Transportation Research Part C 49 (2014) 32–42

Contents lists available at ScienceDirect

Transportation Research Part C

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

Real time detection of driver attention: Emerging solutions based on robust iconic classifiers and dictionary of poses

G.L. Masala [*], E. Grosso

POLCOMING Department at University of Sassari, Italy

ARTICLE INFO

Article history:
Received 26 May 2014
Received in revised form 12 September 2014
Accepted 13 October 2014
Available online 5 November 2014

Keywords:
Automotive applications
Monitoring of driver attention
Driver assistance systems
Neural networks

ABSTRACT

Real time monitoring of driver attention by computer vision techniques is a key issue in the development of advanced driver assistance systems. While past work mostly focused on structured feature-based approaches, characterized by high computational requirements, emerging technologies based on iconic classifiers recently proved to be good candidates for the implementation of accurate and real-time solutions, characterized by simplicity and automatic fast training stages.

In this work the combined use of binary classifiers and iconic data reduction, based on Sanger neural networks, is proposed, detailing critical aspects related to the application of this approach to the specific problem of driving assistance. In particular it is investigated the possibility of a simplified learning stage, based on a small dictionary of poses, that makes the system almost independent from the actual user.

On-board experiments demonstrate the effectiveness of the approach, even in case of noise and adverse light conditions. Moreover the system proved unexpected robustness to various categories of users, including people with beard and eyeglasses. Temporal integration of classification results, together with a partial distinction among visual distraction and fatigue effects, make the proposed technology an excellent candidate for the exploration of adaptive and user-centered applications in the automotive field.

1. Introduction

Since late 90s, undesirable or unusual driver conditions have been clearly identified as a primary cause of car crashes and road deaths (Kircher et al., 2002; European Transport Safety Council, 2001). This problem attracted the interest of the scientific community, which has begun to study the development of intelligent and adaptive systems, namely Advanced Driver Assistance Systems (ADAS), suitable to monitor the diver's state of vigilance and give real-time support in accident avoidance (Liang et al., 2007; Batista, 2005).

As pointed out by Liang and Lee (2014), the nature of driver inattention can vary: fatigue and related symptoms like drowsiness and frequent nodding are very common in real cases but distraction from safe driving can also have a visual or cognitive cause.

Visual distraction has often to do with the on-board presence of electronic devices or tools like mobile phones, navigation and multimedia systems, requiring active control from the driver (for example pushing buttons or turning knobs); visual distraction can be also related to the presence of salient visual information away from the road, thus causing spontaneous

[*] Corresponding author at: POLCOMING: Department of Political Science, Communication, Engineering and Information Technologies, Computer Vision Laboratory, Viale Mancini, 5, 07100 Sassari, Italy. Tel.: +39 (79) 229486; fax: +39 (79) 229482.

E-mail address: gilmasala@uniss.it (G.L. Masala).

http://dx.doi.org/10.1016/j.trc.2014.10.005
0968-090X/© 2014 The Authors. Published by Elsevier Ltd.

off-road eye glances and momentary rotation of the head. Cognitive distraction happens whenever the mind of the driver is not sufficiently focused on the critical task of safe driving; symptoms of cognitive distraction are less apparent, and difficult to be detected or quantified by objective indicators. Most of times the analysis of cognitive distraction is therefore based on long behavioral patterns and sophisticated statistical techniques (Liang and Lee, 2014).

Focusing on fatigue and visual distraction, the paper investigates the design and the development of a fully automated driver assistance system based on advanced techniques coming from image analysis and related fields like pattern recognition and biometrics (Zhao et al., 2003).

In previous studies, computer vision techniques have been often proposed to detect driver attention (Batista, 2005; Singh and Papanikolopoulos, 1999) both by standard and day-night infrared cameras. In particular, these techniques have been adopted to detect signs of visual distraction, like off-road gaze direction and persistent rotation of the head, and changes in the facial features which characterize persons with reduced alertness due to fatigue: longer blink duration, slow eyelid movement, small degree of eye opening, nodding, yawns and drooping posture are among the most interesting conditions which has proved to be captured by vision-based approaches (Bergasa et al., 2008).

A common processing scheme, well discussed in Senaratne et al. (2011, 2007) includes the following steps:

- face localization;
- localization of facial features (e.g. eyes or mouth);
- estimation of specific cues related to fatigue or distraction;
- fusion of cues in order to determine the global attention level.

Concerning face localization, very robust techniques based on neural networks have been developed in late 90s (Rowley et al., 1998; Sung and Poggio, 1998). In 2004 Viola and Jones (Viola and Jones, 2004) proposed a new high performance algorithm based on integral images and robust classification; this algorithm is a de facto standard for real-time applications. Both the above approaches belong to the image-based subclass of the face detection techniques. More recently also feature-based approaches demonstrated a reasonable level of efficiency. In particular, Particle Swarm Optimization (Kennedy and Eberhart, 1995) has been proposed for locating and tracking a limited number of facial landmarks.

