Machine Vision and Applications (2016) 27:1063–1070
DOI 10.1007/s00138-016-0794-2

ORIGINAL PAPER

Pose-invariant descriptor for facial emotion recognition

Seyedehsamaneh Shojaeilangari[1,3] · Wei-Yun Yau[2] · Eam-Khwang Teoh[1]

Received: 18 May 2015 / Revised: 17 May 2016 / Accepted: 25 June 2016 / Published online: 19 July 2016

Abstract Most facial emotion recognition algorithms assume that the face is near frontal and the face pose fixed during the recognition process. However, such constrain limits the adoption for real-world applications. To solve this, pose-invariant descriptor for emotion recognition is required. This work proposes a novel pose-invariant dynamic descriptor that encodes the relative movement information of facial landmarks. The proposed feature set is able to handle speed variations and continuous head pose variations, while the subject is expressing an emotion. In addition, the proposed method is fast and thus can be realize real-time implementation for real-world application. Performance evaluation done using three publicly available databases; Cohn-Kanade (CK$^+$), Amsterdam Dynamic Facial Expression Set (ADFES), and Audio Visual Emotion Challenge (AVEC 2011) showed that our proposed method outperforms the state-of-the-art methods.

Keywords Pose-invariant · Spatio-temporal descriptor · Emotion recognition · Facial expression

1 Introduction

Facial emotion recognition under natural conditions has a wide range of potential applications such as human-computer interaction, gaming, behavioral study and monitoring of disorder condition such as autism. Despite significant progress in this field, there are still challenges related to real-world application in unconstrained situations. Most of the previous research works only focus on the frontal or near frontal face images and assume that the frontal face can be captured. Relying only on the frontal faces limits the continuity of the emotion recognition process. In addition, processing the static images alone for emotion analysis may fail in real-world application whereby the expressions are mostly subtle and spontaneous. Also study has shown that human visual system is able to detect an expression more accurately when its temporal information is taken into account [1]. Furthermore, in real-world applications, frontal or near frontal view of facial images may not be available all the time as the face moves. Thus it is necessary to solve emotion recognition under multiple face poses.

This paper focuses on facial emotion recognition challenge from videos containing arbitrary view. We propose a novel dynamic representation which encodes the relative movement of facial landmarks such that the representation is invariant to the face's pose. The proposed approach also does not require any face alignment. The basic idea is motivated by the fact that facial landmarks vary symmetrically, and there is consistent pattern with respect to the mid-point of the face. Thus, we first find the facial landmarks for all frames of a video sequence. Then the motion vectors of the landmarks are calculated for all subsequent frames. After extracting the relative movement descriptors, the statistics of the features is retained for further processing. The proposed approach will be further elaborated in Sect. 3, while Sect. 4 describes the comparison done to evaluate the proposed approach.

✉ Seyedehsamaneh Shojaeilangari
seyedehs1@e.ntu.edu.sg

[1] School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Ave, S1-B4C-14, Singapore 639798, Singapore

[2] Institute for Infocomm Research, A*STAR, Singapore, Singapore

[3] Present Address: Research Center for Science and Technology in Medicine, Tehran University of Medical Sciences, Emam Khomeini Hospital Complex, Tehran, Iran

2 Related works

Although facial expression recognition has been extensively studied in the past, most of the approaches [2,3] focus on frontal facial images. Small changes in the facial pose may reduce the effectiveness of these methods. Only a few researchers have attempted to solve the facial pose challenge.

A probabilistic method based on 2D geometrical features was proposed by Rudovic et al. [4] for pose-invariant facial expression. The locations of 39 landmarks were extracted from a facial image with an arbitrary head pose. The Coupled Scaled Gaussian process regression model was then applied to normalize the facial pose. In this model, the mapping between the facial landmarks for each pair of non-frontal faces and frontal one is learned independently, and then their coupling was performed to achieve the dependencies among them. This approach was shown to be better than the state-of-the-art methods for head pose normalization. Furthermore, it was also claimed to be the first work that is able to deal with -45 to $+45^\circ$ head pan and -30 to $+30^\circ$ tilt movement. Although the model was trained based on only a few discrete head poses, the method was able to deal with continuous head pose variation within the above-mentioned limits. However, the method is not applicable for dynamic process of facial emotions from videos.

