classification", second edition, A Wiley-Interscience Publication John Wiley & Sons, 2001.

[8] P. Viola and M. Jones , Robust "Real Time Object Detection", International Journal of Computer Vision, Volume 57, Issue 2 Pages: 137 – 154, 2004

[9] G. L. Masala, E. Grosso, Detecting Driver Inattention by Rough Iconic Classification*, Proceedings of IEEE Intelligent Vehicles Symposium (IV 2013), June 23-26,* 2013.

[10] G. L. Masala, U. Bottigli, A. Brunetti, M. Carpinelli, N. Diaz, P. L. Fiori, B. Golosio, P. Oliva, G. Stegel, Automatic cell colony counting by region-growing approach, Nuovo Cimento C, vol. 30(6), p. 633-644, 2007.

[11] T. D. Sanger , "Optimal Unsupervised Learning in a Single-Layer Linear Feedforward Neural Network", Neural Networks, vol. 2, pp. 459-473, 1989.

[12] M. Turk, A. Pentland, "Eigenfaces for recognition", Journal of Cognitive Neuroscience, 71-86, March 1991.

[13] Pekalska E., Duin R.P.W. "Classifiers for dissimilarity-based pattern recognition" 3nd International Conference on Pattern Recognition Barcelona, vol 2, Pattern Recognition and Neural Networks, pp.12-16, September 2000.

[14] S.W. Kim, R. Duin. "Dissimilarity-Based classifications in eigenspaces". In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer Berlin Heidelberg, p. 425-432, 2011.

[15] Bottigli, U., Golosio, B., Masala, G.L., Oliva, P., Stumbo, S., Cascio, D., Fauci, F., Magro, R., Raso, G., Bellotti, R., De Carlo, F., Tangaro, S., Mitri, D., De Nunzio, G., Quarta, M., Preite Martinez, A., Tata, A., Cerello, P., Cheran, S.C., Lopez Torres, E. Dissimilarity application for medical imaging classification WMSCI 2005 - The 9th World Multi-Conference on Systemics, Cybernetics and Informatics, Proceedings, 3, pp. 258-262, 2005.

[16] IDIAP Head Pose Database; url: http://idiap-head-pose-db.sspnet.eu/

[17] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi, "The FERET evaluation methodology for face recognition algorithms", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 10, October 2000.

[18] R. Senaratne, B. Jap, S. Lal, A. Hsu, S. Halgamuge, and P. Fischer. Comparing two video-based techniques for driver fatigue detection: classification versus optical flow approach. Mach. Vision Appl. 22, 4. July 2011.

[19] H. Rowley, S. Baluja, and T. Kanade: Neural Network-Based Face Detection. *IEEE Tr. On PAMI,* January 1998.