Facial Recognition System for Driver Vigilance Monitoring*

H.J. Dikkers BSc
Faculty of Electrical Engineering,
Mathematics and Computer Science
Delft University of Technology
Delft, The Netherlands
harmen@ch.tudelft.nl

M.A. Spaans BSc
Faculty of Electrical Engineering,
Mathematics and Computer Science
Delft University of Technology
Delft, The Netherlands
mike@ch.tudelft.nl

Co-authors
D. Datcu BSc – Delft University of Technology
prof. dr. M. Novák - Czech Technical University of Prague
drs. dr. L.J.M. Rothkrantz – Delft University of Technology

Abstract - *Many automobile accidents are related to drivers lacking required levels of vigilance to properly control their vehicles. In this paper we present a system that monitors the activity of parts of the face, in particular the eyes, in order to predict expressions of somnolence. The input to the system is a sequence of images of the face of a car driver, captured by a video camera. The system makes an assessment based on the movement and position of the eyes and eyelids. The system is tested in a car simulation environment. The results will be presented.*

Keywords: Facial recognition, micro-sleeps prediction, facial expression classification.

1 Introduction

A great many people are injured or killed each year due to accidents involving automobiles. Statistics indicate that over 40% [2] off all traffic accidents are related to lack of vigilance on behalf of the driver. At several institutions around the world, modern research on car safety has shifted focus to understand the phenomena of somnolence in full. At the Faculty of Transportation Sciences at the Czech Technical University in Prague (CTU) researchers use electroencephalography (EEG) to attempt to understand the activity in the brain at the onset of sleep. Brain wave patterns are considered a primary factor in the determination (and prediction) of sleep as brain wave patterns will change instantly at the onset of sleeps, as compared to secondary factors such as cardiac rhythm or respiratory activity, which change more slowly once sleep sets in. In this paper we present a proof-of-concept (or feasibility study) of a system that monitors the activity of the face, in particular the eyes, and can determine/predict expressions of somnolence. Our research complements the research done at CTU's Faculty of Transportation Sciences in such a way that data obtained from our system can be combined with EEG and other data to provide a complete map of human physiology at the onset of sleep.

2 Related Literature

The scientific field of research of facial expression classification is closely related to the problem of somnolence detection in facial images. Typical classification systems are capable of recognizing around six different facial expressions from visual data (happiness, anger, etc.). Determining expressions of somnolence can be considered a sub-problem of this, since the data involved need only be classified into two distinct expressions: sleepy—notsleepy. All facial expression recognition models can be placed within a three-tier framework, as described by Pantic and Rothkrantz [6]. This framework consists of the following steps:

- Face detection
- Facial expression data extraction
- Facial expression classification

In the rest of this section these steps will be described in more detail, in order to extend the reader's background knowledge of the relevant subject matter.

2.1 Face Detection

We distinguish two different models to represent the face: the holistic model, in which the face is represented as a whole, and the analytical model, in which the face is represented as a set of facial features. A holistic approach has been proposed by Huang and Huang [3], who use something called a Canny edge detector to roughly

* **0-7803-8566-7/04/$20.00 © 2004 IEEE.**

estimate the location of the face in the image. This detector observes the valley in pixel intensity that lies between the lips and the two symmetrical vertical edges representing the outer vertical boundaries of the face. This system has limitations with regard to rigid head movements, facial hair and glasses, and has several illumination constraints. Pantic and Rothkrantz [7] use dual view facial images in their holistic approach to facial expression classification. They analyze the horizontal and vertical histograms of the frontal view image in order to determine the boundaries of a rectangle around the face. To determine the contour of the face, they use an algorithm based on the HSV color model. No facial hair or glasses are allowed in their system and they require a camera to be mounted on the subject's head. Kobayashi and Hara [5] use an analytical approach: they use a CCD camera in monochrome mode to obtain brightness distribution data of the face. Their system determines the position of the irises in real-time, by comparing the brightness distribution of the currently examined data to an average distribution obtained by data averaged over ten subjects. The subjects face the camera at a distance of approximately one meter; no rigid head rotations are allowed. Kimura and Yachida [4] propose using a potential net for face representation. First, they normalize an image by using the centers of the eyes and the center of mouth, which are found by an integral projection method based on color and edge information. Then, the potential net is fitted to the normalized image to model the face and its movement. Analyzed faces are without facial hair and glasses and parallel to the direction of the camera; no rigid head rotations are allowed.