Jeni et al. [5] used the facial shape information as a robust representation to pose variations. The 3D landmark positions were estimated on face images using constrained local models at the first stage. Then, the rigid transformation from the obtained 3D shape was removed. Finally, SVM classifier was applied to the projected shape onto 2D space. Although the shape information was shown to be efficient for facial expression detection, the pose-invariant ability of the proposed system was not evaluated in more challenging scenarios where the head pose variation occurs while expressing the emotion.

Songfan and Bhanu [6] introduced a homogeneous reference face model called avatar reference to capture the nature of the whole dataset. Then, a video sequence with any length was condensed into a single image representation. This representation is able to aggregate the temporal facial emotion information and also compensate for large rigid head motion as well as removing the person-specific information. However this approach requires proper alignment of the face to the reference avatar model which itself is a source of error.

A novel representation approach for facial images using the regional covariance matrix was proposed by Zheng et al. [7]. A dimensionality reduction step is then applied to the resulting features based on the theory of discriminant analysis. An effective approach was further proposed to find the optimal discriminant vectors. The key advantage of this method is that it does not need any facial alignment and feature point localization, which are both challenging tasks. However, this method is only applicable for static analysis of facial expression from images.

A simple method named variable-intensity template was proposed by Kumano et al. [8] to obtain a person-specific model for describing various facial expressions. The variable intensity templates define how the intensity of multiple facial points varies for an observed expression. The method is able to detect facial expression and estimate the pose simultaneously within the framework of a particle filter. Simple modeling and low computational cost are the advantages of this method. However, the method is quite sensitive to errors in interest point localization and misalignment. Additionally, the method was not evaluated on public databases to check its performance.

3 Proposed methodology

In this section, we present our proposed framework for view invariant facial feature extraction and emotion recognition from video sequences. We noticed that in expressing an expression, the facial landmarks move consistently with respect to a stable point of the face. For example, when smiling, the mouth's corners curve inward with respect to the tip of the nose. Similarly, when frowning, they curve outward. Such observations are used as the basis to develop pose-invariant features to encode facial expression. The method comprises three main steps; (1) facial point's localization and tracking, (2) pose-invariant feature extraction, (3) feature encoding, and (4) emotion classification.

3.1 Landmark localization and tracking

We used the algorithm[1] proposed in [9] to locate and track the facial landmarks. The procedure is based on a method named Supervised Descent Method (SDM) to optimize a nonlinear least squares (NLS) function. In training, the SDM learns a sequence of descent directions that minimizes the mean of NLS functions sampled at different points and during testing, SDM minimizes the NLS objective function using the learned descent directions without computing the Jacobian and Hessian. In other words, SDM learns the generic descent directions in a supervised manner which makes it able to overcome many drawbacks of second-order optimization approaches like non-differentiability and expensive computation of the Jacobians and Hessians.

A total of 42 facial landmarks are used as shown in Fig. 1. To describe the relative movement of these points within the face, we require a stable reference point that does not move or change when an emotion is expressed. For this purpose,

[1] The source code for nose point detection is available at: http://humansensing.cs.cmu.edu/intraface/download.html.

