Towards Automatic Recognition of Narcolepsy with Cataplexy

Ilaria Bartolini
Department of Computer Science and Engineering (DISI)
Alma Mater Studiorum, University of Bologna
Bologna, Italy

Andrea Di Luzio
Department of Computer Science and Engineering (DISI)
Alma Mater Studiorum, University of Bologna
Bologna, Italy

ABSTRACT

Narcolepsy with cataplexy is a severe lifelong disorder characterized, among the others, by sudden loss of bilateral face muscle tone triggered by emotions (cataplexy). The current approach followed by neurologists for the classification of such abnormal motor behavior is based on a completely *manual* analysis of video recordings of patients undergoing emotional stimulation made *on-site* by medical specialists. With the double aim of supporting neurologists in such a delicate task and facilitating the experience of patients, avoiding them to conduct video recordings at hospitals, in this position paper we advocate the use of *automatic* video content analysis techniques and *mobile multimedia* technologies in order to solve the problem. In particular, we propose a medical tool, built on top a general and extensible framework for the effective and efficient management of video collections, that allows patients (i.e., video producers) to record videos through the use of smart devices and neurologists (i.e., video consumers) to automatically identify the presence of the disease by means of a user friendly GUI. Preliminary results achieved on the recognition of one of most recurrent cataplexy motor behaviours pattern (namely, ptosis) and conducted on real data demonstrate the effectiveness of the proposed solution and encourage further investigations in this direction.

CCS CONCEPTS

• **Computing methodologies** → **Visual content-based indexing and retrieval**; • **Applied computing** → **Health care information systems**;

KEYWORDS

Automatic video content analysis, Motor behaviour patterns, Mobile multimedia technologies

ACM Reference Format:
Ilaria Bartolini and Andrea Di Luzio. 2017. Towards Automatic Recognition of Narcolepsy with Cataplexy . In *MoMM '17: The 15th International Conference on Advances in Mobile Computing & Multimedia, December 4–6, 2017, Salzburg, Austria.* ACM, New York, NY, USA, 5 pages.
https://doi.org/10.1145/3151848.3151875

MoMM '17, December 4–6, 2017, Salzburg, Austria

ACM ISBN 978-1-4503-5300-7/17/12…$15.00
https://doi.org/10.1145/3151848.3151875

1 INTRODUCTION

Narcolepsy with cataplexy is a rare disorder mainly arising in young adults/children characterized by daytime sleepiness, sudden loss of muscle tone while awake triggered by emotional stimuli (cataplexy), hallucinations, sleep paralysis, and disturbed nocturnal sleep [1]. The current approach for the recognition and classification of such abnormal motor behaviour is based on a completely *manual* analysis of video recordings of patients undergoing emotional stimulation made *on-site* by medical specialists. This is due to the complete absence of automatic technological solutions able to properly support neurologists in a such delicate task [1].

Few scientific studies have considered the video-polygraphic features of cataplexy in adult age and only recently the motor phenotype of childhood cataplexy has been described exclusively using video recordings of the attacks evoked by watching funny cartoons [1]. These studies showed that in the context of the physiological response to the laughter there are the distinctive elements of cataplexy, called *motor behaviours patterns*, particularly evident at the level of the *facial expression changes* that are, however, still to be manually detected by neurologists [1].

From above arguments, it is evident that a system able to detect the "correct" facial expression changes from video recordings of patients would be able to automatically identify the presence of the disease. Motivated by this observation, in this position paper we advocate the use of automatic video content analysis techniques in order to provide a novel mobile medical tool supporting neurologists in the delicate task of the disease recognition. The tool exploits the visual content of video recordings made on patients undergoing emotional stimulation through the vision of funny movies designed to evoke the laughter. In details, by means of an intuitive and user friendly GUI, the medical tool effectively supports neurologists with (1) automatic recognition of motor behaviors patterns that determine with certainty the presence of childhood narcolepsy

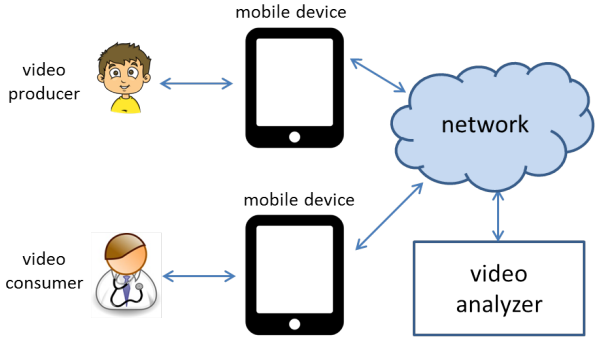

Figure 1: Medical tool architecture.

with cataplexy, and (2) video recordings searching and browsing facilities.

