Automated Characterization of Mouth Activity for Stress and Anxiety Assessment

A.Pampouchidou [♦1], M.Pediaditis [△2], F.Chiarugi [△3], K.Marias [△4], P.Simos [⋈5], F.Yang [♦6], F.Meriaudeau [♦*7], M.Tsiknakis [△◇8]

[♦] *Le2i Laboratory, University of Burgundy, Le Creusot, France*
[1] anastasia.pampouchidou@gmail.com, [6] fanyang@u-bourgogne.fr,

[△] *Institute of Computer Science, Foundation for Research & Technology - Hellas, Heraklion, Crete, Greece*
[2] mped@ics.forth.gr, [3] chiarugi@ics.forth.gr, [4] kmarias@ics.forth.gr

[⋈] *Division of Psychiatry, School of Medicine, University of Crete, Heraklion, Crete, Greece*
[5] akis.simos@gmail.com

[*] *CISIR, Electrical Engineering Department, Universiti Teknologi Petronas, Malaysia.*
[7] fmeriau@u-bourgogne.fr

[◇] *Technological Educational Institute of Crete, Department of Informatics Engineering, Heraklion, Crete, Greece*
[8] tsiknaki@ie.teicrete.gr

***Abstract*—Non-verbal information portrayed by human facial expression, apart from emotional cues also encompasses information relevant to psychophysical status. Mouth activities in particular have been found to correlate with signs of several conditions; depressed people smile less, while those in fatigue yawn more. In this paper, we present a semi-automated, robust and efficient algorithm for extracting mouth activity from video recordings based on Eigen-features and template-matching. The algorithm was evaluated for mouth openings and mouth deformations, on a minimum specification dataset of 640x480 resolution and 15 fps. The extracted features were the signals of mouth expansion (openness estimation) and correlation (deformation estimation). The achieved classification accuracy reached 89.17%. A second series of experimental results, for the preliminary evaluation of the proposed algorithm in assessing stress/anxiety, took place using an additional dataset. The proposed algorithm showed consistent performance across both datasets, which indicates high robustness. Furthermore, normalized openings per minute, and average openness intensity were extracted as video-based features, resulting in a significant difference between video recordings of stressed/anxious versus relaxed subjects.**

***Index Terms*—mouth gesture recognition, image processing, stress, anxiety, automatic assessment**

I. Introduction

Non-verbal communication conveyed by the human face, besides portraying individuals emotional status [1], encompasses important information, which can contribute to a reliable psychophysical evaluation [2] [3] if properly analyzed. Different mouth activities can be interpreted as signs of several conditions; e.g. people suffering from depression smile less and generally show reduced mouth activity [4], while those in fatigue have droopy mouth corners [5]. Furthermore, lip deformation [6] [7], mouth opening [8], lip parting [9], as well as mouth movements in generally [10] have been shown to increase during stress.

The work described herein, presents a semi-automated, algorithm for extracting mouth activity from video recordings based on Eigen-features and template-matching. Experiments were conducted for mouth openings and mouth deformations, with the extracted features being the two signals of mouth expansion and correlation respectively. Frequency, average intensity as well as the duration of openness can easily be derived from the two signals in order to provide video-based features. Preliminary findings for stress/anxiety assessment are also presented.

II. State-of-the-Art

Mouth activity is being investigated a lot as part of systems that employ action units (AUs) classification. According to Ekmans Facial Action Coding System (FACS) [1] any given combination of AUs is associated with the presence of specific emotions. An example of such an approach is that of the Computer Expression Recognition Toolbox (CERT) [11]. Another area in which mouth activity classification has found an application is visual speech analysis [12].

A basic prerequisite for any of the aforementioned applications, is an accurate mouth detection. The most popular object detector is that of Viola and Jones (V&J) [13] based on Haar features. Another approach commonly followed, in order to obtain and track the mouth region, is the active contour method [14].

Feature extraction methods for mouth activity recognition found in literature include appearance features derived from Gabor wavelets, scale-invariant feature transform (SIFT) descriptors [15], deformable models [6] and motion based features such as space-time interest points [16]. The proposed work lies in the area of biometrics [17], and exploits geometrical and textural information. Although it is of significantly lower complexity than the aforementioned approaches, yet manages to achieve comparable experimental results.