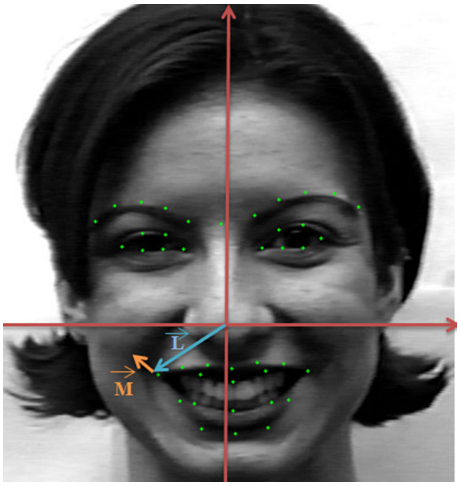

Fig. 1 Face image showing the facial landmarks, reference point and relative features extracted from the reference point

we choose the point at the tip of the nose. This point is also used to estimate the head movement. Subsequently this is subtracted from the motion vector of each landmark point to obtain the effective net movement due to emotion alone. This is necessary to reduce the effect of head motion. Furthermore, face alignment is not required. These points are continuously tracked throughout the video sequence.

3.2 Pose-invariant feature extraction

After localizing the facial landmarks at every frame, the movement of each point can be tracked for any successive frames and denoted as the motion vector. Let $M(P; t_k)$ represent the motion vector $(u; v)$ at landmark location $P = (x; y)$ at time t_k. Then the motion components are defined as follows:

$$u\,(P, t_k) = x\,(P, t_k) - x\,(P, t_{k+1}) \tag{1}$$

$$v\,(P, t_k) = y\,(P, t_k) - y\,(P, t_{k+1}) \tag{2}$$

where $x\,(P, t_k)$ and $y\,(P, t_k)$ are the horizontal and vertical position of each landmark P at time t_k, respectively.

Two pose-invariant feature sets are proposed for encoding the relative motion information of facial points.

$$f^1\left(P_{ij}, t_k\right) = \frac{\nabla u\left(P_{ij}, t_k\right)}{\nabla x\left(P_{ij}, t_k\right)} + \frac{\nabla v\left(P_{ij}, t_k\right)}{\nabla y\left(P_{ij}, t_k\right)} \tag{3}$$

$$f^2\left(P_{ij}, t_k\right) = \frac{\nabla v\left(P_{ij}, t_k\right)}{\nabla x\left(P_{ij}, t_k\right)} - \frac{\nabla u\left(P_{ij}, t_k\right)}{\nabla y\left(P_{ij}, t_k\right)} \tag{4}$$

where $\nabla u\left(P_{ij}, t_k\right)$, $\nabla v\left(P_{ij}, t_k\right)$, $\nabla x\left(P_{ij}, t_k\right)$, and ∇y $\left(P_{ij}, t_k\right)$ are defined as follows:

$$\begin{aligned}
\nabla u\left(P_{ij}, t_k\right) &= u\,(P_i, t_k) - u\left(P_j, t_k\right)\\
\nabla v\left(P_{ij}, t_k\right) &= v\,(P_i, t_k) - v\left(P_j, t_k\right)\\
\nabla x\left(P_{ij}, t_k\right) &= x\,(P_i, t_k) - x\left(P_j, t_k\right)\\
\nabla y\left(P_{ij}, t_k\right) &= y\,(P_i, t_k) - y\left(P_j, t_k\right)\\
&\text{for}\quad i, j = 1 : N, j \neq i
\end{aligned} \tag{5}$$

and N is the total number of landmarks used.

The first descriptor (f^1) was defined like divergence of the flow field from any two landmarks that measures the amount of local expansion or contraction of the facial landmarks.

The second descriptor (f^2) that captures the local spin around the axis perpendicular to the motion plane is like Curl feature. It is useful to measure the dynamics of the local circular motion of the facial landmarks. The proposed features f^1 and f^2 encode both the relative motion information and geometrical position of the landmarks.