Further, in order to facilitate the experience of patients, avoiding them to conduct video recordings at hospitals and/or research centers, we propose the use of mobile multimedia technology to improve the availability of the system. In such architecture, patients play the role of "video producers" by exploiting a mobile device (e.g., an iPad/iPhone), while neurologists are the "video consumers" since they receive the output of visual analysis performed on their patients (e.g., automatic "alert" generation in case of positive classification of the disease on patients). Video analysis is performed by a back-end service and is built on top of SHIATSU, a general and extensible framework for video retrieval which is based on the (semi-)automatic hierarchical semantic annotation of videos exploiting the analysis of their visual content [8]. The overall architecture of the system is depicted in Figure 1.

To the best of our knowledge, this work represent the first attempt to tackle the problem of the automatic recognition of narcolepsy with cataplexy. In particular, we start our investigation by focusing on one of the most recurrent identified motor phenomena, namely *ptosis*, that is often displayed by children affected by the deases. Technically speaking, ptosis is a drooping or falling of the upper eyelid [1] that we are able to model exploiting the facial landmark detector OpenFace [2].

The rest of the paper is as follows: Section 2 provides useful background for the complete understanding of the proposed medical tool. In Section 3 we describe how the video analyzer is able to automatically detect ptosis, while Section 4 shows the GUI of the medical front-end. In Section 5 we comment some preliminary experimental results based on real data and in Section 6 we conclude by drawing some interesting directions for future studies.

2 BACKGROUND

This section provides basics on SHIATSU and OpenFace.

2.1 The SHIATSU Framework

SHIATSU is a general and extensible framework for content-based video retrieval. Its engine consists of three main components: (1) the visual features extractor is used to automatically extract visual features from video frames; (2) the annotation processor implements algorithms for the automatic tagging of videos by exploiting video features; (3) the query processor contains the logic for retrieval of videos based on semantics (tags) and/or similarity (features) [8].

The basic assumption at the core of automatic tagging in SHIATSU is that frames with a similar visual content also convey the same semantic content. According to this principle, tagging of a video in SHIATSU is performed in a hierarchical way: (1) videos are automatically segmented into shots (containing key-frames with a same visual content), i.e., sequences of consecutive frames that share a common visual content; (2) so-obtained shots are then (semi-)automatically labelled using high-level concepts, according to the above mentioned similarity principle; (3) finally, tags assigned to videos shots are used to appropriately annotate the whole video [8].

Tagging and retrieval are based on *multidimensional* taxonomies; this allows to connect each tag with its intended meaning, exploiting the coexistence of multiple, independent classification criteria. According to this multidimensional approach, labels belonging to different dimensions may have separate meanings, while each dimension will represent the meaning of high-level concepts contained therein, providing a disambiguation of their semantics [8].

2.2 OpenFace

OpenFace is one of the most popular open source facial landmark detector used to localize facial features, like eyes and eyelid contours [2]. OpenFace employs a novel instance of the Constrained Local Model (CLM) framework called Constrained Local Neural Field (CLNF) dealing with the issues of feature detection in complex scenes reaching state-of-the-art performances when detecting facial landmarks across different illuminations. A Constrained Local Model (CLM) is a class of methods of locating sets of keypoints (constrained by a statistical shape model) on a target image (or video frame). The general CLM approach consists of three basic steps: (1) sample a region from the image around the current estimate; (2) for each keypoint, generate a *response image* giving a cost for having the point at each pixel; (3) searching for a combination of points which optimises the total cost, by manipulating the shape model parameters [2].

CLM is made by three main components: (1) a point distribution model (PDM); (2) local detectors that evaluate the probability of a landmark being aligned at a particular pixel location; (3) the fitting strategy used.

CLNF is an instance of a CLM that uses more advanced local detectors and an optimization function [2]. The two basic components of CLNF are: (1) PDM which captures landmark shape variations; (2) local detectors which capture local appearance variations of each landmark.

3 VIDEO ANALYZER BACK-END

To characterize features for ptosis pattern we exploit OpenFace facial landmarks (refer Section 2). As described in the Section 1, ptosis is a drooping or falling of the upper eyelid. This, however, should not be mistaken as a (regular) eye blink. For this reason, we need to detect closures of the eyes that last longer than a typical blink. This is performed by analyzing facial features on each video frame to detect whether eyes are open or closed: we measure the time of each eyelid drooping as the number of consecutive frames for which the eye closure is detected times the frame rate.