Research on facial features extraction mainly focused on eyes and mouth (Zhao et al., 2003); Gabor and SVM techniques have been successfully proposed to this aim (Senaratne et al., 2007). In order to work under low light conditions, researchers also proposed the use of infrared illuminators, exploiting high reflection of the pupils (Senaratne et al., 2011); as noted in Lenskiy and Lee (2012), however, IR based approaches show malfunctions during daytime and require the installation of additional hardware.

It is worth noting here that most of the literature defines the PERCLOS as the main cue for the estimation of driver's fatigue. PERCLOS is a measure of the time percentage during which eyes remain closed 80% or more; in order to compute this cue, every image frame is usually classified into two classes (closed eyes or open eyes): k-NN techniques, SVMs and Bayes approaches have been successfully applied to this purpose (Rowley et al., 1998). Other cues commonly used are head pose, eye blinking detection (Lenskiy and Lee, 2012), slouching frequency and postural adjustment. To the aim of this work the estimation of the head pose certainly represents the most interesting issue (Sung and Poggio, 1998); this information can be derived by applying both 2D and 3D approaches (Murphy-Chutorian and Trivedi, 2009).

Overall, previous studies show that the problem of detecting visual distraction and fatigue can be faced with fairly good results in driving simulators or constrained conditions. However, the application on a real moving vehicle presents new challenges like changing backgrounds and sudden variations of lighting. 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.

In order to tackle the real problem and to reach a sufficient level of accuracy and performance, we propose here a driver assistance system based on robust iconic classifiers. Starting from a preliminary image data reduction step, and from a priori knowledge related to known head poses and known patterns (like, for instance, closed/open eyes), we show that iconic classifiers perform well with respect to changes in pose and facial features configuration, while ignoring unessential details like glasses, hairstyle and lighting conditions. As explained later in the text, the conceptual boundary between raw input data, feature extraction and classification can be somewhat arbitrary; moreover the proper classification of the input data can be heavily influenced by the collection of poses and patterns used in the learning phase. For this reason we propose a binary classification of poses and features, where the collection of possible configurations is simply categorized in "attentive" versus "inattentive" classes.

Following sections are organized as follows: Section 2 briefly introduces the adopted attention model and the fundamental methods applied for the various processing steps; Section 3 details the experimental setup, the data collection phase and experimental results. Finally Section 4 draws some conclusions and analyses possible outcomes of this research.

2. Approach and methods

Even though the adoption and the fusion of different cues usually shows some increase of performance, recent work (Masala and Grosso, 2013) demonstrates that this approach can be efficiently replaced by an alternative "fully iconic" approach, based on a generalized model of the "inattentive driver". This iconic generalization, derived by processing and classifying off-line a sequence or a selected set of images of a generic real user, is denoted here as "dictionary of poses" because it captures essential iconic information related to the position of the head and the state of the eyes of a driver both during attentive and distracted or fatigued driving. As in the Viola Jones face detector (Viola and Jones, 2004), this pre-learned pattern that can be usefully exploited for on-line processing, achieving high levels of accuracy and real time performance.

Note that for the Driver Assistance Systems the distinction among visual, cognitive and fatigue effects is not unessential; having knowledge about the origin of the distraction can help the system to implement adaptive and more intelligent behaviors. For this reason, while we define in the following as "inattentive driver" a subject showing visual distraction, fatigue effects or both (which is totally coherent with the final goal to detect the diver's state of vigilance), we also try to maintain some level of information about the type of processing which generates the inattentive classification.

2.1. Outline of the model

The proposed attention model is based on a two-layer classifier where the single frames are processed and associated to the "attentive" or "inattentive" state of the driver. The first layer is devoted to the detection of the 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 Fig. 1.

Note that the Viola Jones face detector (Viola and Jones, 2004) is preliminary applied to each frame in order to extract a small region of interest (ROI) containing face-candidates. The Viola Jones detector relies on a large set of simple Haar-like features, and uses the AdaBoost learning algorithm to reduce this over-complete set . The detector is applied to gray-scale images, producing fairly regular results; 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". (See Figs. 2 and 3)

Both the following layers work on extracted ROIs: these ROIs are first scaled to a fixed dimension (280 × 280 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 temporal integration block, deciding conveniently about the alarm state of the system.

2.2. Correct/wrong pose detection

The first classification layer is specialized on the detection of a wrong head pose in a single frame. The input is the ROI extracted from the Viola Jones face detector. ROI are first processed by histogram equalization, then a binarization filter is applied. In order to reduce the dimensionality of the image, a Sanger neural network is used. Then a dissimilarity representation is computed taking as reference a small dictionary of poses. The final classifier is a feed forward neural network (FF-Bp) which processes the dissimilarity representation and decides about the attention state of the driver due to head movement. If the head of the driver has the correct pose (A) the original ROI is passed to the secondary layer, otherwise it is labeled as inattentive I_2.