Two additional pose-invariant features are defined regarding the projection and rotation information of each landmark with respect to the nose point. For this purpose, each landmark location is calculated with respect to the new coordinate system constructed using the chosen reference point as shown in Fig. 1. A sample motion vector (M) and landmark location vector (L) for lip corner are illustrated in this figure. The features are defined as follows:

$$f^3\,(P, t_k) = \overrightarrow{M} \cdot \hat{L} = u\hat{l}_x + v\hat{l}_y \tag{6}$$

$$f^4\,(P, t_k) = \hat{L} \times \overrightarrow{M} = v\hat{l}_x - u\hat{l}_y \tag{7}$$

where $\vec{M}$ is the motion vector, $\hat{L}$ is the unit position vector of the landmarks $\left(\hat{L} = \frac{\vec{L}}{\left|\vec{L}\right|} = (\hat{l}_x, \hat{l}_y)\right)$ originating from the nose point.

Indeed f^3 measures the amount of expansion or contraction of each point with respect to the nose point and f^4 captures the amount of clockwise or anti-clockwise rotation of each landmark movement with respect to the position vector.

3.3 Feature encoding

The final descriptor is obtained by accumulating the features extracted at each landmark. Two types of histograms are used to accumulate the features of each landmark in temporal domain. A Weighted Histogram (WH) is used to characterize the magnitude of emotion. It consists of two bins - positive and negative bins, and the magnitude of the associated features is used to vote for each bin. The Un-Weighted Histogram (UWH) ignores the magnitude of the emotion and attempts to characterize its dynamics. It involves three bins containing positive, negative, and zero features. Equal vote is assigned for each bin, which means that the total number

Table 1 Brief information of the databases used

Database	# Sequence	# Subject	Emotion type	Head movement	Expression type
CK^+	593	123	6 Discrete	No	Posed
ADFES	648	22	10 Discrete	Yes	Posed
AVEC	95	NA	4 Affect dimension	Yes	Natural

Table 2 Comparison of the proposed descriptor to other methods for CK$^+$ database

Method	No. feature	Detection rate	Time complexity (S)
LBP-TOP [2]	17700	89.31	5.22
LPQ-TOP [3]	76800	89.17	2.19
(STHLP + STHLO) [15]	10800	91.08	0.5
LPLO [16]	20736	91.87	0.5
HDPC [17]	10800	92.88	4.20
OF-based Pose-Invariant Descriptor [18]	8000	94.48	13.12
Landmark-based Pose-Invariant Descriptor	17220	95.76	0.08

of positive, negative, and zero features are counted. Since the magnitude of emotion is ignored in UWH, it is able to handle variations in the speed of the expression.

WH and UWH are computed for each landmark based on the four features (f^1, f^2, f^3, and, f^4) described earlier. The concatenation of both these histograms produces the required descriptor of the corresponding landmark.

Since $f^1$and f^2 are encoded the relative movement of any two landmarks, for each point we should calculate the features $f^1$and f^2 relative to 41 remaining points. It means that 41 measures for each landmark are encoded as WH and UWH across all the frames. For example, for the first landmark, the extracted measures are $\{f^1_{1,2}, f^1_{1,3}, .., f^1_{1,41}\}$ and $\{f^2_{1,2}, f^2_{1,3}, .., f^2_{1,41}\}$ for every subsequent frames. Accumulating the measures across the frames to construct WH will result in 82 (41×2) measures and 123 (41×3) for UWH. As $f^3$and f^4 are measured for each landmark relative to the nose point, each landmark is described by only one measure and then encoded with WH and UWH across the temporal domain giving a total of 5 measures. The final concatenated feature for each landmark has a total of 420 dimensions ($205 \times 2 + 5 \times 2$).

3.4 Emotion classification

For emotion classification, Support Vector Machine (SVM) with polynomial kernel function is selected. We used one-against-all technique that constructs binary SVM classifiers to categorize each emotion against all the others. Classification of a new instance is done by a winner-takes-all strategy, where the classifier with the highest output function assigns the final class. Regarding parameter selection of SVM, we carried out grid-search on the parameters as suggested in [10]. The parameters producing the best result are chosen.