2.2 Facial Expression Data Extraction

Eye closure and narrowing eyelids are the most obvious signs of the onset of somnolence. In our research we have decided to focus on the eyes in determining expressions of somnolence. We are assuming other facial features are less relevant to the classification problem, though they can provide extra information. At the onset of sleep the position of corners of the mouth, for instance, are likely to be slightly lower than their normal position. As we have distinguished two approaches in face detection systems, we can also distinguish two different approaches in facial expression data extraction: the holistic and the analytical approach. Additionally, it is possible to combine the two approaches in a hybrid approach. Using the analytical approach and selecting appropriate (i.e. eye-based) features is the most obvious solution to our particular problem. In the holistic and hybrid approaches many non-relevant features are taken into account which make results less reliable. In their analytical system Kobayashi and Hara [5] use a geometric face model of 30 Facial Characteristic Points (FCPs), 16 of which concern the eyes. A set of brightness distributions of 13 vertical lines crossing these FCPs is used on a normalized image. The data thus obtained is fed to a neural network as input. No facial hair and no glasses are allowed in their system. It is able to operate in real-time. Cohn [1] uses a model of facial landmark points near the facial features. In the first frame of a sequence of recorded images, the landmark points are selected manually. For the other frames an optical flow method is used. The displacement of each landmark point is calculated by subtracting its normalized position in the first frame from its current normalized position. All frames of an input sequence are normalized manually. The displacement vectors, calculated between the initial and the peak frame, represent the facial information used for recognition of the displayed facial actions. Analyzed faces are without facial hair and glasses, no rigid head motions are allowed and the face in the first frame must be neutral (or expressionless).

2.3 Facial Expression Classification

With regard to the facial expression classification problem three methods can be distinguished:

- Template-based methods
- Neural-network-based methods
- Rule-based methods

Template-based methods compare an arbitrary image to prototypic templates in each expression category, and classify this image to the category with the best match. In general, it is difficult to achieve quantified template-based recognition of non-prototypic images, which means that images can only be exclusively categorized into one class without a level of uncertainty. The fact that each person has a unique maximal intensity of displaying a certain facial action makes this even more difficult. It is possible to consider neural networks as template based methods due to their black-box behavior. Pantic et al. [6] however distinguishes neural networks from template-based methods, as they perform quantified facial expression categorization. When performing a neural-network-based classification, a facial expression is classified according to the categorization process that the network learned during a training phase. Recognition of non-prototypic facial expressions is feasible if each neural network output is associated with a weight from the interval [0, 1], instead of being associated with either 0 or 1. Kobayashi and Hara [5] apply a 234 * 50 * 6 back-propagation neural network. The units of the input layer correspond to the brightness distribution data extracted from an input facial image (see previous section). Each unit of the output layer corresponds to an emotion category. The rule-based systems surveyed by Pantic classify the examined facial expressions into the basic emotion categories, based on previously encoded facial actions (a method to describe the face). In order to achieve this, prototypic expressions are first described in terms of facial actions, after which an

examined expression can be compared to the prototypic expressions defined for each of the emotion categories and classified in the optimal fitting category. Pantic and Rothkrantz [7] use the localized contours of the face to extract model features. The difference between the currently extracted model features and the same features extracted from an expressionless face of the same person is calculated and compared to prior acquired knowledge. This generates a set of production rules, which can thus classify the images into the appropriate classes.

3 HADES

In this section we present the HADES system (Hybrid Approach to Determining Expressions of Somnolence), a system designed to analyze a stream of images taken by a video camera of a subject's face, and detect expressions of somnolence. As was described in the previous chapter, the problem of emotional expression classification consists of three basic steps. Although the problem of detecting an expression of somnolence is a sub-problem of general emotional classification, the steps are in effect similar. To the three basic steps, face detection, facial expression data extraction and facial expression classification we have added two extra steps, image acquisition and determination of somnolence, relevant to our more particular problem.