2.3. Open/closed eyes detection

The second layer of classification is specialized in detecting the state of the eyes in a single frame; only ROIs labeled as A from the first layer are considered. In this case, first a small image rectangle centered on the eyes region (220 × 120 pixels) is

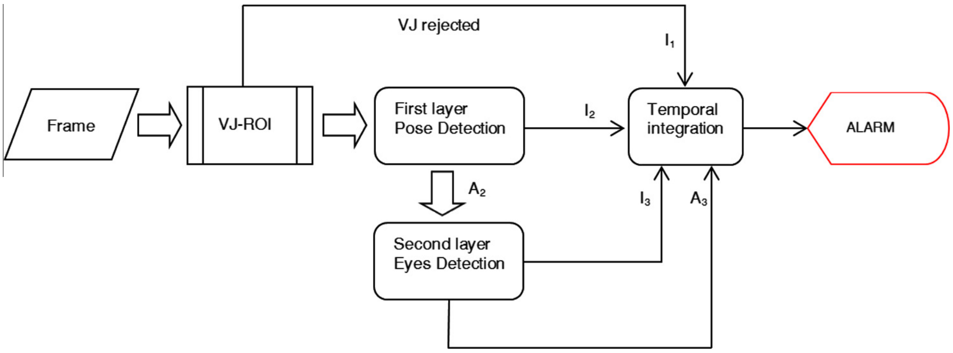

Fig. 1. In the block diagram the input frame is first processed by the Viola-Jones algorithm; frames rejected from VJ are considered as inattentive I_1 while accepted regions of interest (ROI) are passed to the first layer which decides about the attentive (A_2) or inattentive (I_2) poses of the driver. The second layer works only on frames A_2 deciding about the attentive (A_3) or inattentive (I_3) eyes state of the driver. The temporal integration stage considers sequences of the inattentive states I and provides a final alarm signal.

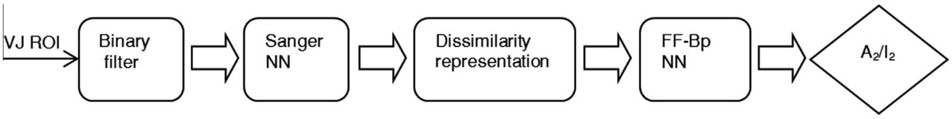

Fig. 2. In the block diagram the input (VJ ROI) is the region of interest extracted by the Viola-Jones algorithm; the ROI is first binarized and then coded through an Sanger neural network into a vector of 16 components. A dissimilarity representation based on the Sanger components is finally computed and passed to a trainable classifier which decides about the attentive (A_2) or inattentive (I_2) state of the driver.

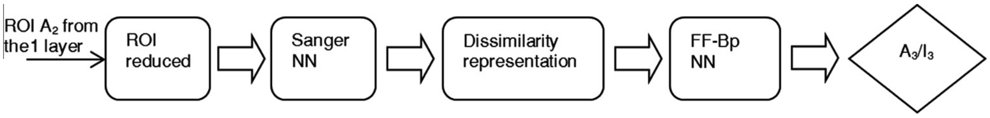

Fig. 3. In the block diagram, the original VJ ROI, labeled as A from the first layer, a small fixed area around the eyes is extracted and coded through an Sanger neural network into a vector of 16 components. A dissimilarity representation based on the Sanger components is finally computed and passed to a trainable classifier which decides about the attentive (A_3) or inattentive (I_3) state of the driver.

extracted and processed by histogram equalization, then a Sanger neural network similar to the previous stage is used to reduce the dimensionality. Also in this layer a dissimilarity representation is used in order to improve the subsequent classification step. Only the frames of the previous dictionary where both eyes are clearly recognizable are used as reference. The final classifier is again a feed forward neural network (FF-Bp) which processes the dissimilarity representation and decides about the attention state of the driver derived from the condition of the eyes. If the driver has normally open eyes the frame is associated to an attentive state (A_3); otherwise it is labeled as inattentive state (I_3).

2.4. Algorithm details

2.4.1. Reducing the dimensionality

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 (Sanger, 1989) . 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. Each network is trained as an auto-encoder (Duda et al., 2001; Masala et al., 2007), 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 (Turk and Pentland, 1991) and using a small number of training frames (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.

2.4.2. Representing dissimilarity

Representation based on dissimilarity is a well-known concept in the pattern recognition literature (Pekalska and Duin, 2000; Bottigli et al., 2005; Kim and Duin, 2011) and it is a very good alternative to the traditional feature-based description whenever relations between objects must be captured (Pekalska and Duin, 2000). A dissimilarity value expresses the difference between two objects or features and becomes zero only when the two objects are identical. In general, dissimilarity measures are applied directly to raw data (for instance images or temporal signals) but it is not rare the use of pre-processing steps aimed at reducing the dimension of the feature space. A very powerful pre-processing method, well investigated by authors in Kim and Duin (2011) is based on principal component analysis. In particular, it has been shown that computing dissimilarity on principal eigenvectors helps to face intractable problems like distortion, illumination changes and noise.

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

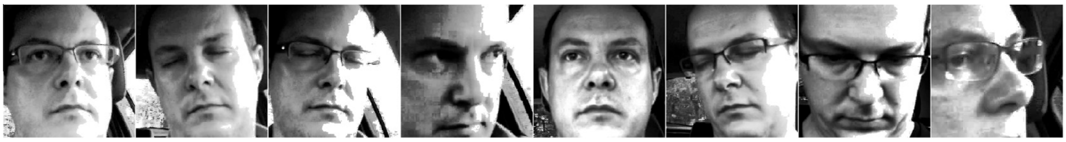

Fig. 4. Some examples of poses extracted from the dictionary.

