A Unified Framework for Monocular Video-Based Facial Motion Tracking and Expression Recognition

Jun Yu(✉)

Department of Automation, University of Science and Technology of China,
Hefei 230026, China
harryjun@ustc.edu.cn

Abstract. This paper proposes a unified facial motion tracking and expression recognition framework for monocular video. For retrieving facial motion, an online weight adaptive statistical appearance method is embedded into the particle filtering strategy by using a deformable facial mesh model served as an intermediate to bring input images into correspondence by means of registration and deformation. For recognizing facial expression, facial animation and facial expression are estimated sequentially for fast and efficient applications, in which facial expression is recognized by static anatomical facial expression knowledge. In addition, facial animation and facial expression are simultaneously estimated for robust and precise applications, in which facial expression is recognized by fusing static and dynamic facial expression knowledge. Experiments demonstrate the high tracking robustness and accuracy as well as the high facial expression recognition score of the proposed framework.

Keywords: Facial motion tracking · Facial expression recognition

1 Introduction

Facial motion and expression enable users to communicate with computers using natural skills. Constructing robust systems for facial motion tracking and expression recognition is an active research topic.

Generally, 2D approaches [1, 2] or 3D approaches [3, 4] can be conducted for this task. Compared with 2D methods, 3D methods are more qualified for the view-independent and illumination insensitive tracking and recognition situations [5]. For 3D methods, a 3D facial mesh model or a depth camera is often used [4–8]. Because the high cost-effectiveness of single video cameras, they are used as inputs here by a 3D facial mesh model, which is served as the priori knowledge and constraints.

For facial motion tracking [9], appearance-based techniques [12, 13] are more robust compared with feature-based ones [10, 11], and often implemented statistically to increase the robustness. Offline statistical appearance-based models [3, 14], such as 3D shape regression [15, 16], use a face image dataset taken under different conditions to learn the parameters of appearance model, while online statistical appearance models (OSAM) [17–19] are more flexible and efficient than the offline ones by updating the

© Springer International Publishing AG 2017
L. Amsaleg et al. (Eds.): MMM 2017, Part II, LNCS 10133, pp. 50–62, 2017.
DOI: 10.1007/978-3-319-51814-5_5

learned dataset progressively. In addition, an adequate motion filtering strategy should be adopted to obtain the true value. Particle filtering [19] has been widely used for the global optimization ability by using the Monte Carlo technique.

For facial expression recognition [20–22], static techniques use spatial ones or spatio-temporal features related to a single frame [23] to classify expressions by several statistical analysis tools [24–30], such as neural network, while dynamic techniques use the temporal variations of facial deformation to classify expressions by several statistical analysis tools [31–33], such as dynamic Bayesian networks.

In this paper, a framework (Fig. 1) is proposed for pose robust and illumination insensitive facial motion tracking and expression recognition on each video frame base on the work in [34].

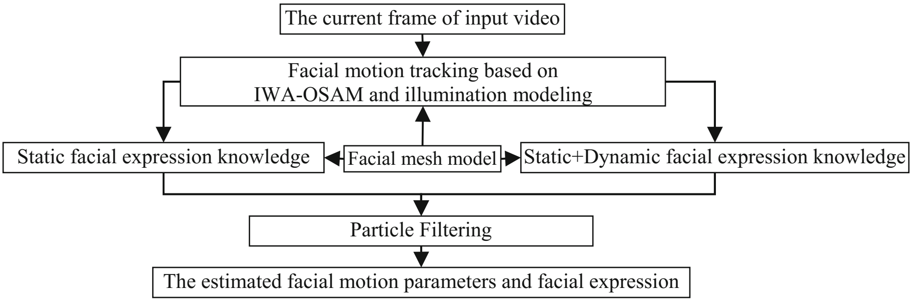

The current frame of input video
Facial motion tracking based on IWA-OSAM and illumination modeling
Static facial expression knowledge
Facial mesh model
Static+Dynamic facial expression knowledge
Particle Filtering
The estimated facial motion parameters and facial expression

Fig. 1. Framework.

Firstly, facial animation and facial expression are obtained sequentially in particle filtering. To alleviate the illumination variation difficulty during facial motion tracking, OSAM is improved to illumination weight adaptive online statistical appearance model (IWA-OSAM), in which 13 basis point light positions are constructed to model the lighting condition of each video frame. Then facial expressions are recognized by the static facial expression knowledge learned from the anatomical definitions in [35].

Secondly, because both the temporal dynamics and static information are important for recognizing expressions [36], they are combined and fused here by tracking facial motion and recognizing facial expression simultaneously in particle filtering. Compared with the sequential approach discussed above, particles are not only generated by the resampling, but also predicted by the dynamic knowledge; thus resulting into a more accurate recognition result.