The first step of the pattern characterization process consists in detecting and extracting patients' facial landmarks of interest from each frame of the video; this is necessary because each patient might have facial features different from each other.

In details, for each frame, a 12-dimensional (12-D) feature vector is extracted. The vector contains the list of coordinates (x,y) of the six landmarks that characterize the shape of the eye (see Figure 2 for a real example).

Formally speaking, let us denote with V_h each video recording of patients, with $VF_{h,j}$ the j-th frame of V_h, and with $\mathbf{p}_j = (px_j, py_j)$ each landmark. The set of features representing video V_h

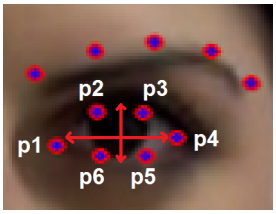

Figure 2: Example of landmarks of an open eye.

is $V_h = \{(\mathbf{p}_1, \ldots, \mathbf{p}_6)_1, \ldots, (\mathbf{p}_1, \ldots, \mathbf{p}_6)_N\}$ with N the total number of frames in V_h and $FV_{h,j} = (\mathbf{p}_1, \ldots, \mathbf{p}_6)_j$ the feature vector associated to frame $FV_{h,j}$ of V_h.

The *Eye Aspect Ratio* (EAR) is defined as the ratio of the eye *height* to the eye *width* [5], that is:

$$EAR = \frac{\|\mathbf{p}_2 - \mathbf{p}_6\| + \|\mathbf{p}_3 - \mathbf{p}_5\|}{2\|\mathbf{p}_1 - \mathbf{p}_4\|}, \tag{1}$$

where $\mathbf{p}_1, \ldots, \mathbf{p}_6$ are the landmarks and $\|\mathbf{x}\|$ denotes the norm of vector $\mathbf{x}$.

The semantics of EAR are as follows: when an eye is closing, EAR approaches zero, whereas when the eye is completely open, EAR attains its maximum value (which varies from person to person). EAR is partially invariant to head pose and fully invariant to uniform image scaling and in-place face rotation [5]. Moreover, for a more stable characterization of EAR, for each frame we use the average of the left and right eye EAR values.

We perform the EAR extraction process on the video recordings at both base conditions (termed *baseline*) and while the patients were undergoing emotional stimulation. This is necessary to characterize precisely the facial features of each patient under normal condition. In this way, by measuring, for example, the median value ($\overline{EAR}$) of EAR values for the baseline video recordings, we obtain the characterization of the patient open eyes (because we can safely assume that, during baseline recordings, patient eyes are normally open [4].

Terminating the EAR process extraction, for each video we obtain a N-D time series. We denote with EAR_B and EAR_{ES} the time series corresponding to baseline and undergoing emotional stimulation videos, respectively.

Given the j-th frame of the recorded stimulation video, we consider the eyes as closed if the normalized difference between the EAR for current frame, EAR_{ES_j}, and $\overline{EAR}$ is higher than a threshold:

$$\frac{|\, EAR_{ES_j} - \overline{EAR}\,|}{\overline{EAR}} > T \tag{2}$$

Since threshold T is an important parameter that strongly influences the accuracy of the detector, we evaluate through experimentation its optimal value (see Section 5). Clearly, we suppose here that "normally" patient eyes are opened, so that whenever inequality 2 holds we are observing the eyes closure.

Due to the noise usually present in the EAR_{ES} time series, an approximate version of them is derived through triangular smoothing: each point of the original time series, EAR_{ES_j}, is replaced with the weighted average of its 5 adjacent points, $EAR^S_{ES_j}$, computed as:

$$EAR^S_{ES_j} = \frac{EAR_{ES_{j-2}} + 2EAR_{ES_{j-1}} + 3EAR_{ES_j} + 2EAR_{ES_{j+1}} + EAR_{ES_{j+2}}}{9}, \tag{3}$$

with j in $[3, N-2]$ (for other values of j the derivation is similar and not detailed here). Such value of $EAR^S_{ES_j}$ replaces EAR_{ES_j} in Equation 2 (the use of $EAR^S_{ES_j}$ in place of EAR_{ES_j} is experimentally demonstrated in Section 5).

Finally, we define the presence or absence of ptosis by measuring the length of the time series corresponding to a "long enough" sequence of frames with closed eyes.