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 for non-metric dissimilarities, as they often rely on 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. The model reference set **R** is composed of 72 images ($r = 72$) for the first layer and 48 images ($r = 48$) 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 example of the images of the dictionary are given in Fig. 4.

The dissimilarity representation is computed over the Sanger components; the dimension n is therefore equal to 16.

In summary, we have:

$$\boldsymbol{t}_i = (t_{i1}, t_{i2}, \dots, t_{i16}) \quad \mathrm{i} = 1, \dots, \mathrm{t} \tag{1}$$

$$\boldsymbol{r}_k = (r_{k1}, r_{k2}, \dots r_{k16}) \quad \mathrm{k} = 1, \dots, \mathrm{r} \tag{2}$$

where t_i and r_k are generic frames of the training and reference set. The generic element of the dissimilarity matrix will be:

$$d_{\mathrm{ik}} = \|\boldsymbol{t}_{\mathrm{i}} - \boldsymbol{r}_k\| \tag{3}$$

A single row of the dissimilarity matrix will express all the distances of the generic training element $\boldsymbol{t}_{\mathrm{i}}$ with respect to the reference set. 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.

2.4.3. Classifiers

For the classification step we used a Feed Forward Back Propagation Neural Network (FF-Bp) (Duda et al., 2001; Haykin, 1999). FF-Bp provides 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. In our case, the classifiers are trained on a training set and tested on a validation set to determine the optimal parameterization. As detailed in the next section, the total number of images used in the training/validation stages depend on the subject but it is always significant, ranging from about 1800 to nearly 2300 images per session.

Concerning the configuration of the classifiers, the following have been used:

1. First layer: a FF-Bp with 72 input neurons and 2 output neurons; note that the input neurons correspond to the dimensionality of the dissimilarities representation of data, while output neurons correspond to the attention states considered in this layer (0-correct, 1-wrong pose).
2. Second layer: a FF-Bp with 48 input neurons and 2 output neurons; in this case the input neurons correspond to a subset of the dictionary of poses where the eyes are clearly recognizable, while output neurons correspond to the attention states considered in this layer (0-open, 1-closed eyes).

2.4.4. Temporal integration

As detailed in the experimental section, the result of a binary classifier for a given condition can be easily defined in terms of true (correct) and false (wrong) rate of detected items.

Denoting by p and $(1 - p)$ the probability of correct/wrong detection of a generic Bernoulli trial (representing the classification of a single frame for which the ground truth state is "inattentive"), the related binomial distribution $B(n,p)$ defines the discrete probability of a number k of correct detections in a sequence of n independent trials. More precisely, if the X random variable follows the binomial distribution, we can write this probability as:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

The cumulative distribution of a random variable X following a binomial distribution is defined in turn as:

$$P(X \leq k) = \sum_{i=1}^{k} \binom{n}{i} p^i (i - p)^{n-i}$$

The meaning of the above expression has to do with the probability of having a number of correct detections less than or equal to k in a sequence of n independent trials where p denotes the probability of a correct detection for a single trial and provided that the truth state remains "inattentive".

It is well known, for the law of large numbers, that taking n sufficiently big and $k = n/2$, $P(X \leqslant k)$ tends to one if $p < 0.5$; conversely, $P(X \leqslant k)$ tends to zero if $p > 0.5$. Therefore $P(X \leqslant k)$ is a robust estimate of the observed condition when independent measures of the same condition can be performed (and obviously the condition does not change during the measurement process).

The temporal integration scheme, commonly used in recognition methods, exploits these theoretical considerations extending the output of the classifiers over a span of n consecutive frames; a majority voting is usually adopted in order to decide about the final classification.

The temporal span must obviously respect the dynamic of the observed event; in particular the duration of the observation window:

- cannot exceed the typical duration of the event (an event lasting typically t_1 ms cannot be correctly detected with a larger temporal window because the measures would refer to different conditions);
- must be greater than the minimum duration for which the event is considered significant (if the event becomes significant after t_2 ms the temporal window must be larger in order to avoid false alarms).

In the proposed approach, we adopted identical temporal windows, 400 ms long, for the first and second layer. Using a frame rate of 10 Hz this corresponds to the integration of 5 frames and a majority voting of 3 over 5 consecutive frames.

3. Experiments

The acquisition of a small database has been considered an essential requirement in order to validate the proposed approach. In fact, even though several important databases are available for testing face and head pose recognition techniques (i.e. IDIAP Head Pose Database (IDIAP Head Pose Database), Feret (Phillips et al., 2000) and others) video sequences of persons driving a car, captured by on-board cameras, are very few in number and hardly available.

The experimental setup has been conceived having in mind the need of collecting images during effective driving; for this reason a USB camera has been installed on the windshield of a car in a position convenient and compatible with a smooth ride. The camera allows the recording of several minutes of video during typical driving situations.