978-1-5090-1817-8/16/$31.00 ©2016 IEEE

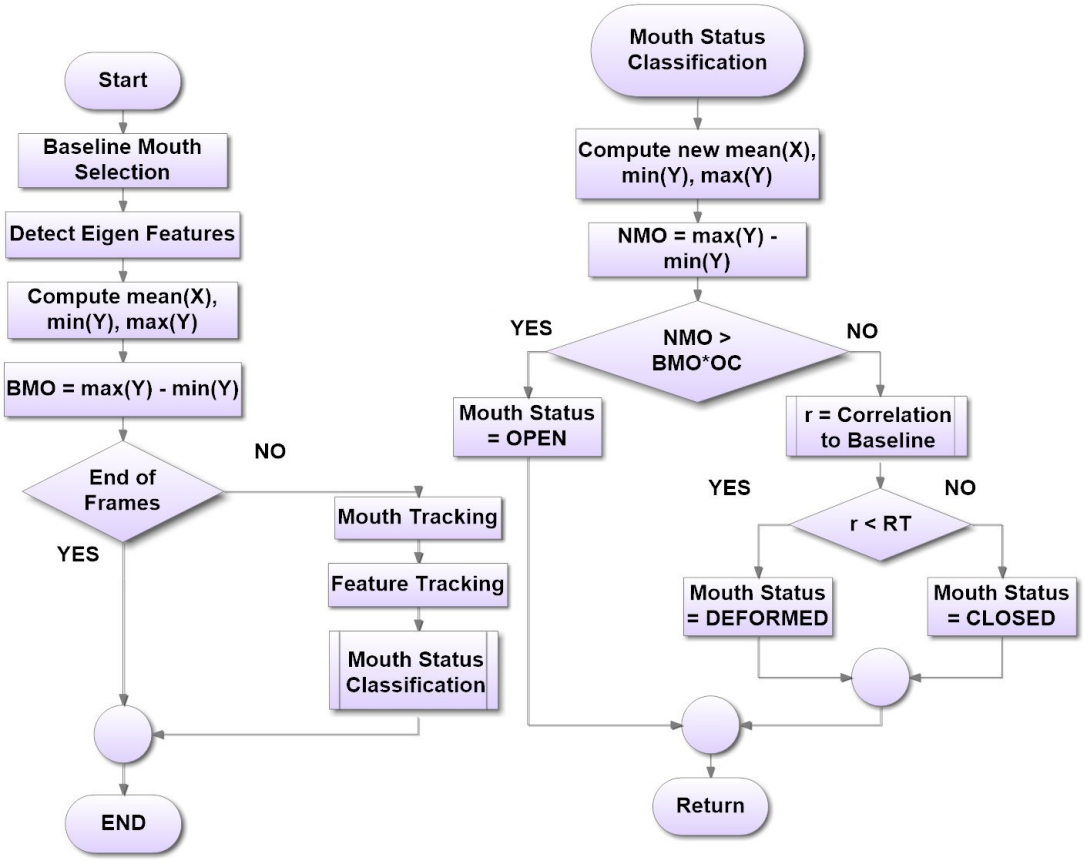


Fig. 1: System flow chart (BMO: Baseline Mouth Openness, NMO: New Mouth Openness, R: correlation coefficient)

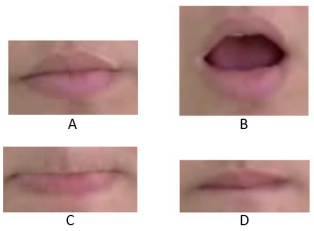


Fig. 2: **A)** Closed/baseline mouth, **B)** Open, **C) & D)** examples of deformation

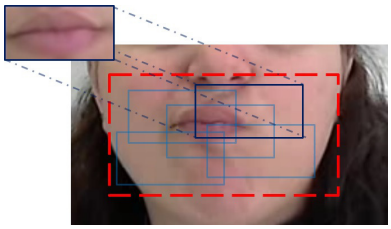


Fig. 3: Template-matching tracking

III. Methodology

The rationale for the selected method lies on a user-specific baseline of the closed mouth and on the idea that when the mouth is open it is not deformed. Based on the previous assumptions, at this phase the problem is dealt as a three class classification problem {CLOSED, OPEN, DEFORMED} (cf. Fig. 2). In order to ascertain that the baseline is defined accurately, manual selection of the mouth region of interest (ROI) is performed. Eigen-feature points are automatically located within the mouth ROI with the method described in [18], which detects reliable patterns (e.g. corners) to be tracked; for the present work the strongest thirty points are considered sufficient. In order to define the baseline mouth openness (BMO), two points are computed from the previously selected strongest thirty, as follows:

$$Q_1 = [mean(X), min(Y)],\ Q_2 = [mean(X), max(Y)] \quad \text{and } BMO = Q_1 - Q_2 \tag{1}$$

where X and Y are the two vectors with the image coordinates of the Eigen-feature points. Having obtained the baseline mouth ROI, the features and the BMO, the prerequisite parameters of the algorithm are all set. For testing the algorithm, and in order to avoid manual selection of the mouth ROI every time, a template-matching tracking algorithm (illustrated in Fig. 3) was implemented, initialized with the baseline mouth ROI. Considering small displacements from one frame to the next, there is no need to scan the whole image, but just around the previously defined ROI. In the example of Fig. 3, the scanning takes place within the largest dashed (red) bounding

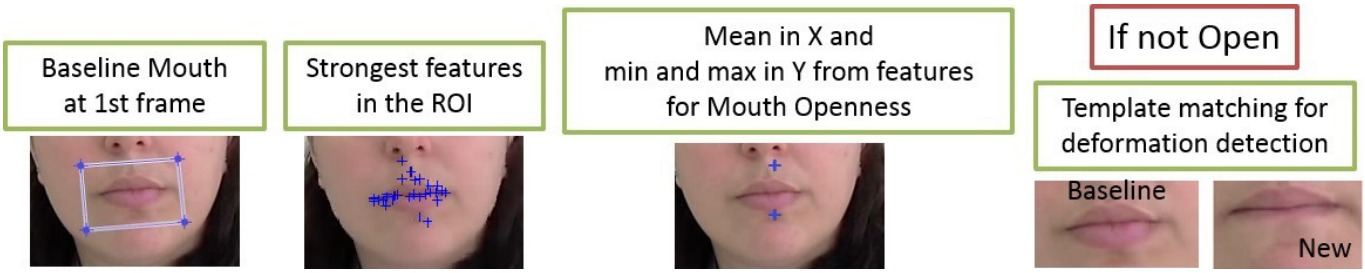


Fig. 4: Visual illustration of the image processing pipeline

box, which is homocentric to the previous ROI, defined by $P1$ and $P2$ as follows:

$$P_1 \left\langle X_1 = max\left[(x - \frac{m}{2}), 1\right], Y_1 = max\left[(y - \frac{n}{2}), 1\right]\right\rangle \quad (2)$$

$$P_2 \left\langle X_2 = min\left[(x + \frac{m}{2}), mf\right], Y_2 = min\left[(y + \frac{n}{2}), nf\right]\right\rangle \quad (3)$$

Where (x, y) are the upper left corner coordinates, m and n the dimensions of the template, and mf and nf the dimensions of the new frame. The box selected as the match, will be the one with the highest r correlation coefficient:

$$r = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{(\sum_m \sum_n (A_{mn} - \bar{A})^2)(\sum_m \sum_n (B_{mn} - \bar{B})^2))}} \quad (4)$$

where A and B the compared images, $\bar{A} = mean(A)$ and $\bar{B} = mean(B)$.

One additional constraint is the fact that r should be higher than 0.95. If this condition is not satisfied the match is not considered reliable; only in this case the V&J based mouth detector is invoked. In the exception of the last case, in order to reduce the number of false positives as much as possible, the face is first detected, and then the mouth is searched only within the bottom-half of the face. Finally, in both cases (template-matching or V&J) the template is updated in every frame by the new mouth ROI.

Feature points previously selected are then tracked within the new mouth ROI, with the Kanade-Tomasi-Lucas (KLT) tracker described in [19]. Subsequently two new points $(Q1, Q2)$ and the new mouth openness (NMO) are computed in accordance to (1). The mouth is considered "OPEN" if the following condition is satisfied:

$$NMO > BMO * OC \quad (5)$$

OC is the opening coefficient, namely the mouth expansion during opening, expressed as a percentage of the baseline mouth; it is set as a constant to the optimal value which was chosen experimentally. If (5) is satisfied then the mouth status is set to "OPEN", and processing moves on to the next frame. If the mouth is not "OPEN" then the correlation (R) between the baseline and the new mouth ROIs is computed as in (4). If R is below a threshold RT then the mouth status is set to "DEFORMED"; otherwise (if it is above RT) it is set to "CLOSED". The general idea of the algorithm is demonstrated in Fig. 1 and Fig. 4. As a final adjustment, in order to remove any potential outliers; a minimum time was set, during which any change on the mouth status is not expected and thus disregarded.