2 Facial Motion Tracking

2.1 OSM-Based Facial Motion Tracking

The model (Fig. 2(a)): *CANDIDE3* [37] is served as the priori knowledge and constraints for facial motion tracking, and defines the facial motion parameters as:

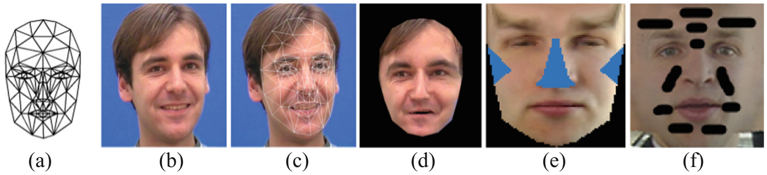


Fig. 2. (a) *CANDIDE3* model. (b) A frame of input video. (c) The projection of CANDIDE3 under $\boldsymbol{b}$. (d) The GNFI. (e) The improved GNFI. (f) Selected facial areas.

$$\boldsymbol{b} = [\theta_{\mathrm{x}}, \theta_y, \theta_z, t_x, t_y, t_z, \boldsymbol{\beta}^T, \boldsymbol{\alpha}^T]^T = [\boldsymbol{h}^T, \boldsymbol{\beta}^T, \boldsymbol{\alpha}^T]^T \tag{1}$$

where $\boldsymbol{h} = [\theta_{\mathrm{x}}, \theta_y, \theta_z, t_x, t_y, t_z]^T$ is global head motion parameters. $\boldsymbol{\beta}, \boldsymbol{\alpha}$ are shape and animation parameter. 10 shape parameter and 7 animation parameter are used.

A face texture is represented as a geometrically normalized facial image (GNFI) [37]. Figure 2(b)–(d) illustrate the process of obtaining a GNFI with an input image. Different facial areas may have different levels of influence on the tracking performance. Because the part above the eyebrows hardly take effect on the facial motion, and is often contaminated by the hair, it is removed from the GNFI. In addition, we found that the top part of the nose and the temples seldom undergo local motions. However, the appearance of these two facial areas is often influenced by head pose change and illumination variation. Therefore, the image regions corresponding to these two facial areas are removed from GNFI. The resulting image, called improved GNFI (Fig. 2(e)), is then used for measurements extraction.

Because the pixel color values are easily influenced by environment, and thus not robust for tracking, a more robust measurement is extracted from the improved GNFI as follows according to the discussion in [34]: for the improved GNFIs of the first and current frames, we obtain the illumination ratio images, and compute Gabor wavelet coefficients on the selected facial areas (Fig. 2(f)) where high frequency appearance changes more likely.

Moreover, illumination variation is one of the most important factors which reduce significantly the performance of face recognition system. It has been proved that the variations between images of the same face due to illumination are almost always larger than image variations due to change in face identity. So eliminating the effects due to illumination variations relates directly to the performance and practicality of face recognition. To alleviate this problem, a low-dimensional illumination space representation (LDISR) of human faces for arbitrary lighting conditions [38] is proposed for recognition. The key idea underlying the representation is that any lighting condition can be represented by 9 basis point light sources. The lighting subspace is constructed not using the eigenvectors from the training images with various lighting conditions directly but the light sources corresponding to the eigenvectors. The 9 basis light positions are shown in Table 1.

However, it is only used for the situation in which the face has only the 2D in-plane rotation, while the face with out-of-plane rotation is a common situation for the face in the real scene. Therefore, we extend the LDISR from 2D to 3D, in which the in-plane rotation and out-of-plane rotation are both considered. The training process is similar to that in the method discussed in [38], and the obtained 13 basis light positions are shown in Table 2.

Table 1. Positions of the 9 basis light sources.

Light	Pitch θ (degree)	Roll φ (degree)
1	0.0	0.0
2	17.5	−47.5
3	25.7	44.4
4	36.0	−108.0
5	44.0	88.0
6	68.6	−3.0
7	33.3	85.0
8	−35.0	−95.0
9	−70.0	22.5

Table 2. Positions of the 13 basis light sources.

Light	Pitch θ (degree)	Roll φ (degree)	Yaw ω (degree)
1	0.0	0.0	0.0
2	17.5	−47.5	−59.5
3	25.7	44.4	−60.7
4	36.0	−108.0	−55.7
5	44.0	88.0	−60.9
6	68.6	−3.0	55.6
7	−33.3	85.0	60.3
8	−35.0	−95.0	65.0
9	−70.0	22.5	60.1
10	−89.2	0.7	−91.3
11	−90.2	0.4	−90.4
12	90.1	−0.3	−90.6
13	91.1	−0.6	91.6

Because different human faces have similar 3D shapes, the LDISR of different faces is also similar. In addition, by using the normalization with GNFI (Fig. 2), it can be assumed that different persons have the same LDISR.