State-of-the-art approach to this is measuring the *PERcentage of eye CLOSure* (PERCLOS), corresponding to the measurement of time the pupils of the eyes are occluded [3, 6]: this would clearly reflect eyelid closures slower than blinks.

Starting from above observation, the approach we propose here defines ptosis as present if the amount of time when eyes are closed exceeds the maximum duration of an eye blink, which can be measured as about 400 ms [7]. This is reflected in Algorithm 1 that detects presence/absence of ptosis on a series of EAR^S_{ES} values.

Algorithm 1 Ptosis detector algorithm

1: **Input**: EAR^S_{ES}, $\overline{EAR}$, T
2: **Output**: P^S_{ES} ▷ Ptosis tag vector
3: **for** $j \leftarrow 1, N$ **do** $EC_j \leftarrow false$ ▷ Eyes closed tag vector
4: **end for**
5: **for** $j \leftarrow 1, N$ **do** ▷ For all frames in EAR^S_{ES}
6: compute $EAR^S_{ES_j}$ as in Equation 3
7: **if** $\frac{|EAR^S_{ES_j} - \overline{EAR}|}{\overline{EAR}} > T$ **then**
8: $EC_j \leftarrow true$ ▷ Eyes is closed at frame j
9: **if** $!blinks(j, w, EC)$ **then** $P^S_{ES_j} \leftarrow true$ ▷ Ptosis detected
10: **else** $P^S_{ES_j} \leftarrow false$
11: **end if**
12: **end if**
13: **end for**

In line 8 of Algorithm 1, the EC vector is used to determine if, at the j-th frame, eyes were closed. Then, function $blinks(j, w, EC)$ checks whether, in the w frames preceding j (i.e., $j-w, j-w+1, \ldots j-1$) at least one of the EC values is false, i.e., eyes were opened in at least one of the preceding w frames (clearly, the value of w depends on the frame rate f and can be computed as $400/f$).

Finally, it has to be highlighted the fact that, during videos with emotional stimulation, it can be the case that landmark extraction fails: this happens, for example, due to head drops, which can be considered as part of the ptosis (this is generally true, according to visual observations made on patients). Therefore, in Algorithm 1 if, for a given frame j, we cannot compute the value of EAR_{ES_j}, the ptosis tag is copied from the previous frame, i.e., $P^S_{ES_j} \leftarrow P^S_{ES_{j-1}}$.