3.1 Image acquisition

Image acquisition in the HADES system is done using a digital video camera in a controlled fashion, that is, under ideal circumstances with regard to lighting, etc. The results of our research, presented in the next chapter, are based on two different camera setups: a close-up of the eyes and a view of the head in its entirety. In both cases the subject is sitting in front of a white screen, to eliminate background disturbances, and does not tilt or rotate his head. The system can be considered as performing analysis of static image sequences according to the classification give by Pantic et al. Frames will only be captured on requests by other parts of the system. The video renderer filter draws the live feed to a window. Figure 1 illustrates the HADES system and its subsystems: Argos, Hypnos and Persephone.

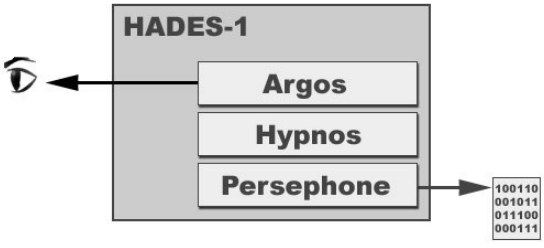

Figure 1: HADES system diagram

3.2 Face Detection

Our research has focused primarily on the eyes as a measure of somnolence detection. The detection of the eyes within the facial images is considered outside the scope of our research and is done by hand. The system operator monitors the incoming video stream and using a pointing device, marks the upper-left corner of the eyes and the bottom-right corner, creating a rectangle marking the area of interest. According to the Pantic et al. survey [6] this is the analytical approach; only the individual features of the face, in our case the eyes, are considered, instead of the face as a whole.

3.3 Facial Expression Data Extraction

Using the marked rectangle of interest (ROI) mentioned previously, we extract a vector of data based on three methods that will be described further on in this chapter. These methods are based on the light/color intensity of pixels in the image. The HADES system can easily be adapted to include different methods of data extraction, based on the rectangle of interest. The vector of data is continuously refreshed as the camera captures a new image. This is an analytical approach according to Pantic et al. [6]

3.4 Facial Expression Classification

When a vector of facial data has been obtained, we need to determine whether an expression of somnolence is present. In order to do this the HADES system requires two calibration images: calib 0 and calib 1. The first calibration image is essentially an image of the subject in neutral condition, without emotion and with open eyes. The second calibration image is an image of the subject in sleepy condition, that is, with eyes closed. From these calibration images two data vectors are obtained using the same rectangle of interest mentioned previously and using the same method. HADES can now determine whether an expression of somnolence is present in the input images by comparing the data vector with the vectors obtained from the calibration images. This principle is illustrated in figure 2.

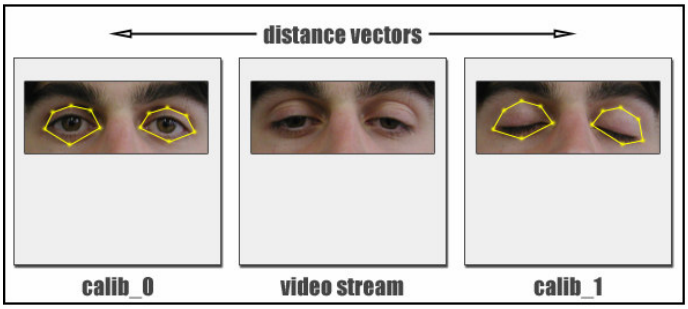

Figure 2: comparison of data vectors to determine somnolence

Essentially we have three vectors in an n-dimensional space. HADES calculates the distance between the data vector and the calib 0 vector and the distance between the data vector and the calib 1 vector. Of the two distances thus obtained the shortest is chosen to classify the expression. If the distance to the second calibration image, the subject with eyes closed, is smallest, an expression of somnolence is determined; if the distance to the first calibration image is smallest, the subject's expression is neutral. In our research we have assumed that the difference between the two calibration images (i.e. eyes-open and eyes-shut) is large enough for us to use in determining expressions of somnolence. The next section will provide data to underline this assumption. According to Pantic et al. in [6] our chosen approach is a template-based method.