Suppose the 13 basis images obtained under 13 basis lights are $L_i, i = 1, \cdots, 13$, the LDISR of human face can be denoted as $\boldsymbol{A} = [L_1, L_2, \cdots, L_{13}]$. Given an image of human face $\boldsymbol{I}_x$ under an arbitrary lighting condition, it can be expressed as:

$$\boldsymbol{I}_x = \boldsymbol{A} \cdot \boldsymbol{\lambda} \tag{2}$$

where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \cdots, \lambda_{13}]^T, 1 \leq \lambda_i \leq 1$ is the lighting parameters of image $\boldsymbol{I}_x$, and can be calculated by minimizing the energy function $E(\lambda)$ as:

$$E(\lambda) = \|\boldsymbol{A} \cdot \boldsymbol{\lambda} - \boldsymbol{I}_x\|^2 \tag{3}$$

Then the lighting parameters can be obtained as:

$$\lambda = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T \cdot \boldsymbol{I}_x \tag{4}$$

In practice, the image pixel will be less influenced by lighting if the light positions are distributed more evenly. In this case, the values of $\lambda_1, \lambda_2, \cdots, \lambda_{13}$ should be close for the 13 basis light positions discussed above. With this fact, an index can be defined to evaluate the lighting influence on the pixel values of each triangular patch of the improved GNFI. To achieve this goal, these triangular patches are first split to 13 areas corresponding to 13 basis point light positions, and then the lighting influence weight of the $kth, k = 1, \cdots, 13$ area is given for the tth video frame as follows:

$$w_t^k(j) = abs\left(\lambda_k - \frac{\lambda_1 + \cdots + \lambda_{13}}{13}\right) \Big/ \sum\nolimits_{k=1}^{13} abs\left(\lambda_k - \frac{\lambda_1 + \cdots + \lambda_{13}}{13}\right) \tag{5}$$

where j is the index of the pixel in the kth area of the triangular patches in Fig. 1(e).

Based on the lighting influence weight discussed above, OSAM is extended to illumination weight adaptive online statistical appearance model (IWA-OSAM). The details are as follows.

$\boldsymbol{m}(\boldsymbol{b}_t)$ with size d, abbreviated as $\boldsymbol{m}_t$, is the concatenation of pixel color value at time t, and it is modeled as a Gaussian Mixture stochastic variable with 3 components, s, w, l, as Jepson et al. [17] does. $\left\{\boldsymbol{\mu}_{i,t}; i = s, w, l\right\}$ is the mean vector. $\{\boldsymbol{\sigma}_{i,t}; i = s, w, l\}$ is the vector composed of the square roots of the diagonal elements of the covariance matrix. $\{\boldsymbol{k}_{i,t}; i = s, w, l\}$ is the mixed probability vector. The observation likelihood is $p(\boldsymbol{m}_t/\boldsymbol{b}_t)$, which is represented by the sum of the Gaussian distributions of 3 components, s, w, l, weighted by $\{\boldsymbol{k}_{i,t}; i = s, w, l\}$.

The IWA-OSAM represents the stochastic process of all observations until time t-1: $\boldsymbol{m}_{1:t-1}$. In order to enable IWA-OSAM to track target, $\{\boldsymbol{k}_{i,t}; i = s, w, l\}$ and $\boldsymbol{\mu}_{s,t}$, $\boldsymbol{\sigma}_{s,t}$ are updated when $\boldsymbol{b}_t$, $\boldsymbol{m}_t$ are got [18]. The following equations are valid for $j = 1, 2, \cdots, d$. $c = 0.2$ is forgetting factor.

$$\begin{aligned}
k_{i,t}(j) &= \left((1-c) + cN\left(w_{t-1}^k(j)m_{t-1}(j); \mu_{i,t-1}(j), \sigma_{i,t-1}^2(j)\right)\right)k_{i,t-1}(j) \\
\mu_{s,t}(j) &= (1-c)\mu_{s,t-1}(j)/k_{s,t}(j) + cw_{t-1}^k(j)m_{t-1}(j)k_{s,t-1}(j)/k_{s,t}(j) \\
\sigma_{s,t}^2(j) &= (1-c)\sigma_{s,t-1}^2(j)\Big/k_{s,t}(j) + c\left(w_{t-1}^k(j)m_{t-1}(j)\right)k_{s,t-1}(j)/k_{s,t}(j) - \mu_{s,t-1}^2(j)
\end{aligned} \tag{6}$$

Moreover, the methods discussed in [19] are used to reduce the influences of occlusion and outlier here.

Once the solution $\boldsymbol{b}_t$ is solved, the corresponding pixels in the resulting synthesis texture will be used to update IWA-OSAM. While IWA-OSAM are not updated for outlier or occlusion pixels, thus the outlier and occlusion cannot deteriorate IWA-OSAM.

3 Facial Expression Recognition

Static knowledge and dynamic knowledge are extracted to cope with the complex variability of facial expression.