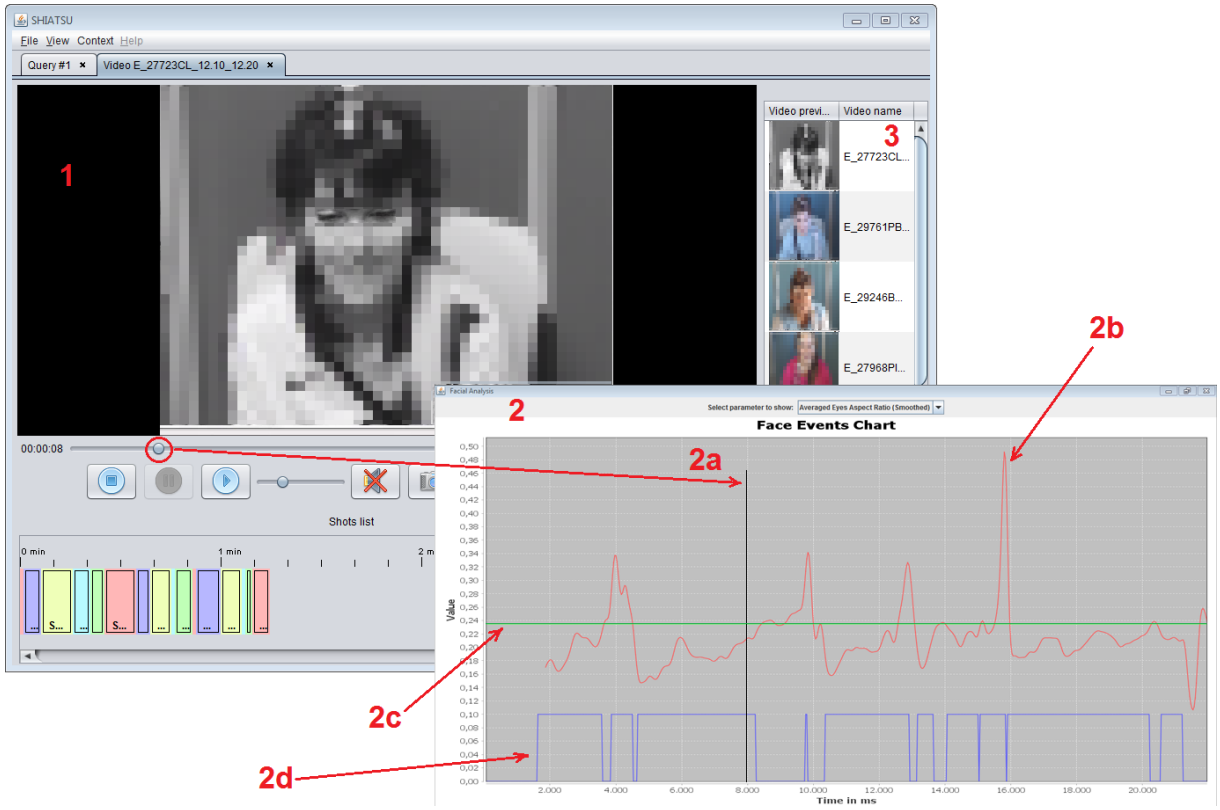

Figure 3: Medical front-end GUI: presence of ptosis is indicated by the blue line (2d).

4 MEDICAL FRONT-END GUI

The proposed tool is implemented on top of the SHIATSU framework (see section 2). This gives us a number of advanced services ranging from video frame splitting, frame feature extraction, and feature-based retrieval to data persistence and visualization. We can use such services for free and therefore focus our effort on medical aspects only with guarantee of the creation of an advanced and complete medical instrument.

Figure 3 presents the intuitive and easy-to-use GUI designed for video consumers. In details, the GUI allows the neurologists to study ptosis presence in a patient by providing specific functionality for inspecting both the original video recordings and the computed ptosis time series. Further, neurologists have the possibility to search for video recordings of patients presenting similar motor phenomena to a given one and to quickly identify possible correlations regarding the motor phenomena in different patients.

The GUI basically consist of three different sections: the central part (1) represents the video play area with its control buttons; the external panel (2) shows values of $EAR^S_{ES_j}$ (2b), $\overline{EAR}$ (2c), and $P^S_{ES_j}$ (2d) for the current patient (the vertical time-line (2a) highlights the currently displayed time series values); finally, the vertical panel on the right side (3) is the area where videos similar to the current (query) one are shown (in descending order of the percentage of frames where ptosis is present).

When the neurologist user starts video play, the time series panel (2) is opened. The vertical time-line (2a) is synchronized with the video playback in order to allow the neurologist to compare the automatically detected motor phenomena (area 2d) with corresponding facial expressions displayed by the patient in the currently played video recordings.

The physical separation of the panel containing the graph of motor phenomena from the main window of SHIATSU is designed to allow the simultaneous viewing of graphs belonging to different videos, or belonging to the same video in case neurologists want to simultaneously observe various parameters of the same patient.

5 EXPERIMENTAL EVALUATION

Here we present preliminary results performed on real data obtained from video recordings related to six patients. Such video recordings amount for a total of 10 minutes and 17150 frames (videos were recorded with a frame rate $f = 30$ frame/s, thus obtaining a corresponding window $w = 15$ for evaluating the presence of a eye blink). Each of the underlying emotional stimulated videos comes with provided tags (manually associated by neurologists) indicating the time periods where the patient is undergoing narcolepsy with cataplexy. This allow us to objectively evaluate the performance of our analyzer by using classic precision/recall values. Precision p is defined as the fraction of frames that are correctly classified as characterizing the disease, while recall r is the

fraction of frames showing the disease that are correctly classified as containing narcolepsy with cataplexy. We highlight here the fact that, unlike classical information retrieval systems, reducing the number of false negatives (i.e., of patients affected by the disease but not recognized by the system) is of extreme importance here, with respect to reducing the number of false positives (i.e., of patients that are incorrectly classified as showing the disease), thus higher recall values are to be preferred wrt higher precision values.

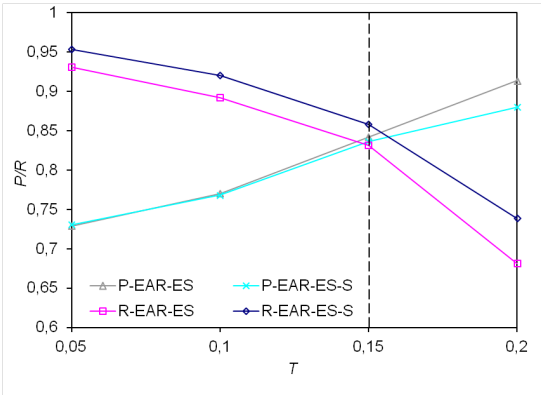

Figure 4: Precision (P) and Recall (R) curves varying the threshold T for EAR_{ES} and EAR^{S}_{ES} time series, respectivel.