TABLE I: Mouth detection/tracking accuracy

True Positives	6361
True Negatives	0
False Positives	55
False Negatives	49
Total Frames	6465
Accuracy	**98.57%**

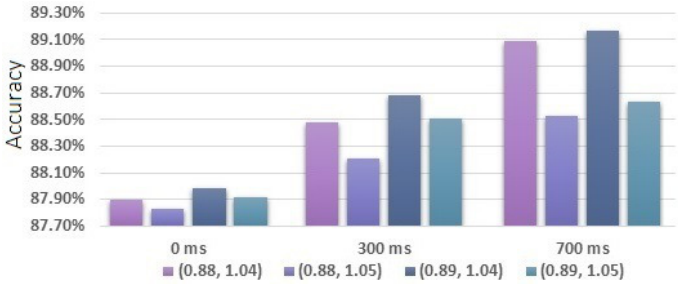


Fig. 5: Comparison of results before applying the constraint and after, for the best combinations of (RT, OC)

IV. Algorithm Evaluation

For the purpose of evaluating the present algorithm an adhoc dataset was created, consisting of 25 subjects, Caucasian, both males and females, aged from 24 to 50, with different

TABLE II: Confusion Matrix for (RT=0.89, OC=1.04), window **not** applied, accuracy 87.99%

	CLOSED	OPEN	DEFORMED
CLOSED	1200	29	127
OPEN	42	1772	96
DEFORMED	22	394	2228

TABLE III: Confusion Matrix for (RT=0.89, OC=1.04), window 700 msec, accuracy 89.17%

	CLOSED	OPEN	DEFORMED
CLOSED	1244	21	91
OPEN	62	1768	80
DEFORMED	0	386	2258

TABLE IV: Recall measure per class for (RT=0.89, OC=1.04) before and after windowing

	0 msec	700 msec	Difference
CLOSED	89.75%	91.74%	1.99% ↑
OPEN	88.92%	92.57%	3.65% ↑
DEFORMED	82.63%	85.40%	2.77% ↑

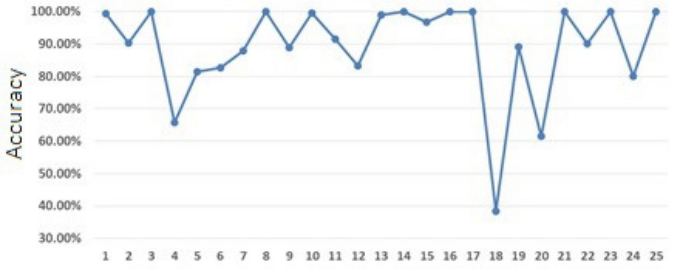
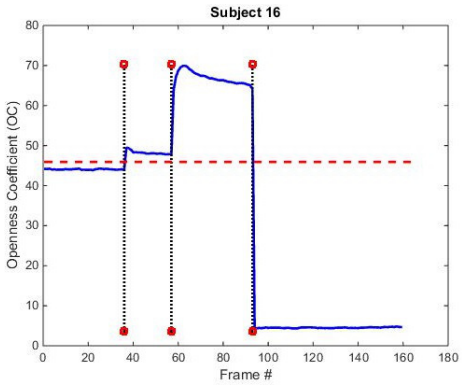
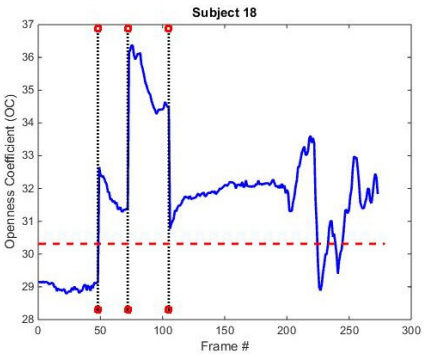


Fig. 6: Classification accuracy for all subjects

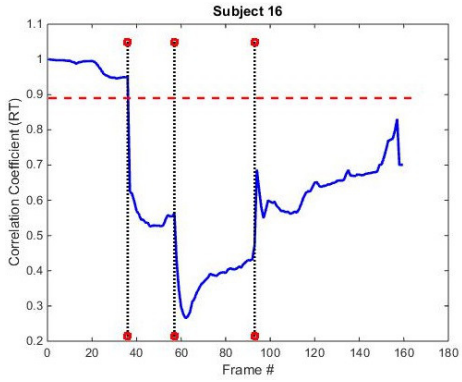


(a)