3.1 Static Facial Expression Knowledge

The retrieved $\boldsymbol{\alpha}$ of one frame from the input video can be seen as a description of facial muscles activations of the person in that frame according to the definitions of action units in [36]. Therefore, the relationship between $\boldsymbol{\alpha}$ and facial expression modes is established, namely 7 typical vectors $\left\{\boldsymbol{\alpha}_{su}, \boldsymbol{\alpha}_{di}, \boldsymbol{\alpha}_{fe}, \boldsymbol{\alpha}_{ha}, \boldsymbol{\alpha}_{sa}, \boldsymbol{\alpha}_{an}, \boldsymbol{\alpha}_{ne}\right\}$ are chosen as the representatives of 7 universal facial expressions: surprise, disgust, fear, happy, sad, angry and neutral. They are set as the static knowledge for facial expression recognition.

When $\boldsymbol{\alpha}$ of one frame of input video is retrieved, the Euclidian distances between it and each of $\left\{\boldsymbol{\alpha}_{su}, \boldsymbol{\alpha}_{di}, \boldsymbol{\alpha}_{fe}, \boldsymbol{\alpha}_{ha}, \boldsymbol{\alpha}_{sa}, \boldsymbol{\alpha}_{an}, \boldsymbol{\alpha}_{ne}\right\}$ are computed, and the facial expression corresponding to the minimum distance is set as the recognition result.

3.2 Dynamic Facial Expression Knowledge

For each expression γ, a three layer Radial Basis Function (RBF) network is trained for describing the temporal evolution of facial animations $\boldsymbol{\alpha}_t$ as:

$$\boldsymbol{\alpha}_t = \boldsymbol{W} \cdot \boldsymbol{\Phi}(\boldsymbol{\alpha}_{t-1}) + \boldsymbol{B} \tag{7}$$

The middle layer contains 400 nodes. $\boldsymbol{W}(7 \times 400)$, $\boldsymbol{B}(7 \times 1)$ are the weight matrix and bias vector of the output layer. The *ith* node of middle layer is given by RBF:

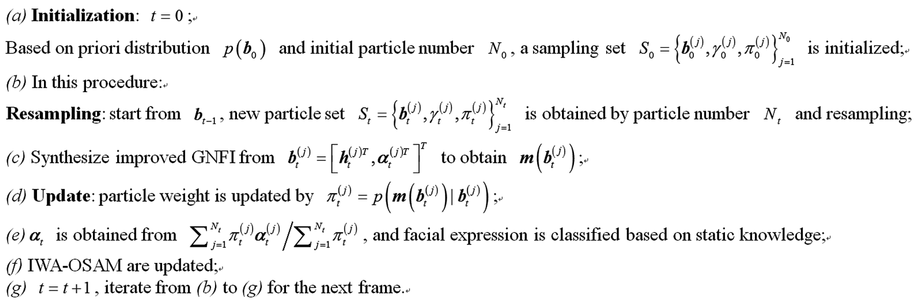
$$\boldsymbol{\Phi}_i(A) = exp(-(\|\boldsymbol{\alpha} - \boldsymbol{IW}_i\| \times B_{mi})^2) \tag{8}$$

where $\boldsymbol{IW}_i$ is the *ith* row component of the weight matrix of the middle layer $\boldsymbol{IW}(400 \times 7)$, and represents the mean value of the *ith* RBF. $\|\boldsymbol{\alpha} - IW_i\|$ represents the distance between $\boldsymbol{\alpha}$ and IW_i. B_{mi} is the *ith* component of bias vector of the middle layer $\boldsymbol{B}_m(400 \times 1)$, and its reciprocal represents the variance of the *ith* RBF. The dynamic of facial animations associated with the neutral expression is simplified as $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}$.

We define a transition matrix $\boldsymbol{T}$ whose entries $T_{\gamma',\gamma}$ describe the probability of transition between two expressions γ' and γ. The transition probabilities are learned from a database [39]. Then the RBF network is trained on the 60% of the Extended Cohn-Kanade (CK+) database 25 and the database [39]. The corresponding facial animations $\boldsymbol{\alpha}_t$ are tracked by IWA-OSAM. Therefore, the RBF network is set as the dynamic knowledge for facial expression recognition.

3.3 Framework

Given a facial video, we would like to estimate $\boldsymbol{b}_t$ and the facial expression for each frame at time *t*, by particle filtering, given all the observations up to time *t*. Therefore, we create a mixed state $\left(\boldsymbol{b}_t^T, \gamma_t\right)^T$, where $\gamma_t \in \{1, \cdots, 7\}$ is a discrete state, representing one of 7 universal expressions. For the estimation of $\left(\boldsymbol{b}_t^T, \gamma_t\right)^T$, two schemes are proposed. The first scheme (Fig. 3) is to infer $\left(\boldsymbol{b}_t^T, \gamma_t\right)^T$ sequentially, where facial expression is recognized by static facial expression knowledge. The second scheme (Fig. 4) is to infer $\left(\boldsymbol{b}_t^T, \gamma_t\right)^T$ simultaneously, where facial expression is recognized by fusing the static and dynamic facial expression knowledge.