3.5 Determination of Somnolence

This final step is necessary when the HADES system is implemented in environments where vigilance is of critical importance. In such situations an alarm, or other device, might need to be triggered to alert a driver or operator to his lack of attention. This alarm can not be triggered on the first determined expression of somnolence by the HADES system, for it is possible when capturing many frames (i.e. data vectors per time instance) that a blink of an eye can be classified as an expression of somnolence. The way this is handled by HADES is described further on in this section.

3.5.1 HypnosEuclidMean

The data vectors as they are calculated by the HypnosEuclidMean class are the averages of pixel intensity along horizontal and vertical scanlines, inside the rectangle of interest (i.e. the eyes). The main assumption is again the fact, that the intensity vector of the sleepy calibration image will differ substantially to the intensity vector of the neutral calibration image, thus allowing accurate comparison. To determine the distance between the data vectors, HypnosEuclidMean uses the Euclidean distance metric as given by equation 1. The Euclidean distance is calculated separately for the means in the horizontal and vertical direction. A correction factor called the λ factor is used to attach more importance to either the horizontal or vertical distances. The HypnosEuclidMean algorithm is simple and fast and proves very effective which will be demonstrated in the next chapter.

$$d_M = \left\{ \sum_{i=1}^{p} (x_i - y_i)^m \right\}^{\frac{1}{m}} \quad (1)$$

3.5.2 HypnosEuclidFull

HypnosEuclidFull is very similar to HypnosEuclid-Mean. It uses the same distance metric, as given by equation 1. The main difference is that HypnosEuclidFull uses the full rectangle of interest to obtain its data vectors and does no do any statistical pre-processing on them. In performance HypnosEuclidFull is also quite similar to HypnosEuclidMean.

3.5.3 HypnosDelta

The HypnosDelta subclass uses a different principle to calculate its data vectors: it determines the difference between subsequent pixel intensities on each scanline (vertical and horizontal). The main reason for using intensity differences is to attempt to make the system less sensitive to head movement. A slight movement of the head causes a severe drop in classification accuracy of the HypnosEuclidMean and HypnosEuclidFull algorithms, for their data vectors are based on actual pixel intensity values. The HypnosDelta data vector contains only differences as values. We found the Euclidean distance metric unsatisfactory for comparing the HypnosDelta data vector with the HypnosDelta vectors taken from the calibration images. Instead an absolute sum is used for comparison. This is done because there are very little boundary (i.e. 'large') values in the sleepy calibration image, thus producing a low vector sum. When compared to an arbitrary data vector, a relative lack of boundary values (i.e. low sum) will immediately indicate a small distance to the sleepy data vector, meaning an expression of somnolence is detected. The HypnosDelta subclass performs more calculations than HypnosMean and HypnosDelta per captured frame and is therefore slightly slower than its brethren classes.

4 Test Results

In order to measure and analyze the performance of our system, we have subjected HADES to two thorough testings. In the first set-up a close-up of the eyes was taken, whereas the second provided us with an image of the face as a whole. In both tests, the lighting conditions were controlled and the subject kept his head reasonably still. The test setup is shown in figure 3. This section presents the data obtained from the first set-up only. The results of the second set-up were surprisingly similar and will not be described in detail. One of the system requirements is operation in real-time or quasi real-time. Measurements of the speed of the different algorithms we used justify the conclusion that the HADES system is indeed quasi real-time: it uses about 16–17 frames per second. The results obtained from the HypnosEuclidFull algorithm are depicted in figure 4.

Figure 3: Simulation environment

The x-axis represents the elapsed time in seconds, whereas the y-axis represents the distance of the examined image to the calib 0 and calib 1 image. The HADES system classifies an image as 'sleepy' when the distance to the neutral face surpasses the distance to the sleepy face.