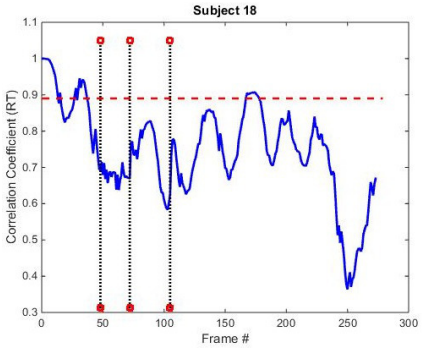


(b)

Fig. 7: Openness coefficient (OC) over time. (a) Subject 16 with 100% classification accuracy, and (b) subject 18 with classification accuracy 38.46%



(a)



(b)

Fig. 8: Correlation threshold (RT) over time. (a) Subject 16 with 100% classification accuracy, and (b) subject 18 with classification accuracy 38.46%

facial features (mustache, beard, glasses), who posed along prompt {closed mouth, half-open, full-open, deform} (for the testing half-open and full-open were considered both as a unified class: open). The video was captured at a resolution of 640x480 pixels, with a frame rate of 15 frames per second. Dataset requirements were intentionally set low with the purpose of testing the algorithm at minimum specifications.

The methods for detecting and tracking the mouth ROI had to be evaluated separately so that any false status classification can be attributed exclusively to the proposed algorithm and not to false ROI. Detection and/or tracking results are shown in Table I.

A subsequent series of experiments took place in order to establish the optimal values of OC and RT that gave the best classification accuracy. During these experiments the presence of outliers was observed. In order to eliminate these outliers a minimum time window was set, during which no changes were expected to occur, and thus were disregarded. At the second series of experiments not all combinations of OC and RT were tested; analysis focused on those with the optimal combination of the first series. As it becomes apparent in Fig. 5, the best classification result is 89.17% for a window of 700 ms, openness (OC) threshold set to 1.04 (4% expansion of the baseline), and correlation (RT) threshold 0.89 to the baseline (matching above 89%). Comparing the

TABLE V: Evaluation of the proposed algorithm for stress/anxiety assessment

			Stressed/Anxious		Relaxed	
Window (msec)	**OC**	**Detection sensitivity %**	**Normalized Openings per minute**	**Average Openness Intensity %**	**Normalized Openings per minute**	**Average Openness Intensity %**
300	1.040	87.97	1.477 ↑	7.57 ↑	0.565 ↓	5.08 ↓
400	1.040	87.97	1.350 ↑	7.72 ↑	0.565 ↓	5.09 ↓
700	1.040	83.97	1.055 ↑	7.03 ↑	0.411 ↓	5.62 ↓
500	1.020	87.97	1.478 ↑	4.85 ↑	0.616 ↓	4.64 ↓
500	1.035	87.97	1.182 ↑	6.76 ↑	0.565 ↓	5.06 ↓
500	1.040	87.97	1.182 ↑	7.92 ↑	0.514 ↓	4.96 ↓
500	1.045	83.97	1.013 ↑	7.70 ↑	0.360 ↓	5.38 ↓
500	1.050	83.97	1.056 ↑	8.12 ↑	0.308 ↓	5.48 ↓
500	1.055	80.03	0.971 ↑	8.26 ↑	0.359 ↓	5.62 ↓

two approaches, before and after applying the window, an improvement of 1.18% was achieved with the window of 700 ms. Comparison of the confusion matrices in Tables II and III before and after the windowing respectively, shows a reduction of about 70 false positives. Recall measures in Table IV also show an average improvement of 2.8% on true positive accuracy.

Experimental results, for the optimal combination of the parameters mentioned before, separately for each subject show that the majority of the subjects were classified correctly (see Fig. 6), with only three out of the total twenty-five being below 70%, and almost 1/3 at 100%.