(a) **Initialization**: $t = 0$;

Based on priori distribution $p(\boldsymbol{b}_0)$ and initial particle number N_0, a sampling set $S_0 = \left\{\boldsymbol{b}_0^{(j)}, \gamma_0^{(j)}, \pi_0^{(j)}\right\}_{j=1}^{N_0}$ is initialized;

(b) In this procedure:

Resampling: start from $\boldsymbol{b}_{t-1}$, new particle set $S_t = \left\{\boldsymbol{b}_t^{(j)}, \gamma_t^{(j)}, \pi_t^{(j)}\right\}_{j=1}^{N_t}$ is obtained by particle number N_t and resampling;

(c) Synthesize improved GNFI from $\boldsymbol{b}_t^{(j)} = \left[\boldsymbol{h}_t^{(j)T}, \boldsymbol{\alpha}_t^{(j)T}\right]^T$ to obtain $\boldsymbol{m}\left(\boldsymbol{b}_t^{(j)}\right)$;

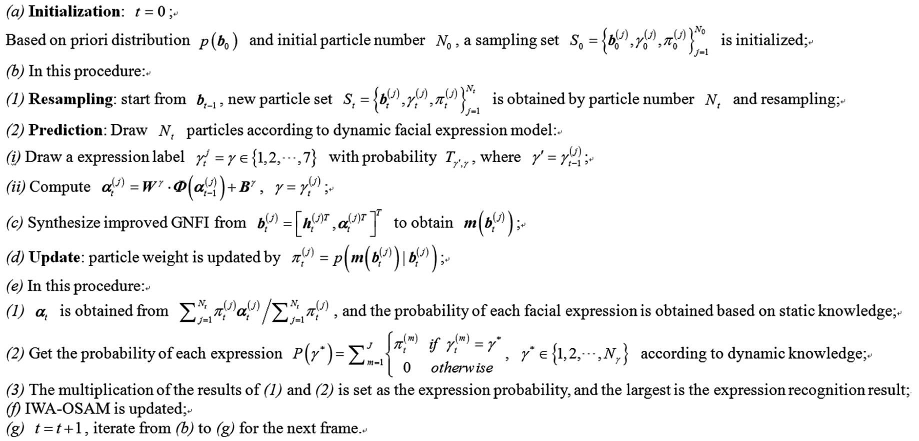
(d) **Update**: particle weight is updated by $\pi_t^{(j)} = p\left(\boldsymbol{m}\left(\boldsymbol{b}_t^{(j)}\right) \mid \boldsymbol{b}_t^{(j)}\right)$;

(e) $\boldsymbol{\alpha}_t$ is obtained from $\sum_{j=1}^{N_t} \pi_t^{(j)} \boldsymbol{\alpha}_t^{(j)} \Big/ \sum_{j=1}^{N_t} \pi_t^{(j)}$, and facial expression is classified based on static knowledge;

(f) IWA-OSAM are updated;

(g) $t = t+1$, iterate from *(b)* to *(g)* for the next frame.

Fig. 3. Inferring the facial motion and facial expression sequentially.



(a) **Initialization**: $t = 0$;

Based on priori distribution $p(\boldsymbol{b}_0)$ and initial particle number N_0, a sampling set $S_0 = \left\{\boldsymbol{b}_0^{(j)}, \gamma_0^{(j)}, \pi_0^{(j)}\right\}_{j=1}^{N_0}$ is initialized;

(b) In this procedure:

(1) **Resampling**: start from $\boldsymbol{b}_{t-1}$, new particle set $S_t = \left\{\boldsymbol{b}_t^{(j)}, \gamma_t^{(j)}, \pi_t^{(j)}\right\}_{j=1}^{N_t}$ is obtained by particle number N_t and resampling;

(2) **Prediction**: Draw N_t particles according to dynamic facial expression model:

(i) Draw a expression label $\gamma_t^j = \gamma \in \{1, 2, \cdots, 7\}$ with probability $T_{\gamma', \gamma}$, where $\gamma' = \gamma_{t-1}^{(j)}$;

(ii) Compute $\boldsymbol{\alpha}_t^{(j)} = \boldsymbol{W}^{\gamma} \cdot \boldsymbol{\Phi}\left(\boldsymbol{\alpha}_{t-1}^{(j)}\right) + \boldsymbol{B}^{\gamma}$, $\gamma = \gamma_t^{(j)}$;

(c) Synthesize improved GNFI from $\boldsymbol{b}_t^{(j)} = \left[\boldsymbol{h}_t^{(j)T}, \boldsymbol{\alpha}_t^{(j)T}\right]^T$ to obtain $\boldsymbol{m}\left(\boldsymbol{b}_t^{(j)}\right)$;

(d) **Update**: particle weight is updated by $\pi_t^{(j)} = p\left(\boldsymbol{m}\left(\boldsymbol{b}_t^{(j)}\right) \mid \boldsymbol{b}_t^{(j)}\right)$;