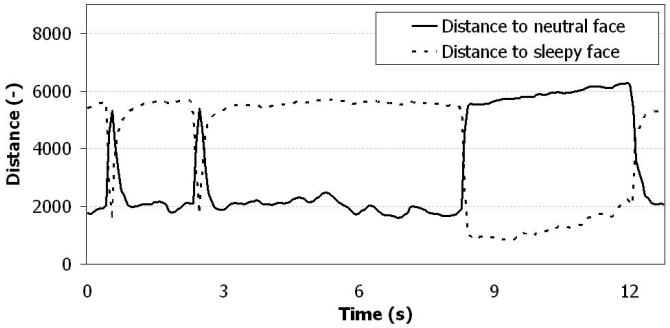

Figure 4: HypnosEuclidFull results

Comparing these results to our personal observations we can conclude that the system did not misclassify any image in this particular image sequence. The blinking of the eye at around T=0.5 and T=2.3 can clearly be seen in the graph. This confirms our assumption that eye closure in only one image does not necessarily mean the test subject is in a state of somnolence. HypnosDelta, which focuses on differences between subsequent pixel intensities, shows us different results. As is apparent by comparing figure 5 to figure 4, the HypnosDelta algorithm also classifies closed eyes correctly in non-transitionary, stable situations.

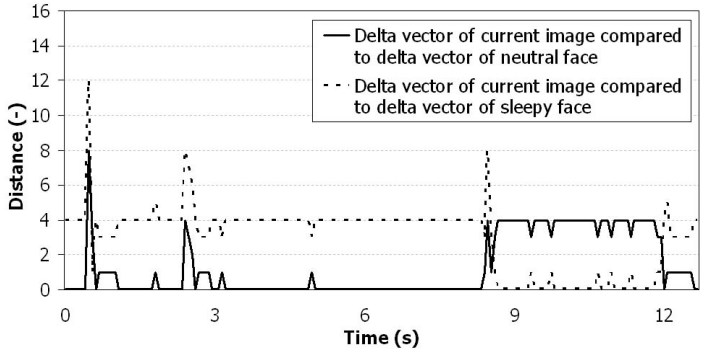

Figure 5: HypnosDelta results

The delta vector of the current image compared to the delta vector of the neutral face, as well as the delta vector of the current image compared to the delta vector of the sleepy face, both show peak values in transition periods from eyes opened to eyes closed. In the HypnosEuclidFull algorithm the distance between the vectors in transition periods shows both a peak and a drop in value. The double-peak values in the HypnosDelta algorithm transitions lead to unreliable classification, although the transition data could be valuable for purposeful study of transitionary behavior.

Table 1: Excerpt of Hypnos data at a transition

Time	HypnosEuclidFull			HypnosDelta		
	N	S	sleepy	N	S	sleepy
t1	4205	1518	true	4	0	true
t2	4195	1539	true	4	0	true
t3	3734	1943	true	0	4	false
t4	3035	2774	true	0	4	false
t5	1927	3865	false	1	3	false
t6	1916	3915	false	1	3	false

In non-transitionary states where the eyes are either open or closed, both algorithms perform equally well. When comparing the results from the different algorithms, the transitions between states provide the most interesting data. Table 1 lists the data of a single transition between closed and open eyes and table 2 lists the data of the first eye blink in the first test, which is actually a double transition. In these tables the 'N' signifies the distance from the acquired image to the calib 0 image, and 'S' signifies the distance to the calib 1 image. As was mentioned in the previous section, HypnosDelta does not classify this blink as 'closed eyes'.

Table 2: Excerpt of Hypnos data at eye blink

Time	HypnosEuclidFull			HypnosDelta		
	N	S	sleepy	N	S	sleepy
t1	1707	3926	false	0	4	false
t2	1788	4135	false	0	4	false
t3	3074	1674	true	8	12	false
t4	3000	2481	true	1	5	false
t5	2304	3345	false	1	3	false
t6	2054	3592	false	1	3	false

The Euclidean based algorithm does, in fact, detect an eye closure. Although the algorithms use the exact same images as a source, they might classify these images totally different. This is apparent from row 3 in the same table, where the HypnosDelta classifies a 'neutral' with full confidence, and HypnosEuclidFull does this same classification two images later (which means 250 ms). Our observations of the two implemented algorithms lead us to conclude that neither of the algorithms performs

significantly better than the other, which is not surprising considering their similarity.