For each driver, data from two acquisition sessions, in different moments of the day and various lighting conditions, 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 min of video recording, manually classified as follows:

- about one minute of normal driver behavior: the driver looks at the road straightaway or to rear view mirrors;
- about one minute of simulated fatigue effects: the driver closes the eyes and simulates nodding;
- about one minute of distracted behaviors; the driver looks up, down or laterally focusing on eight fixed markers around the car.

Currently, the database is composed of 15 registered users driving the same car, for a total of 30 sessions and about 90 min of video recording. The database includes subjects of different gender, subjects wearing glasses, subjects with beard; common expressions due to smiling and talking are also included. Some ROIs extracted from the various sessions are shown in Fig. 5. 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.

In Fig. 6 we show how our database can be decomposed by gender and by additional characteristics that can potentially condition the accuracy of the system.

It is well known in the pattern recognition community that a crucial step in the experimental phase concerns the identification of three different sets of data (training set, validation set and test set). In fact, a good random distribution of the samples in these data sets guarantees a correct measure of the system performance, compensating for possible biases. We used a Self Organizing Map (SOM) (Duda et al., 2001; Haykin, 1999; Masala and Grosso, 2014) 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. All the images of the second session compose the test set (or blind set) which is therefore used only to measure the performance of the system.

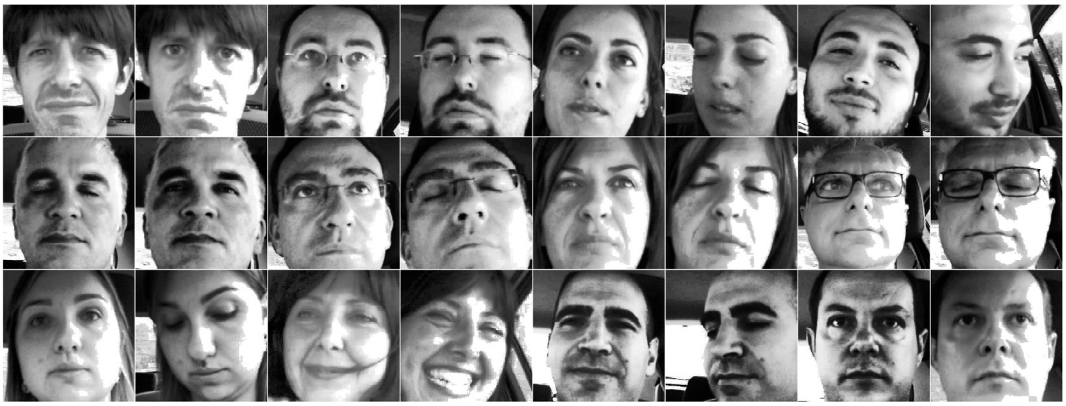

Fig. 5. Samples of the pictures captured during the acquisition sessions, after the Viola Jones ROI extraction; subjects show different poses and different degrees of attention. Note that for each person two sample frames have been selected, one for each acquisition session; differences in lighting and saturation effects of the camera are clearly visible.

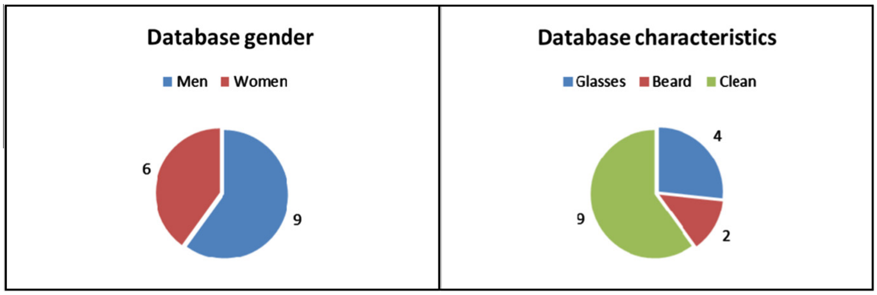

Fig. 6. Database decomposition by gender and by some additional characteristics; in particular we denote with "clean" a person without beard and not wearing glasses. Only one person has both beard and glasses.

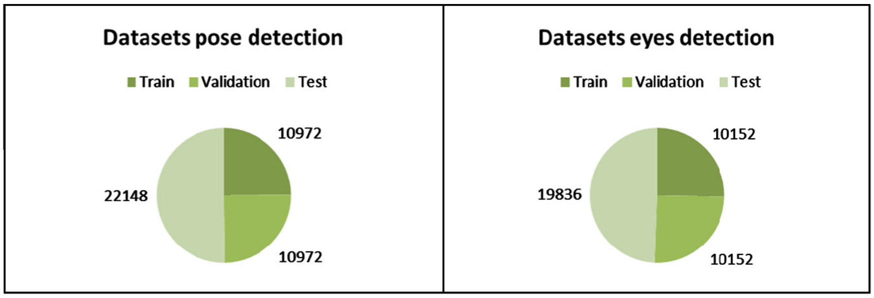

Fig. 7. Consistency of the datasets, after the VJ stage, for the two classification layers. Training + Validation and Test (blind) sets are composed by frames belonging to two independent sessions.

Fig. 7 shows the overall distribution of the resulting data sets used in the experimental phase. In the training, validation and test of the second layer (eyes detection), only A_2 frames are considered; images related to inattentive head pose of the driver are therefore removed in the counts.