Our first experiment aims at showing (1) how the threshold T in Equation 2 can be experimentally assessed and (2) that the use of smoothed values (Equation 3) improves performance. Figure 4 shows P/R values for different values of the threshold T. The obvious breakpoint value, i.e., the value of T maximizing both precision and recall values, is obtained for $T = 0.15$. Thus, this is the value that will be use in the following. Moreover, Figure 4 shows graphs for original (EAR_{ES}) and smoothed (EAR^{S}_{ES}) values of the EAR descriptors. It is clear that, for all considered values of the threshold T, smoothed values of EAR achieve better performance, proving the higher accuracy of the proposed descriptors.

patient	P	R
1	0.70	0.81
2	0.97	0.99
3	0.99	0.77
4	0.82	0.79
5	0.98	0.93
6	0.55	0.81
avg.	0.84	0.86

Table 1: Effectiveness of the proposed method for different patients.

Table 1 reports the performance of the proposed classification tool on individual videos. Results show that recall values are usually higher than precision values, as requested by the particular case study at hand; only for "easy" cases (i.e., for patients with high precision values) R is not higher than P. Clearly, this is only a preliminary result based on a dataset which only contains patients exhibiting the disease, thus the quite good results obtained by the proposed method should be evaluated also for cases where the disease is absent.

Finally, we include a brief discussion about efficiency of the proposed technique. All experiments were run using a 2C/4T @ 2.20Ghz CPU equipped with 4G RAM. On this setup (which is typical of a low-end machine), we were able to extract EAR descriptors in real time. Clearly, this is the more time consuming operation in Algorithm 1, thus it is proven that the whole process of automatic ptosis detection can be performed during a single emotional stimulated video recording session.

6 CONCLUSIONS

In this paper, we proposed a novel mobile architecture for the automatic classification of narcolepsy with cataplexy. For this, we exploit the visual content of video recordings made on patients caraterized by OpenFace features and the SHIATSU framework. Our preliminary study focused on ptosis, one of most recurrent cataplexy motor behaviours pattern. Experiments conducted on real data demonstrated the accuracy of the proposed solution and encourage further investigations on this direction. We highlight that presented results were obtained using low quality video recordings: we are in the process of applying our detection techniques on a larger set of high resolution videos recorded by technicians after providing ad doc instructions to patients. In the future, we plan to investigate other motor phenomena that characterize narcolepsy with cataplexy by providing a holistic model using correlations among such phenomena.

REFERENCES

[1] G. Plazzi, F. Pizza, V. Palaia, C. Franceschini, F. Poli, K. K. Moghadam, P. Cortelli, L. Nobili, O. Bruni, Y. Dauvilliers, L. Lin, M. J. Edwards, E. Mignot and K. P. Bhatia. Complex movement disorders at disease onset in childhood narcolepsy with cataplexy. *Brain: A Journal of Neurology*, 2011.

[2] T. Baltrušaitis, P. Robinson, and L-P. Morency. OpenFace: an open source facial behavior analysis toolkit. *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[3] W.W. Wierwille, L.A. Ellsworth, S. S. Wreggit, R. J. Fairbanks, and C. L. Kirn. Research on vehicle based driver status/performance monitoring: development, validation, and refinement of algorithms for detection of driver drowsiness. *National Highway Traffic Safety Administration Final Report: DOT HS 808 247*, 1994.

[4] F.M. Sukno, S-K. Pavani, C. Butakoff, and A. F. Frangi. Automatic assessment of eye blinking patterns through statistical shape models. *ICVS '09*, 2009.

[5] T. Soukupová and J. Čech. Real-time eye blink detection using facial landmarks. *CVWW '16*, 2016.

[6] R. Knipling and P. Rau. PERCLOS: a valid psychophysiological measure of alertness as assessed by psychomotor vigilance. *Washington: Office of Motor Carriers*, 1998.

[7] H. R. Schiffman. Sensation and perception: an integrated approach. John Wiley and Sons, 2001.

[8] I. Bartolini, M. Patella, and C. Romani. SHIATSU: tagging and retrieving videos without worries. *Multimedia Tools and Applications*, 2013.