5 Further Research

Upon examining the HADES system, we discover several areas for improvement, which reflect the steps in the three-tier framework as introduced in section 2: face detection, facial expression feature extraction and facial expression classification. The face or eye detection is of great interest, since the current system misclassifies expressions when there is substantial movement of the head. Other methods of feature extraction than current algorithms might enhance the system and the data comparison methods in our algorithms can be improved. Replacement of the user-selected rectangle of interest by an automated eye detection module would greatly enhance the system. This module can utilize the fact that drivers tend to keep their heads practically in the same position. The spatio-temporal relation between images can be used for tracking movements of the head, instead of processing each image independently. When it comes to facial expression feature extraction, many different approaches have been discussed by Pantic et al. [6]. Kobayashi and Hara. [5] use brightness distributions at certain fixed vertical lines in the face, which is similar to the HADES system. Changing the current color model from RGB to HSV, a model based on brightness values, might enable us to retrieve more information from the images. Other, holistic, approaches such as the fitting of elastic graphs to facial images, or utilization of eigenfaces based on PCA, are possible and described in Pantic et al. The classification of the images can also be improved. The Mahalanobis distance is a commonly used metric in the field of image classification. Compared to the Euclidean distance it corrects for correlation between the different features, since it is very sensitive to intervariable changes in the template images. Besides changing the distance calculations it is possible to use alternative classification methods, such as neural networks or rule-based systems. Especially the former is commonly used in the world of facial expression recognition.

6 Conclusion

In this paper we have presented a system for recognizing expressions of somnolence in the human face. When assessing its performance, we may conclude that our system works well under ideal circumstances. In ideal situations the subject under scrutiny keeps his head straight and faces directly forward continuously. Even slight movements of the head however severely affect correct classification of somnolence. Even more demanding circumstances, such as person independency, have not been taken into account at all in the current HADES system. Considering the current level of technology we consider the application of a HADES-like system inside a moving vehicle, or anywhere else for that matter, where constant vigilance is of critical importance, well within the scope of possibilities.

References

[1] J.F. Cohn, A.J. Zlochower, J.J. Lien, and T. Kanade, “Feature-Point Tracking by Optical Flow Discriminates Subtle Differences in Facial Expression”, Proceedings of the International Conference on Automatic Face and Gesture Recognition, pp. 396–401, 1998.

[2] J. Faber, M. Nov´ak, P. Svoboda, and V. Tatarinov, “Electrical Brain Wave Analysis During Hypnagogium”, Neural Network World, vol. 1, pp. 41–54, 2003.

[3] C.L. Huang and Y.M. Huang, “Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification”, Journal of Visual Communication and Image Representation, vol. 8, no. 3, pp. 278–290, 1997.

[4] S. Kimura and M. Yachida, “Facial Expression Recognition and Its Degree Estimation”, Proceedings of Computer Vision and Pattern Recognition, pp. 295–300, 1997.

[5] H. Kobayashi and F. Hara, “Recognition of Six Basic Facial Expressions and Their Strength by Neural Network”, Proceedings of the International Workshop on Robot and Human Communication, pp. 387–391, 1992.

[6] M. Pantic and L.J.M. Rothkrantz, “Automatic analysis of facial expressions: the state of the art”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1424–1445, 2000.

[7] M. Pantic and L.J.M. Rothkrantz, “An Expert System for Multiple Emotional Classification of Facial Expressions”, Proceedings of the International Conference on Tools with Artificial Intelligence, pp. 113–120, 1999.

[8] M.A. Spaans and H.J. Dikkers, Facial recognition system for driver vigilance monitoring, Res. Rep.No. LSS 169/03, Czech Technical University, Prague, June 2003.