Towards Hand-over-Face Gesture Detection

Gábor Révy, Dániel Hadházi, Gábor Hullám
Department of Measurement and Information Systems
Budapest University of Technology and Economics
Budapest, Hungary
Email: revy.gabor@edu.bme.hu, {hadhazi, hullam.gabor}@mit.bme.hu

***Abstract*—Facial microexpressions are immediately appearing reactions on the face that indicate various details about people's mental and emotional states. Their most important property is that their interpretation is identical or very similar for people all over the world. At present, their identification requires a psychologist expert. Thus automating this task would enable a broader application.**

The goal of this research is the detection of microexpressions using hybrid expert algorithms. Our algorithms mainly rely on landmark point detectors. Based on their output, several expert algorithms are utilized to extract key features and changes appearing on the face of a subject. These algorithms usually include several steps of image processing and time series analysis algorithms.

In this paper, a component responsible for detecting hand gestures and hand pose is introduced. This component helps other algorithms to eliminate false positive detections by detecting the hands over the face. In addition, the recognizability of hand-over-face gestures is investigated. Finally, the implemented face occlusion detector method is evaluated on videos.

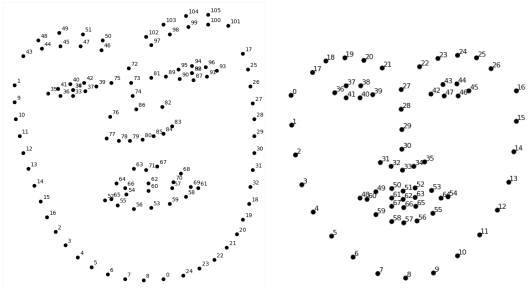
***Index Terms*—microexpression, image processing, landmark points, expert system, facial expressions, hand-over-face gestures**

I. Introduction

Microexpressions are the visible features of emotions appearing on the face for a very short time, e.g. an involuntary reaction to a question. Automating the detection of facial expressions would allow a wide range of uses, e.g. to study reactions to an advertisement or to assist in the diagnosis of mental disorders. Experts usually distinguish 7 different basic emotions: anger, disgust, fear, happiness, sadness, surprise and contempt. In order to categorize reactions in videos, proper detection of microexpressions is required.

The first step is to detect the motion of the muscles, the so-called action units on the face. These parts of the face are described in detail in the FACS system [1]. Emotions can be determined based on the activated action units.

Previously, we have developed several algorithms to identify these muscle movements appearing on the face. These methods operate on videos, since our goal is to detect the change in the facial features. Our algorithms utilize a facial landmark point detector to localize key points in the face. We combine two facial landmark detectors to make the localization more accurate and robust. One is the neural network-based PFLD [2] and the other, that can be found in the Dlib [3] library is utilizing an ensemble of regression trees. These identify 106 and 68 key points on the face, respectively, as shown in Figure 1. Using the landmark detectors on videos also allows for smoothing between the frames. This is useful because the output of the detectors often oscillates, it is not accurate, and some objects can even occlude the face.



(a) PFLD [2] (b) Dlib [3]

Fig. 1: Facial landmark points identified by the facial landmark detectors utilized.

Furthermore, the fact whether the hand is occluding the face, can be an additional source of information to determine emotions [4]. This already has a meaning in itself, but it can also help other image processing algorithms to indicate that there may be anomalies or outliers due to this "noise". Some algorithms already exist to detect hand-over-face (HOF) gestures. Mahmoud et al. utilized histograms of oriented gradients (HOG) and local space-time features [5], [6] to extract features and an SVM (Support Vector Machine) for classification to detect HOF gestures. In addition, Mahmoud et al. [7] compared the use of local binary patterns and Gabor filters in detecting face occlusions. Ghanem et al. constructed a neural network called MPSPNet [8] to segment the hands over the face.

In this paper a simple, but surprisingly accurate approach is proposed for hand over face detection. Furthermore, the pose of the hand may reveal additional information about the mental state of the person in the video [4]. This area however, requires further investigation and it is out of scope of this paper. Here we only focus on the recognizability of different hand poses.

This research was supported by the ÚNKP-21-5-BME-362 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund, and the János Bolyai research scholarship.

Figure 2 provides an overview of the tasks described in this paper.

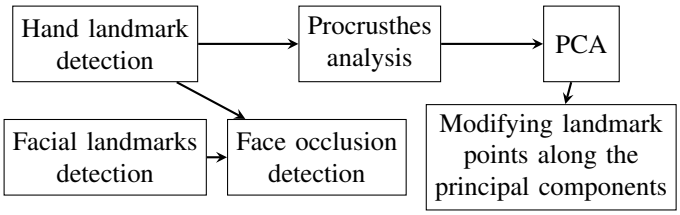


Fig. 2: Steps of the tasks related to hand-over-face gesture detection.