3.1. 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; more in detail we denote as:

- True Positives (TP) – the number of the outcomes related to a positive prediction (inattentive driver detected) when the actual condition is also positive (inattentive driver);
- False Positives (FP) – the number of the outcomes related to a positive prediction (inattentive driver detected) when the actual condition is negative (attentive driver);
- True Negatives (TN) – the number of the outcomes related to a negative prediction (attentive driver detected) when the actual condition is also negative (attentive driver);
- False Negatives (FN) – the number of the outcomes related to a negative prediction (attentive driver detected) when the actual condition is positive (inattentive driver).

The above values usually compose the confusion matrix, a 2 × 2 table which relates each actual condition with the test outcome. Table 1 shows the results for the blind test of the first classification layer. For sake of clarity the values displayed refer to the average of the 15 subjects considered; moreover values are expressed in percentage terms with respect to the total number of images P and N belonging to positive (inattentive) and negative (attentive) sets, respectively.

As detailed by the global accuracy level, the overall performance of the classifier is satisfactory. In particular, good levels of TP/P and TN/N denote a good discriminative power for both conditions, even though at single frame level.

Table 2 shows the results for the second classification layer. Note that also in this case the global accuracy is good, with a good balance between TP/P and TN/N values.

Classification results can be further analyzed with respect to sub-classes represented in Figs. 7–9 show the decomposition of the experimental results both by gender and additional characteristics; average values of accuracy, sensitivity and specificity are displayed for each sub-class, together with the extension of the range of values.

Note that results are not significantly dependent on the considered sub-classes; the only remarkable variations concerns the presence of beard or glasses that do not affect pose but cause imbalance on eyes detection between sensitivity and specificity values.

Table 3 summarizes accuracy, sensitivity (TP/P) and specificity (TN/N) values at different stages of the model for the whole dataset. 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 31% 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 3 shows values related to the overall system; in this case accuracy, sensitivity and specificity are computed just averaging inattentive and attentive frames and without caring about the rejection stage.

Improvements related to the use of temporal integration are shown by Table 4 for the whole dataset. As detailed in Section 2.4.4 the temporal integration module works on a 400 ms window, applying majority voting over 5 frames. The expected improvement of accuracy, sensitivity and specificity is correctly detected in the experimental data.

3.2. Analysis and discussion

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

Table 1
Classification results for the first layer (pose detection) on single frames.

	Confusion matrix for head pose detection (mean 15 persons) Detection		
Condition	Inattention %	Attention %	Accuracy %
Inattention	83.7 (TP/P)	16.3 (FN/P)	
Attention	6.7 (FP/N)	93.3 (TN/N)	
			92.0 ± 9.5

Table 2
Classification results for the second layer (eyes detection) on single frames.

	Confusion matrix for eyes detection (mean 15 persons) Detection		
Condition	Inattention %	Attention %	Accuracy %
Inattention	88.0 (TP/P)	12.0 (FN/P)	
Attention	27.2 (FP/N)	72.8 (TN/N)	
			81.0 ± 9.0

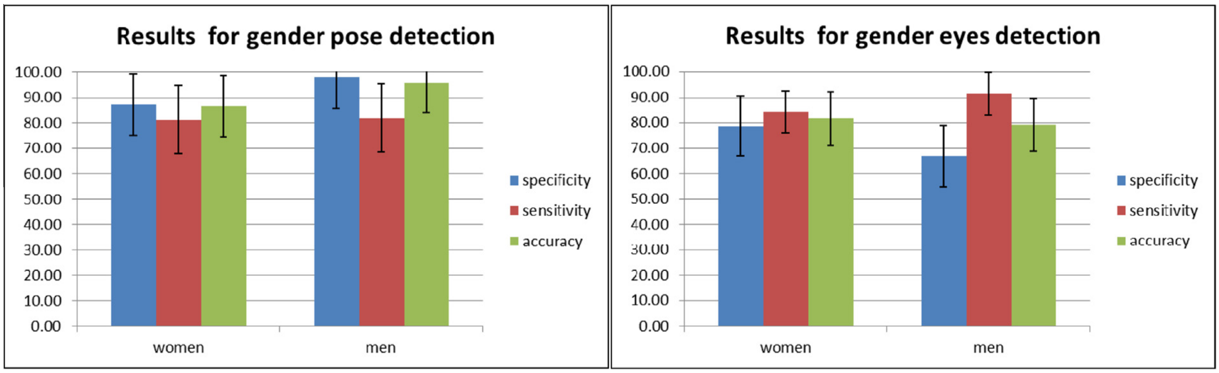

Fig. 8. Decomposition by gender of the classification results for the first layer (pose detection, left) and the second layer (eyes detection, right).