(e) In this procedure:

(1) $\boldsymbol{\alpha}_t$ is obtained from $\sum_{j=1}^{N_t} \pi_t^{(j)} \boldsymbol{\alpha}_t^{(j)} \Big/ \sum_{j=1}^{N_t} \pi_t^{(j)}$, and the probability of each facial expression is obtained based on static knowledge;

(2) Get the probability of each expression $P(\gamma^*) = \sum_{m=1}^{J} \begin{cases} \pi_t^{(m)} & \mathit{if}\ \gamma_t^{(m)} = \gamma^* \\ 0 & \mathit{otherwise} \end{cases}$, $\gamma^* \in \{1, 2, \cdots, N_\gamma\}$ according to dynamic knowledge;

(3) The multiplication of the results of *(1)* and *(2)* is set as the expression probability, and the largest is the expression recognition result;

(f) IWA-OSAM is updated;

(g) $t = t+1$, iterate from *(b)* to *(g)* for the next frame.

Fig. 4. Inferring the facial motion and facial expression simultaneously.

4 Evaluation

A workstation with Intel i7-6700K 4.0G, 8G memory and NVIDIA GTX960 is used.

4.1 Testing Dataset and Evaluation Methods for Facial Motion Tracking

A facial image sequence [5] and the IMM face database [40] with the ground truth landmarks available are used. They support point based comparison for errors, and a texture based test could also be performed on it. Besides, as the pose coverage and illumination condition variations of above databases are not large enough, the 13 videos, including Carphone and Forman image sequences, in the MPEG-4 testing database and 78 captured videos from 48 subjects with the resolution 352 × 288 are also used. The ground truth landmarks of them are obtained by manual adjustment.

Root Mean Square (RMS) landmark error measures the Root Mean Square Error (RMSE) between the ground truth landmark points and the fitted shape points after tracking, and is defined as:

$$\sum\nolimits_{i=1}^{N} sqrt\left(C_{fit}^{i} - C_{grd}^{i}\right) \Big/ N \tag{9}$$

where C_{est}^{i}, C_{grd}^{i} are the x or y coordinate of the *ith* fitted shape point and the *ith* ground truth landmark point.

4.2 Testing Dataset and Evaluation Methods for Facial Expression Recognition

The Extended Cohn-Kanade (CK+) database [25] and the database [39], not including the training part, are used. The database also presents the baseline results using AAM and a linear support vector machine classifier.

For evaluation, the facial expression recognition score is used, and a confusion matrix between different facial expression is used.

4.3 Facial Motion Tracking for Monocular Videos

Figure 5 shows the facial motion tracking results of several publically and captured videos. By computing the evaluation criteria, we can say that accurate tracking is

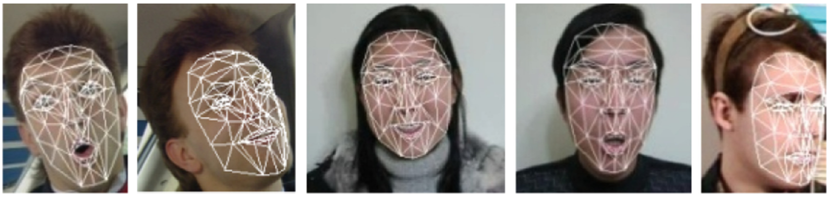


Fig. 5. Facial motion tracking results.

obtained even in the presence of perturbing factors including significant head pose and illumination as well as facial expression variations.

Based on the testing dataset, the comparison between different tracking algorithm is conducted. It should be stated that the work in [3] an offline method, and its performance is highly dependent on the training data. However, Active Shape Model (ASM)/Active Appearance Model (AAM) based trackers, such as that in [3], are the mainstream, and after learning the model in a dataset, the tracker can track any face without further training. Therefore, we compare this work with that in [3]. The images of the training dataset, which approximately correspond to every 4th image in the sequences of the testing dataset, are used to construct the AAM in [3], and the number of bases in the constructed AAM is chosen so as to keep 95% of the variations. The performances is evaluated on the testing data except for the training images.

By computing the evaluation criteria, Table 3 shows the superiority of our algorithm. This is because the IWA-OSAM in our proposed approach can learn the variation of facial motion effectively, and our proposed approach has the ability to alleviate lighting influence. Moreover, this is also because the improved GNFI is less influenced by the head pose change and illumination variation.

Table 3. Performance evaluation of different facial motion tracking algorithms.

	RMS landmark error
Our facial motion tracking algorithm	2.54
Our facial motion tracking algorithm not using illumination modeling	4.23
The method in [3]	4.33
The method in [34]	3.65
The method in [37]	5.29

4.4 Facial Expression Recognition

Figure 6 shows the facial expression recognition results on the testing database. As can be seen from it, facial expressions can be recognized effectively by our proposed algorithm in the presence of perturbing factors including significant head pose and illumination.