II. Hand over face detection

There are several tasks related to hand-over-face detection (HOF) including face occlusion detection and hand pose detection. After the hand detection, it can be informative for other algorithms to know, whether a specific part of the face is occluded. This is because covering the face acts as noise in algorithms that focus on different areas of the face. In addition, it also carries information about emotions, i.e. the covered area of the face can indicate certain types of emotions. Detecting hand pose along with the position could reveal even more information about emotions.

To detect hands and hand landmarks the MediaPipe [9] library was utilized. MediaPipe is an open source cross-platform library providing customizable machine learning solutions for visual media (i.e. images and videos). It contains several landmark and object detection algorithms as well as segmentation algorithms. MediaPipe Hands is a hand and hand landmark detector. It detects the ROI (region of interest) and 21 3D key points of the hands in an RGB image (see Figure 3). The handedness of the hand ROI is also determined.

Fig. 3: 21 landmark points (and their skeleton) detected by the MediaPipe [9] Hands detector.

A. Face occlusion detection

To detect the occlusion of different areas of the face, the face is divided based on its landmark points. 3 areas were selected on the face, that are separated by the bottom of the nose and the upper line of the eyes. The occlusion is simply determined by checking if a hand landmark point falls into such an area. Examples of the output of the face occlusion algorithm are shown in Figure 4. Note, that depending on the analyzed microexpressions, additional facial regions of interest could be defined.

Fig. 4: The output of the face occlusion detection: the occluded part is marked with red. The source of the original images: BAUM-1s dataset [10]

Using this algorithm jointly with other feature detectors allows the removal of several false positive detections. An example of this is shown in Figure 5: the lip compression detector algorithm made a false positive detection, but using the HOF detector, this can be removed automatically.

B. Towards hand gesture recognition

The detectability of hand gestures was also investigated. A series of hand gestures were recorded along with their landmark points. In order to eliminate the effect of the scaling, rotation and translation of the landmark points and investigate only the relevant information of the shape, the coordinates of the landmark points must be normalized appropriately. This can be performed by the Procrustes superimposition [11]. The Procrustes superimposition algorithm is usually applied before comparing shapes. It minimizes the so-called Procrustes distance between two shapes by scaling, translating

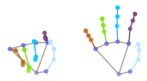
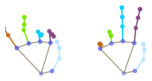
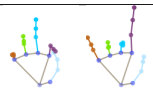
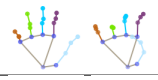
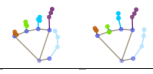
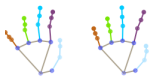
Fig. 5: The lip compression detector made the mistake of a false positive detection, but using the face occlusion detector, it can be filtered out. Source: BAUM dataset [10]

and rotating them. In our case the Procrustes superimposition algorithm minimized the Euclidean distance between the palm landmark points of the recorded hand and a simple hand model. After the uniformization, PCA (Principal Component Analysis) was applied on the coordinates of the landmark points. The motivation of the utilization of PCA is to explore and coordinate the manifold of the landmarks in order to extract features, which can efficiently describe the pose of the hands. Note that PCA extracts features along which the deviation of the samples is the highest. Therefore, we can adapt these features to our goal by expanding the dataset with specific samples.

The most relevant features (principal components) were selected to explain 95% of the variance in the data, i.e. the first 8 components. After projecting to these components and modifying their values, the reconstructed landmark points were investigated. For each principal component, Table I shows the changes observed when examining the restored landmarks of the hand.

As Table I shows, the most relevant principal components can be interpreted semantically fairly well. The detection of the hand gestures is currently in progress. It could be utilized later to look for signs of different mental states. One such sign is nail biting, which is an indicator of stress. Other detectable states include thinking, evaluation, skepticism, boredom, happiness and deception [12]. A further option is to take multiple aspects of hand features into consideration such as hand shape and hand action [4]. Hand shapes range from closed hand through holding out one or more fingers to a fully open hand, while hand action includes holding (the chin), leaning (the face on the hand), tapping and touching (parts of the face). Using these "dimensions" as descriptors, a *thinking* state could be characterized with using the index finger (or other fingers) to touch the chin or the forehead, whereas a *bored* state would be better characterized by either leaning the face on an open or closed hand or holding the chin. The challenge with this approach however is that there are no properly annotated videos for hand gesture detection. Existing datasets either consist of only still images or they lack the necessary annotation that is detailed enough to be of any use. The BAUM dataset [10] could be utilized for hand gesture detection, however this requires additional annotation

TABLE I: The first 8 principal components explaining 95% of the variance in the recorded hand landmark data.

	Description	**Cumulative explained variance**	**Plot**
1	openness of fingers 2-5	0.5	
2	slant of the openness of fingers 2-5	0.6676	
3	openness of the 2nd and 5th fingers with respect to the 3rd and 4th fingers	0.7714	
4	openness of the thumb	0.8415	
5	curvature of the thumb	0.9003	
6	spreadedness of the fingers	0.9225	
7	distance between the fingers	0.9393	
8	lateral position of the 3rd and 4th fingers	0.954	

and further investigation of applicable expert algorithms.