The openness coefficient OC is visualized in Fig. 7; in 7a for subject 16 whose mouth activity was classified 100% correctly, and in 7b for subject 18 whose mouth activity was classified with the worst overall accuracy of 38.46%. The solid (blue) line corresponds to OC signal, vertical dotted (black) lines starting and ending with (red) circles to the transition between classes, and the dashed (red) horizontal line to the openness threshold which is 4% higher than the OC of closed mouth (baseline). For the first case OC for classes {half-open, full-open} lies above the threshold as it is intended, while the other two classes {closed, deformed} lie below. However the case for the other subject is not the same, as deformed is also above the threshold and this misattributes open status and does not continue to compute the correlation. By watching closely the videos, in an effort to clarify this misclassification, it was observed that subject 18 was constantly moving left to right and the opposite, as they were rotating on the swivel chair. Evidently translations appear to have an impact on the result.

The correlation threshold (RT) is visualized in Fig. 8, again in 8a for subject 16 (100% classification accuracy) and in 8b for subject 18 (38.46% classification accuracy). As before, for the first case, {closed} lies above the threshold, and everything else is below; although {half-open, full-open} are also below threshold they were already excluded during the previous step, as they were detected to be open, and thus not examined for correlation, which leaves only deformed input to be correctly classified as deformed. Still for subject 18 it seems that again the movement influences the result, as in some frames even the closed mouth period is below the correlation threshold, and a small part of deformed period is above.

Finally, for both coefficients it can be noted that for subject 16 the signal is smooth, while for subject 18 it is very unstable.

V. Preliminary Evaluation on Stress/Anxiety Assessment

Once the algorithm was evaluated in terms of detection performance, the next step was to evaluate it in terms of stress/anxiety assessment. In this section the dataset that was used for this evaluation is described, along with the experimental results.

A. Stress/Anxiety Dataset

The dataset used for evaluating the performance of the proposed algorithm for stress/anxiety assessment was created for the needs of the the FP7 Specific Targeted Research Project SEMEOTICONS (SEMEiotic Oriented Technology for Individuals CardiOmetabolic risk self-assessmeNt and Selfmonitoring) [3] [20]. The dataset consists of 23 volunteer participants (69.6% male), who were recorded while being presented with stimuli indented to elicit accordingly a.) stress/anxiety and b.) relaxation. The elicitation was realized with respect to ethical constrains, thus the effort was to produce the feelings without being intense and causing extreme stress. Categorization of each individual recordings (stress/anxious versus relaxed) was based both on the participants self-reports, as well as on annotations made by three clinicians for each video recording. The categorization resulted in 12 recordings for the stress/anxious class, and 10 for the relaxed. The video was captured at a resolution of 526x696 pixels, with a frame rate of 50 frames per second.

B. Experimental Results

In order to evaluate the proposed algorithm in stress/anxiety assessment two features were extracted based on the mouth openness in comparison to the baseline; the number of total openings through each video recording normalized per minute, as well as the average openness intensity, corresponding to the mouth expansion. The tests took place for different sets of values for the window size and the OC parameters. The significance of the extracted features is visible in Table V, where a general trend for the stressed/anxious participants to show a higher number of openings, as well as greater openness intensity is shown. These finding are consistent with

the literature, which supports that mouth opening [8] and lip parting [9] are increased under stress.

VI. Discussion

The present work proposed a semi-automatic system for classifying mouth actions {CLOSED, OPEN, DEFORMED}, with an achieved accuracy of up to 89.17%. The improvement of 1.18%, by using windows, brings a delay of the window size (700 ms), as processing starts only after the specified number of frames is reached; this introduces a trade-off between classification accuracy and real time execution. The preliminary evaluation in terms of stress/anxiety assessment presented features with significant difference between the two classes, something which motivates further exploration. The steady high performance of the algorithm over different datasets, both in respect of the acquisition protocol, as well as recording specifications, and with slight modification of the parameters, suggests that these features are highly robust. Although the results are quite satisfactory, further investigation is required in order to address the current limitations, such as the need for a manually selected baseline, and the impact of movements. Finally, the use of a benchmark dataset which includes mouth gestures, such as the EURECOM Kinect Face Dataset [21], could enable a comparison with the state-of-the-art. To the best of the authors' knowledge, however, the specific dataset has been mostly used in the area of face recognition, and as of yet not to evaluate applications similar to the one proposed herein. Approaches in the area of automatic stress and anxiety assessment employ ad-hoc datasets, this does not enable direct comparisons between the methods.

Acknowledgment

Pampouchidou Anastasia was funded by the "Maria Zaousi" scholarship from the Greek State Scholarships Foundation (I.K.Y). Additionally, part of this work was performed in the framework of the FP7 Specific Targeted Research Project SEMEOTICONS, partially funded by the European Commission under Grant Agreement 611516.