4 Results and discussion

We evaluated our proposed dynamic descriptor on three publicly available databases for facial expression recognition; namely Cohn-Kanade (CK^+) [11,12] dataset, Amsterdam Dynamic Facial Expression Set (ADFES) [13] database and Audio Visual Emotion Challenge (AVEC 2011) [14] database. The brief information of the databases is given in Table 1.

4.1 CK$^+$ database

We used all 309 sequences from the dataset that have been labeled with at least one of the six basic emotions. The six prototype emotion types are joy, surprise, anger, fear, disgust, and sadness.

The first experiment was conducted to test the efficiency of the proposed dynamic descriptor. While a number of researchers have reported the performance of their facial emotion recognition algorithms on the CK$^+$ benchmark database, these results are not directly comparable due to the large differences in the experimental setup (e.g., pre-processing steps, feature extraction method, number of sequences used for training and evaluation, etc.). Therefore, to obtain a meaningful comparison of the proposed feature with other, we have evaluated some of the successful techniques reported in the literature using a common experimental setup. Our method is compared to the those successful dynamic approaches [2,3,15–17] in Table 2.

We also compared the new proposed method to our previous work on pose- invariant feature extraction which is based on Optical Flow (OF) [18]. In our OF-based pose-invariant descriptor, after computing the OF of every subsequent frame in a video, we defined four features; divergence, curl, projection (similar to f^3), and rotation (similar to f^4), and then

Table 3 Confusion matrix of the result for CK$^+$ database

Actual	Predict					
	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	97.78	0	0	0	7.14	1.20
Disgust	2.22	100	0	0	0	0
Fear	0	0	84.00	0	0	0
Happiness	0	0	8.00	100	0	1.20
Sadness	0	0	0	0	92.86	0
Surprise	0	0	8.00	0	0	97.59

all features were encoded to WH and UWH for each spatio-temporal segment of the video.

We used polynomial SVM as the classifier for all methods. To report the results, fivefold Cross Validation (CV) was used. We repeated the experiments 10 times and then average the results. The results are shown in Table 2. The time complexity reported is the execution time to extract the features from a volume data of size 100 × 100 × 10. The time for pre-processing, feature selection or landmark localization is not included. Note that we have optimized the parameters of each method via greedy search and report the best results for each method in this table. However, the segmentation of the volume data is the same for all methods (each volume data is divided into 100 blocks).

As shown, the recognition rate of our proposed descriptor is superior to other methods in terms of accuracy and execution time. It is noted that the results shown in this table are different from those reported in the original papers due to different experimental setup such as pre-processing, evaluation measurement, segmentation, number of sequences used, etc.

Table 3 shows the confusion matrix of the results obtained. The detection rate of “fear” and “sadness” are lower than the other emotion. It is due to the inability of some subjects in showing the mentioned expression in this database, not the deficiency of the pose-invariant feature, since we obtained similar results for the other methods [16].

4.2 ADFES database

The ADFES has 10 discrete emotional expressions including anger, disgust, fear, joy, sadness, and surprise, contempt, pride, embarrassment, and neutral. Sample frames taken from a subject showing all 10 emotions are depicted in Fig. 2.

We only used the video samples where subjects turn the face away from the camera for evaluating the efficiency of the proposed method in detecting the emotions from non-frontal faces (216 samples). Figure 3 shows a sample video of the database with face turning away from the camera.