III. Evaluation of hand occlusion detection

To evaluate the hand-over-face (HOF) detection we were looking for annotated datasets specific to this purpose. However, we only found a dataset containing images labeled with hand segmentation. Unfortunately, this is not fit for purpose, as the algorithm that detects the facial landmark points can only be run on videos to smooth the detection between the frames. This is useful, because if the face is occluded by the hand, the landmark points are not detectable by default, but can be determined from the adjacent frames. Thus, the detection of the HOF gestures was evaluated on videos selected from the same BAUM [10] dataset, which we used previously to evaluate other microexpression detection algorithms. A total of 50 videos were selected in which the hand covers the face. The videos contained 79 HOF gestures from which 76 (96.2%) was detected and there were 3 false positive detections. When examining the results, the following could be observed:

- The detector performs well if a large part of the hand is visible and the hand is not too blurred. If the frame rate is high enough, the image will not become blurry even during rapid hand movements.
- In a very small number of cases, the hand detector generates a false detection of hand. In those cases, where the illusionary hand landmark points do not intersect the area covered by facial landmark points, this is not a problem,

as it does not cause a false occlusion (see Figure 6a). In cases where there is an intersection however, this causes false detection (as shown in Figure 6b).

- If a big part of the hand is not visible or the hand is in an uncommon position (or at least a position difficult for the detector to detect), the hand landmark points become inaccurate. This can lead to false detections (as shown in Figure 6c).
- If the hand moves too fast or the frame rate is too low compared to the speed of the hand movement, the image of the hand becomes blurry and the hand is not detected. This can lead to false negative detections as shown in Figure 6d.
- If only a small part of the hand is occluding the face, no occlusion may be detected as can be seen in Figure 6e. This is caused by the nature of the HOF detection algorithm. This could be addressed by creating a hand segmentation based on the hand landmark points and examining the cross section of the hand and the face mask.

(a) Fake hand landmark detections.

(b) False positive detection caused by a fake hand detection.

(c) False positive detection caused by the inaccuracy of the hand landmark detector.

(d) False negative detection caused by a false negative hand detection.

(e) False negative detection caused by the nature of the hand over face detection algorithm.

Fig. 6: Examples of difficult hand over face detection cases. Source: BAUM dataset [10]

IV. Conclusion

In this paper, we introduced a component to detect the pose of the hand over the face. The utilization of this algorithm eliminated several false positive detections by notifying other algorithms that their ROI is covered by the hands. We also investigated the recognizability of the hand-over-face gestures and have shown, that the principal components defined by applying PCA can be interpreted semantically. Finally, we evaluated the face occlusion detector method on videos from the BAUM [10] dataset, which is specialized for emotions. Results have shown, that despite of being a simple approach, the face occlusion detector works quite accurately. This however requires that the two machine learning-based algorithms, (i.e. the face and the hand landmark detectors) we are building on, work correctly. In the future we plan to extend the algorithm with additional features such as taking hand actions into consideration.

References

[1] P. Ekman, J. C. Hager, and W. V. Friesen, *Facial action coding system: the manual*. Research Nexus, 2002.

[2] X. Guo, S. Li, J. Yu, J. Zhang, J. Ma, L. Ma, W. Liu, and H. Ling, "Pfld: A practical facial landmark detector," *arXiv e-prints*, pp. arXiv–1902, 2019.

[3] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[4] M. Mahmoud and P. Robinson, "Interpreting hand-over-face gestures," in *International Conference on Affective Computing and Intelligent Interaction*, pp. 248–255, Springer, 2011.

[5] I. Laptev, "On space-time interest points," *International journal of computer vision*, vol. 64, no. 2, pp. 107–123, 2005.

[6] M. Mahmoud, T. Baltrušaitis, and P. Robinson, "Automatic analysis of naturalistic hand-over-face gestures," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 6, no. 2, pp. 1–18, 2016.

[7] M. Mahmoud, R. El-Kaliouby, and A. Goneid, "Towards communicative face occlusions: machine detection of hand-over-face gestures," in *International Conference Image Analysis and Recognition*, pp. 481–490, Springer, 2009.

[8] S. Ghanem, A. Dillhoff, A. Imran, and V. Athitsos, "Hand over face segmentation using mpspnet," in *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pp. 1–8, 2020.

[9] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv e-prints*, pp. arXiv–1906, 2019.

[10] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "Baum-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 300–313, 2016.

[11] M. Rudemo, "Statistical shape analysis. I. L. Dryden and K. V. Mardia, Wiley, Chichester 1998. no. of pages: xvii+347. ISBN 0-471-95816-6," *Statistics in Medicine*, vol. 19, no. 19, pp. 2716–2717, 2000.

[12] A. Pease and B. Pease, "The definitive book of body language," 2006.