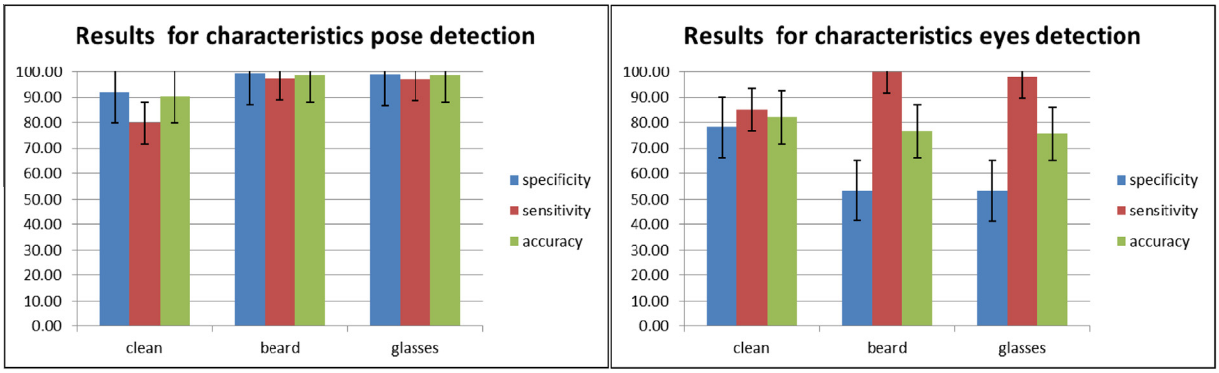

Fig. 9. Decomposition by additional characteristics of the classification results for the first layer (pose detection, left) and the second layer (eyes detection, right).

Table 3
Summary of the performance of the system for each rejection stage. The last row refers to the overall system performance.

Inattentive states (Mean 15 persons)	Results on single frame			
	Accuracy **%**	Sensitivity %	Specificity **%**	Frames I %
VJ rejected (I_1)	96.5	97.8	93.9	31.0
Pose detection (I_2)	92.0	83.7	93.3	38.7
Eyes detection (I_3)	81.0	88.0	72.8	30.3
Mean weighted for total I respect to the number of frames	90.1	89.4	87.3	100.0

Table 4
Summary of the performance of the system for each rejection stage when considering simple temporal integration. The last row refers to the overall system performance.

Inattentive states (Mean 15 persons)	Results on temporal sequence of 5 frames			
	Accuracy %	Sensitivity %	Specificity **%**	Frames I %
VJ rejected (I_1)	97.1	98.4	94.6	31.0
Pose detection (I_2)	95.7	83.3	98.6	38.7
Eyes detection (I_3)	84.3	87.6	81.3	30.3
Mean weighted for total I respect to the number of frames	92.7	89.3	92.1	100.0

The work of Bergasa and colleagues (Bergasa et al., 2008) is, to our knowledge, the only work using long sequences recorded during real driving. Authors describe a quite complex feature-based approach and report results for ten sequences records (10 participants involved): the performance detecting inattentive states like nodding and wrong face pose is 72.5%

Table 5
Comparison of the literature results concerning the detection of inattentive states of a driver.

Paper	Comparative performances of the systems			
	Types	Sequences	Measures	Accuracy %
Paper (Bergasa et al., 2008) Fuzzy rules	Real drive	10	Nodding, face pose, gaze, eye closure duration and blinking frequency	97
Paper (Senaratne et al., July 2011) MoC + Gabor features	Real drive	1	PERCLOS	92
Paper (Rowley et al., 1998) Neural Networks	Driving simulator	6	PERCLOS	94
Paper (Liang and Lee, 2014) Layered algorithm	Driving simulator	9	Eyes movements, spatial and temporal measures, and some driving performance measures	88
Paper (Liang and Lee, 2014) DBNs	Driving simulator	9	Eyes movements, spatial and temporal measures, and some driving performance measures	88
Paper (Liang and Lee, 2014) SVMs	Driving simulator	9	Eyes movements, spatial and temporal measures, and some driving performance measures	90
Paper (Masala and Grosso, 2014) Neural Networks	Real drive	5	Face pose, eyes closure	83
Actual proposed method	Real drive	15	Face pose, eyes closure, temporal integration	93

and 87.5%, respectively, while fusing a large 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 93.1% is reported. Senaratne et al. (2011) also report partial results on real driving, claiming a PERCLOS accuracy around 92%. In Rowley et al. (1998) 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 Liang and Lee (2014) 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 from lane position. These results, again, refer to a simulator-based experiment. Interestingly, in this paper a comparison among different classifiers is also performed, demonstrating the superiority of non-probabilistic linear classifiers like SVMs with respect to Bayesian Networks. A similar comparison among classifiers has been proposed by Masala et al. in a preliminary work (Masala and Grosso, 2014) where Feed Forward Back Propagation Neural Network (FF-Bp), Bayesian Probabilistic Neural Networks (PNN) and deterministic K-Nearest Neighbors (K-NN) are considered; from this comparison FF-BP neural networks seem to provide optimal stability and good robustness with respect to varying people. Table 5 gives a more clear overview of the above remarks.