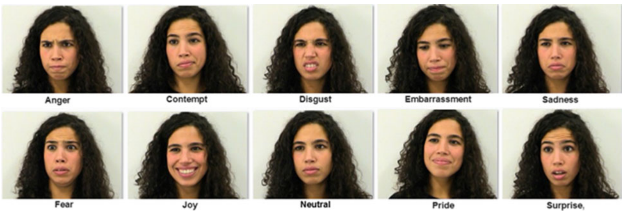

Fig. 2 Sample frames of ADFES database displaying 10 emotions

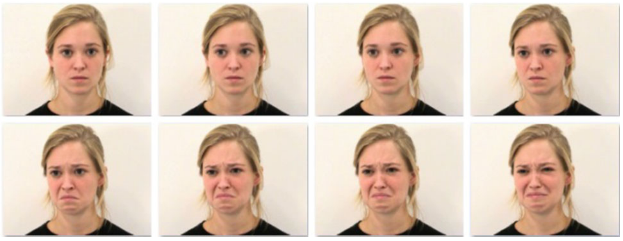

Fig. 3 Sample frames of a video sequence from ADFES database with moving face

Table 4 Comparison of the proposed descriptor to other methods for the ADFES database

Method	Detection rate
LBP-TOP [2]	77.08
OF-based Pose-Invariant Descriptor [18]	80.25
Landmark-based Pose-Invariant Descriptor	88.73

We used SVM as a classifier for all methods

We again applied polynomial SVM classifier and evaluated our proposed pose-invariant descriptor using 5-fold cross-validation. The accuracy of the proposed method compared to the other dynamic descriptors is tabulated in Table 4. As shown, our proposed Landmark-based Pose-Invariant Descriptor has the best performance.

Since we estimate the location of face interest points with respect to the nose point at each frame (nose point is a stable point in expressing an emotion and its movement is related to head motion), some kind of head motion correction is performed before feature extraction. Therefore, we expect that the method can tolerate slight head movement during expressing an emotion. However, for strong and complex head motion, more accurate method for head motion correction may be needed. As our estimation is based on the

Fig. 4 Sample video frames from AVEC2011 database. (Extracted from [19])

projection of 3D head movement on a 2D plane, using the 3D techniques for head movement estimation may be more accurate for real-world applications.

Although the results confirm the efficiency of our method for emotion detection from non-frontal faces, the head movement in the ADFES database is limited and the pose angle for all subjects is about the same. Indeed, the power of the technique will be more apparent when applying more accurate technique for global head motion correction for more challenging multi-view databases.

4.3 AVEC2011 database

The AVEC2011 database uses four affective dimensions - activation, expectation, power and valence for each video frame. The data are divided into 3 subsets: training, development, and testing. The training subset consists of 31 records, while the development subset (for validation of model parameters) consists of 32 sequences and the test subset consists of 11 video sequences.

We partition each video into segments containing 60 frames with 20 % overlap between the segments. Each segment is then down sampled at a rate of 6. Thus, each volume data include only 10 frames. We process only 1550 frames of each video for the training and development subsets (total of 48050 frames for training and 49600 frames for development). Fig. 4 shows a few sample frames of this database.

Table 5 tabulates the comparison of our results to the baseline and the other reported results. Comparison using the development set is not a good measure because the total frames were not included in the experiments and thus the sampling process may not extract the required emotion. However, for the test set, the proposed descriptor outperforms the one used in the baseline. It also shows that new proposed Pose-Invariant Descriptor is superior to our earlier proposed OF-based descriptor.

Our experiments on AVEC database which is a natural expression containing head movement, partial occlusion, and illumination variations confirm that the method can somewhat tolerate such situations. Indeed, the accuracy of our proposed feature extraction dealing with head pose variation, occlusion, and illumination changes highly depends on accurate localization of the interest points. The method used in this work to locate and track the facial landmarks is based on SDM proposed in [9]. The authors in this paper experimentally show that SDM can reliably track facial landmarks with large pose (∓45° yaw, ∓90° roll and, ∓30° pitch), occlusion and illumination. Fig. 5 depicts some frame examples of AVEC database in case of partial occlusion and head movement where SDM is able to accurately localize the interest points, while Fig. 6 shows several cases with extreme head movement and facial occlusion where the method is not able to localize the interest points. In these cases, the landmarks are estimated based on previous frames. So, some error is introduced in point localization and consequently the features will be corrupted.