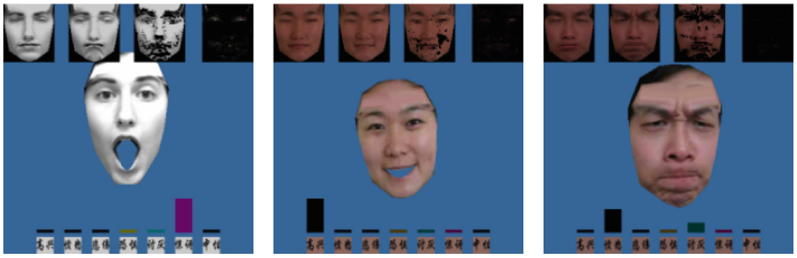


Fig. 6. Facial expression recognition results.

Based on the CK+ database, we compare our proposed algorithm to the methods in [20–22, 31, 34, 37] (Table 4). The recognition score of our facial expression recognition algorithm is higher than those of other algorithms, and also higher than those of our facial expression recognition algorithm not using illumination modeling.

Table 4. The accuracy comparison between several algorithms.

	Recognition score
Our facial expression recognition algorithm	87.4%
Our facial expression recognition algorithm not using illumination modeling	79.8%
The algorithm in [31]	82.2%
The algorithm in [31]	78.4%
The algorithm in [37]	80.1%
The algorithm in [21]	70.4%
The algorithm in [20]	72.7%
The algorithm in [22]	79.2%

According to the benchmarking protocol in the CK+ database, the leave-one-subject-out cross-validation configuration is used, and a confusion matrix is used to document the results. Table 5 shows the high recognition scores by our proposed algorithm. This is because that both static and dynamic knowledge are used, illumination is modeled and removed from improved GNFI to increase the accuracy and robustness of facial motion tracking.

Table 5. The confusion matrix of facial expression recognition by our algorithm.

	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	**89.7**	4.1	2.9	0.0	2.1	1.2
Disgust	0.9	**99.1**	0.0	0.0	0.0	0.0
Fear	3.1	0.0	**86.1**	4.2	0.0	6.6
Happy	0.0	0.0	0.0	**100.0**	0.0	0.0
Sad	6.1	2.8	2.7	0.0	**87.2**	1.2
Surprise	0.0	0.0	0.0	0.0	0.0	**100.0**

5 Conclusion

We propose a unified facial motion tracking and expression recognition framework for monocular video. For retrieving facial motion, an online weight adaptive statistical appearance method is embedded into the particle filtering strategy by using a deformable facial mesh model served as an intermediate to bring input images into correspondence by means of registration and deformation. For recognizing facial expression, facial animation and facial expression are estimated sequentially for fast and efficient applications, in which facial expression is recognized by static anatomical facial expression knowledge. In addition, facial animation and facial expression are simultaneously estimated for robust and precise applications, in which facial expression is recognized by fusing static and dynamic facial expression knowledge. Experiments demonstrate the high tracking robustness and accuracy as well as the high facial expression recognition score of the proposed framework.

In future, the recursive neural network will be used to learn the dynamic expression knowledge.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (No. 61572450, No. 61303150), the Open Project Program of the State KeyLab of CAD&CG, Zhejiang University (No. A1501), the Fundamental Research Funds for the Central Universities (WK2350000002), the Open Funding Project of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No. BUAA-VR-16KF-12), the Open Funding Project of State Key Laboratory of Novel Software Technology, Nanjing University (No. KFKT2016B08).

References

1. Black, M.J., et al.: Recognizing facial expressions in image sequences using local parameterized models of image motion. IJCV **25**(1), 23–28 (1997)
2. Gokturk, S., et al.: A data-driven model for monocular face tracking. In: ICCV, pp. 701–708 (2001)
3. Sung, J., Kanade, T., Kim, D.: Pose robust face tracking by combining active appearance models and cylinder head models. IJCV **80**(2), 260–274 (2008)
4. Dornaika, F., Davoine, F.: On appearance based face and facial action tracking. TCSVT **16** (9), 1107–1124 (2006)
5. Zeng, Z.H., et al.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. TPAMI **31**(1), 31–58 (2009)
6. Wen, Z., Huang, T.S.: Capturing subtle facial motions in 3D face tracking. In: ICCV, pp. 1343–1350 (2003)
7. Sandbach, G., et al.: Static and dynamic 3D facial expression recognition: a comprehensive survey. IVS **30**(10), 683–697 (2012)
8. Fang, T., Zhao, X., et al.: 3D facial expression recognition: a perspective on promises and challenges. In: ICAFGR, pp. 603–610 (2011)
9. Marks, T.K., et al.: Tracking motion, deformation and texture using conditionally Gaussian processes. TPAMI **32**(2), 348–363 (2010)
10. Zhang, W., Wang, Q., Tang, X.: Real time feature based 3-D deformable face tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5303, pp. 720–732. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88688-4_53
11. Liao, W.-K., Fidaleo, D., Medioni, G.: Integrating multiple visual cues for robust real-time 3D face tracking. In: Zhou, S.Kevin, Zhao, W., Tang, X., Gong, S. (eds.) AMFG 2007. LNCS, vol. 4778, pp. 109–123. Springer, Heidelberg (2007). doi:10.1007/978-3-540-75690-3_9
12. Cascia, M.L., et al.: Fast, reliable head tracking under varying illumination: an approach based on registration of texture mapped 3D models. TPAMI **22**(4), 322–336 (2000)