Coming to a more detailed analysis of the experimental results obtained for the proposed approach, it is worth noting that, even without considering temporal integration (Table 3), classification results are very good for pose detection and quite good for eyes detection. Intra-subject and intra-group variations are limited and overall acceptable. Clearly the detection of small features like the eyes is affected by the presence of eyeglasses and beard, which is perfectly explicable in relation to the “iconic” content of the image regions considered. The overall performance of the system proposed in this paper reaches an average accuracy of 92.7% (Table 4) for real sequences captured on-board and in uncontrolled situations. This result states that the proposed technique, though extremely simple with respect to structured feature-based approaches, performs comparably well in different environmental conditions. To this respect, note that current results are strictly related to a simple majority voting scheme, therefore admitting a significant level of improvement related to a more convenient use of additional information pertaining the specific type of inattentive states. This feature will be certainly taken into account in the future design of appropriate alarm strategies.

4. Conclusions

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 (Masala and Grosso, 2013) 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. Moreover, the proposed method allows for a simple generalization of additional inattention states: yaws 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 s, during which the model is adjusted to the iconic appearance of the current user. The approach would also admit an easy extension to the biometric field, serving as a face recognition based security system for the vehicle. In fact the same adaptation procedure could be used to analyze and store peculiar biometric features of the actual user.

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

Batista, J.P., 2005. A Real-Time Driver Visual Attention Monitoring System. In Lecture Notes in Computer Science: Pattern Recognition Image Analysis. Springer, vol. 3522, pp. 200–208.

Bergasa, L.M., Nuevo, J., Sotelo, M.A., Barea, R., Lopez, E., 2008. Visual monitoring of driver inattention. In: Comput. Intel. in Automotive Applications, SCI. Springer, US, pp 25–51.

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., 2005. Dissimilarity Application for Medical Imaging Classification WMSCI 2005 – The 9th World Multi-Conference on Systemics, Cybernetics and Informatics, Proceedings, 3, pp. 258–262.

Duda, O., Hart, P.E., Stark, D.G., 2001. Pattern Classification, second ed. A Wiley-Interscience Publication, John Wiley & Sons.

European Transport Safety Council, 2001. The Role of Driver Fatigue in Commercial Road Transport Crashes. European Transport Safety Council, Brussels.

Haykin, S., 1999. Neural Networks – A comprehensive foundation, second ed. Prentice Hall.

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

Kennedy, J., Eberhart, R., 1995. Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks. IV. pp. 1942–1948.

Kim, S.W., Duin, R., 2011. Dissimilarity-based classifications in Eigenspaces. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer, Berlin, Heidelberg, pp. 425–432.

Kircher, A., Uddman, M., Sandin, J., Vehicle Control and Drowsiness Technical Report VTI-922A, Swedish National Road and Transport Research Institute, 2002.

Lenskiy, A.A., Lee, J., 2012 (Driver's eye blinking detection using novel color and texture segmentation algorithms). In: Int. J. Control Autom. Syst. 10 (2).

Liang, Y., Lee, J.D., 2014. A hybrid Bayesian Network approach to detect driver cognitive distraction. Transport. Res. Part C: Emerg. Technol. 38, 146–155.

Liang, Yulan., Reyes, Michelle L., Lee, John D., 2007. Real-time detection of driver cognitive distraction using support vector machines. IEEE Transact. Intell. Transport. Syst. 8 (2), 340–350.

Masala, G.L., Grosso, E., 2013. Detecting driver inattention by rough iconic classification. In: Proceedings of IEEE Conference on Intelligent Vehicles Symposium, Gold Coast, Australia, June 23–26, 2013.

Masala, G., Grosso, E., 2014. A Driver Assistance System based on Multilayer Iconic Classifiers: Model and Assessment on Adverse Conditions. In: Proceedings of the ITSC 2014, IEEE Conference on Intelligent Transportation Systems, Qingdao, China, October 8–11, 2014.

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

Murphy-Chutorian, Erik., Trivedi, Mohan M., 2009. Head pose estimation in computer vision: a survey. IEEE Trans. Pattern Anal. Mach. Intell. 31 (4), 607–626.

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

Phillips, P.J., Moon, H., Rauss, P.J., Rizvi, S., 2000. The FERET evaluation methodology for face recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. 22 (10).

Rowley, H., Baluja, S., Kanade, T., 1998. Neural network-based face detection. IEEE Trans. PAMI.

Sanger, T.D., 1989. Optimal unsupervised learning in a single-layer linear feedforward neural network. Neural Netw. 2, 459–473.

Senaratne, R., Hardy, D., Vanderaa, B., Halgamuge, S., 2007 (Driver fatigue detection by fusing multiple cues). In: Proceedings of the 4th International Symposium on Neural Networks: Part II-Advances in Neural Networks. Springer-Verlag, Berlin, Heidelberg.

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

Singh, S., Papanikolopoulos, N.P., 1999. Monitoring driver fatigue using facial analysis techniques. In: Proc. Of the IEEE Int. Conf. On Intelligence Transportation Systems, pp. 314–318.

Sung, K., Poggio, T., 1998. Example-based learning for view-based face detection. IEEE Trans. PAMI, January 1998.

Turk, M., Pentland, A., 1991. Eigenfaces for recognition. J. Cogn. Neurosci., 71–86.

Viola, P., Jones, M., 2004. Robust "real time object detection". Int. J. Comput. Vis. 57 (2), 137–154.

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