5 Conclusion

We proposed a novel descriptor for facial emotion analysis from video sequences. The advantages of the descriptor include robustness to continuous head pose variations, insensitivity to emotion speed variations, simple and fast computation. However, the method suffers from errors in interest point localization. If the interest point localization fails, our

Table 5 Comparison of the detection rate for the AVEC2011 database to check the efficiency of the proposed descriptor

Method	Test				
	A	*E*	*P*	*V*	Average
Baseline [14]	42.2	53.6	36.4	52.5	46.2
[20]	65.5	61.7	47.1	69.8	61.0
[21]	56.5	59.7	48.5	59.2	55.9
[22]	56.9	47.5	47.3	55.5	51.8
[16]	75.6	59.8	56.9	56.9	62.3
[17]	57.8	53.5	61.6	61.5	58.6
OF-based Pose-Invariant Descriptor + SVM [18]	57.7	60.0	48.4	60.0	56.5
Landmark-based Pose-Invariant Descriptor + SVM	56.0	66.4	57.2	66.4	61.5

A activation, *E* expectancy, *P* power, *V* valence

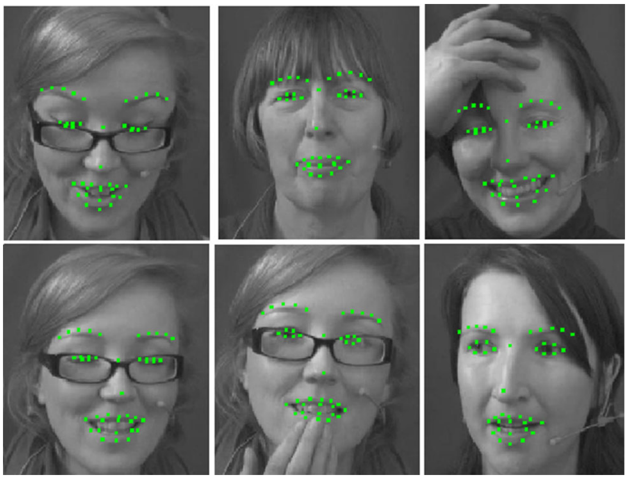

Fig. 5 Example of successful point localization from AVEC database in case of head movement and partial occlusion

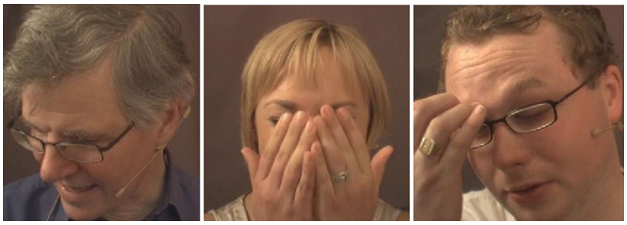

Fig. 6 Example of failure cases in point localization from AVEC database in case of extreme head movement and facial occlusion

proposed method will also fail. Indeed, the accuracy of our proposed feature extraction dealing with head pose variation, occlusion, and illumination changes is dependent on accurate interest point localization.

We validated the performance of the descriptor using 3 publicly available databases. Experimental results on the CK^+ database achieved an accuracy of 95.76 %, while for the ADFES database it achieved 88.73 %. For AVEC2011 database also, we obtained comparable recognition rate to the state-of –the-arts. Our experimental results confirmed the efficiency of the method for multi-view expression recognition even if face is slightly turning during expressing the emotion. The high performance of the method suggests that it is applicable for dynamic video in natural settings.

The method used in this work to locate and track the facial landmarks is based on SDM proposed in [9] which has been shown to reliably track facial landmarks with large pose of ($\mp 45°$ yaw, $\mp 90°$ roll and, $\mp 30°$ pitch). However, this will not be sufficient for fast and large head movements, it will fail. Thus future works include improving the estimation of global head motion via 3D techniques, and a better classifier that can take advantage of the dynamic nature of the emotion.

Acknowledgments This research is supported by the Agency for Science, Technology and Research (A*STAR), Singapore.