13. Fidaleo, D., Medioni, G., Fua, P., Lepetit, V.: An investigation of model bias in 3D face tracking. In: Zhao, W., Gong, S., Tang, X. (eds.) AMFG 2005. LNCS, vol. 3723, pp. 125–139. Springer, Heidelberg (2005). doi:10.1007/11564386_11
14. Liao, W.K., et al.: 3D face tracking and expression inference from a 2D sequence using manifold learning. In: CVPR, pp. 3597–3604 (2008)
15. Cao, C., Lin, Y., Lin, W.S., Zhou, K.: 3D shape regression for real-time facial animation. TOG **32**(4), 149–158 (2013)
16. Cao, C., et al.: Displaced dynamic expression regression for real-time facial tracking and animation. In: SIGGRAPH, pp. 796–812 (2014)
17. Jepson, A.D., Fleet, D.J., et al.: Robust online appearance models for visual tracking. TPAMI **25**(10), 1296–1311 (2003)
18. Lui, Y.M., et al.: Adaptive appearance model and condensation algorithm for robust face tracking. TSMC Part A **40**(3), 437–448 (2010)
19. Yu, J., Wang, Z.F.: A video, text and speech-driven realistic 3-D virtual head for human-machine interface. IEEE Trans. Cybern. **45**(5), 977–988 (2015)
20. Wang, Y., et al.: Realtime facial expression recognition with Adaboost. In: ICPR, pp. 30–34 (2004)
21. Bartlett, M., Littlewort, G., Lainscsek, C.: Machine learning methods for fully automatic recognition of facial expressions and facial actions. In: ICSMC, pp. 145–152 (2004)
22. Zhang, Y., Ji, Q.: Active and dynamic information fusion for facial expression understanding from image sequences. TPAMI **27**(5), 699–714 (2005)
23. Tian, Y.L., et al.: Facial expression analysis. In: Li, S.Z., Jain, A.K. (eds.) Handbook of Face Recognition. Springer, New York (2005)
24. Chang, Y., et al.: Probabilistic expression analysis on manifolds. In: CVPR, pp. 520–527 (2004)
25. Lucey, P., et al.: The extended Cohn-Kande dataset (CK+): a complete facial expression dataset for action unit and emotion-specified expression. In: CVPR, pp. 217–224 (2010)
26. Tian, Y., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. TPAMI **23**, 97–115 (2001)
27. Hamester, D., et al.: Face expression recognition with a 2-channel convolutional neural network. In: IJCNN, pp. 12–17 (2015)
28. Wang, H., Ahuja, N.: Facial expression decomposition. In: ICCV, pp. 958–963 (2003)
29. Lee, C., Elgammal, A.: Facial expression analysis using nonlinear decomposable generative models. In: IWAMFG, pp. 958–963 (2005)
30. Zhu, Z., Ji, Q.: Robust realtime face pose and facial expression recovery. In: CVPR, pp. 1–8 (2006)
31. Cohen, L., Sebe, N., et al.: Facial expression recognition from video sequences: temporal and static modeling. CVIU **91**(1–2), 160–187 (2003)
32. North, B., Blake, A., et al.: Learning and classification of complex dynamics. TPAMI **22**(9), 1016–1034 (2000)
33. Zhou, S., Krueger, V., Chellappa, R.: Probabilistic recognition of human faces from video. CVIU **91**(1–2), 214–245 (2003)
34. A Video-Based Facial Motion Tracking and Expression Recognition System. Multimed. Tools and Appl. (2016). doi:10.1007/s11042-016-3883-3
35. Ekman, P., Friesen, W., et al.: Facial Action Coding System: Research Nexus. Network Research Information, Salt Lake City (2002)
36. Schmidt, K., Cohn, J.: Dynamics of facial expression: normative characteristics and individual differences. In: ICME, pp. 728–731 (2001)
37. Dornaika, F., Davoine, F.: Simultaneous facial action tracking and expression recognition in the presence of head motion. IJCV **76**(3), 257–281 (2008)

38. Hu, Y.K., Wang, Z.F.: A low-dimensional illumination space representation of human faces for arbitrary lighting conditions. Acta Automatica Sinica **33**(1), 9–14 (2007)
39. http://www.chineseldc.org/emotion.html
40. Nordstrøm, M.M., et al.: The IMM face database - an annotated dataset of 240 face images. Technical report, Technical University of Denmark (2004)