References

[1] P. Ekman, W. Friesen, and J. Hager, *Facial Action Coding System: The Manual*. Research Nexus of Network Information Research Corporation, 2002.

[2] A. Pampouchidou, K. Marias, M. Tsiknakis, P. Simos, F. Yang, and F. Meriaudeau, "Designing a framework for assisting depression severity assessment from facial image analysis," in *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. IEEE, 2015, pp. 578–583.

[3] F. Chiarugi, G. Iatraki, E. Christinaki, D. Manousos, G. Giannakakis, M. Pediaditis, A. Pampouchidou, K. Marias, and M. Tsiknakis, "Facial signs and psycho-physical status estimation for well-being assessment," in *Proceedings of the 7th International Conference on Health Informatics (HEALTHINF)*. SCITEPRESS, 2014, pp. 555–562.

[4] H. Ellgring, *Non-verbal communication in depression*. Cambridge University Press, 2007.

[5] T. Sundelin, M. Lekander, G. Kecklund, E. J. Van Someren, A. Olsson, and J. Axelsson, "Cues of fatigue: effects of sleep deprivation on facial appearance," *Sleep*, vol. 36, no. 9, p. 1355, 2013.

[6] D. Metaxas, S. Venkataraman, and C. Vogler, "Image-based stress recognition using a model-based dynamic face tracking system," in *Computational Science-ICCS 2004*. Springer, 2004, pp. 813–821.

[7] D. F. Dinges, R. L. Rider, J. Dorrian, E. L. McGlinchey, N. L. Rogers, Z. Cizman, S. K. Goldenstein, C. Vogler, S. Venkataraman, and D. N. Metaxas, "Optical computer recognition of facial expressions associated with stress induced by performance demands," *Aviation, space, and environmental medicine*, vol. 76, no. Supplement 1, pp. B172–B182, 2005.

[8] W. Liao, W. Zhang, Z. Zhu, and Q. Ji, "A real-time human stress monitoring system using dynamic bayesian network," in *IEEE computer society conference on Computer vision and pattern recognition-workshops*. IEEE, 2005, pp. 70–70.

[9] H. G. Wallbott and K. R. Scherer, "Stress specificities: Differential effects of coping style, gender, and type of stressor on autonomic arousal, facial expression, and subjective feeling." *Journal of Personality and Social Psychology*, vol. 61, no. 1, p. 147, 1991.

[10] N. Sharma and T. Gedeon, "Objective measures, sensors and computational techniques for stress recognition and classification: A survey," *Computer methods and programs in biomedicine*, vol. 108, no. 3, pp. 1287–1301, 2012.

[11] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (CERT)," in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*. IEEE, 2011, pp. 298–305.

[12] S. Stillittano, V. Girondel, and A. Caplier, "Lip contour segmentation and tracking compliant with lip-reading application constraints," *Machine vision and applications*, vol. 24, no. 1, pp. 1–18, 2013.

[13] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[14] S. Usman and J.-L. Dugelay, "Combining edge detection and region segmentation for lip contour extraction," in *Articulated Motion and Deformable Objects*. Springer, 2010, pp. 11–20.

[15] J. M. Girard, J. F. Cohn, and F. De la Torre, "Estimating smile intensity: A better way," *Pattern Recognition Letters*, 2014.

[16] H. Hojo and N. Hamada, "Mouth motion analysis with space-time interest points," in *TENCON 2009-2009 IEEE Region 10 Conference*. IEEE, 2009, pp. 1–6.

[17] S.-L. Wang and A. W.-C. Liew, "Physiological and behavioral lip biometrics: A comprehensive study of their discriminative power," *Pattern Recognition*, vol. 45, no. 9, pp. 3328–3335, 2012.

[18] J. Shi and C. Tomasi, "Good features to track," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'94)*. IEEE, 1994, pp. 593–600.

[19] C. Tomasi and T. Kanade, *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ. Pittsburgh, 1991.

[20] M. Pediaditis, G. Giannakakis, F. Chiarugi, D. Manousos, A. Pampouchidou, E. Christinaki, G. Iatraki, E. Kazantzaki, P. Simos, K. Marias *et al.*, "Extraction of facial features as indicators of stress and anxiety," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 3711–3714.

[21] R. Min, N. Kose, and J.-L. Dugelay, "Kinectfacedb: A kinect database for face recognition," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 11, pp. 1534–1548, Nov 2014.