References

1. Wehrle, T., Kaiser, S., Schmidt, S., Scherer, K.R.: Studying the dynamics of emotional expression using synthesized facial muscle movements. J. Personal. Soc. Psychol. **78**, 105–119 (2000)
2. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 915–928 (2007)
3. Bihan, J., Valstar, M. F., Pantic, M.: Action unit detection using sparse appearance descriptors in space-time video volumes. In: IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)
4. Rudovic, O., Pantic, M., Patras, I.: Coupled Gaussian processes for pose-invariant facial expression recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1357–1369 (2013)
5. Jeni, L.A., et al.: 3D shape estimation in video sequences provides high precision evaluation of facial expressions. Image Vis. Comput. **30**, 785–795 (2012)
6. Songfan, Y., Bhanu, B.: Understanding discrete facial expressions in video using an emotion avatar image. IEEE Trans. Syst. Man Cybern. Part B Cybern. **42**, 980–992 (2012)
7. Zheng, W., Tang, H., Lin, Z., Huang, T.: Emotion recognition from arbitrary view facial images. Comput. Vis. ECCV 2010 **6316**, 490–503 (2010)
8. Kumano, S., Otsuka, K., Yamato, J., Maeda, E., Sato, Y.: Pose-invariant facial expression recognition using variable-intensity templates. Int. J. Comput. Vis. **83**, 178–194 (2009)
9. Xiong, X., De La Torre, F.: Supervised descent method and its applications to face alignment. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 532–539 (2013)
10. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Comput. **27**, 803–816 (2009)
11. Lucey, P., et al.: The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101 (2010)
12. Kanade, T., Cohn, J. F., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG00), Grenoble, France, pp. 46–53
13. Van der Schalk, J., Hawk, S.T., Fischer, A.H., Doosje, B.J.: Moving faces, looking places: the Amsterdam dynamic facial expressions set (ADFES). Emotion **11**, 907–920 (2011)
14. Schuller, B., et al.: AVEC 2011—the first international audio/visual emotion challenge. Affect. Comput. Intell. Interact. **6975**, 415–424 (2011)
15. Shojaeilangari, S., Yau, W.Y., Teoh, E.K.: Dynamic facial expression analysis based on histogram of local phase and local orientation. In: International Conference on Multimedia and Human-Computer Interaction (MHCI), Canada (2013)
16. Shojaeilangari, S., Yau, W.Y., Li, J., Teoh, E.K.: Multi-scale analysis of local phase and local orientation for dynamic facial expression recognition. J. Multimed. Theory Appl. (JMTA) **2**, 1–10 (2014)
17. Shojaeilangari, S., Yau, W.Y., Teoh, E.K.: A novel phase congruency based descriptor for dynamic facial expression analysis. Pattern Recognit. Lett. **49**, 55–61 (2014)
18. Shojaeilangari, S., Yau, W.Y., Nandakumar, K., Li, J., Teoh, E.K.: Robust representation and recognition of facial emotions using extreme sparse learning. IEEE Trans. Image Process. **24**, 2140–2152 (2015)
19. Meng, H., Bianchi-Berthouze, N.: Affective state level recognition in naturalistic facial and vocal expressions. IEEE Trans. Cybern. **44**, 315–328 (2014)

20. Ramirez, G., Baltrušaitis, T., Morency, L.-P.: Modeling latent discriminative dynamic of multi-dimensional affective signals. Affect. Comput. Intell. Interact. **6975**, 396–406 (2011)
21. Cruz, A., Bhanu, B., Yang, S.: A psychologically-inspired match-score fusion model for video-based facial expression recognition. Affect. Comput. Intell. Interact. **6975**, 341–350 (2011)
22. Glodek, M., et al.: Multiple classifier systems for the classification of audio-visual emotional states. Affect. Comput. Intell. Interact. **6975**, 359–368 (2011)