Multimed Tools Appl (2016) 75:12563–12580
DOI 10.1007/s11042-014-2347-x

Visual facial expression modeling and early predicting from 3D data via subtle feature enhancing

Lumei Su · Feng Lu

Received: 16 July 2014 / Revised: 28 October 2014 / Accepted: 31 October 2014 /
Published online: 25 November 2014

Abstract This work investigates a new challenging problem: how to exactly recognize facial expression captured by a high-frame rate 3D sensing as early as possible, while most works generally focus on improving the recognition rate of 2D facial expression recognition. The recognition of subtle facial expressions in their early stage is unfortunately very sensitive to noise that cannot be ignored due to their low intensity. To overcome this problem, two novel feature enhancement methods, namely, adaptive wavelet spectral subtraction method and SVM-based linear discriminant analysis, are proposed to refine subtle features of facial expressions by employing an estimated noise model or not. Experiments on a custom-made dataset built using a high-speed 3D motion capture system corroborated that the two proposed methods outperform other feature refinement methods by enhancing the discriminability of subtle facial expression features and consequently make correct recognitions earlier.

Keywords Facial expression · Feature enhancement · Adaptive wavelet spectral subtraction · Linear discriminant analysis-based SVM

1 Introduction

Facial expression is one of the most important ways that people communicate emotion and other mental signal besides verbal expressions. As an active and challenging research topic in computer vision, facial expression recognition impacts many important applications in areas such as human-computer interaction and human affective recognition, and it can be also combined with other techniques such as human gaze estimation [14–16] for human

L. Su
Xiamen University of Technology, Amoy, Fujian, People's Republic of China
e-mail: sulumei@ut-vision.org

F. Lu (✉)
The University of Tokyo, Tokyo, Japan
e-mail: lufeng@ut-vision.org

behavior analysis. Most research on facial expression analysis has focused on obvious facial expressions near or in the apex phase, not on subtle facial expressions in the onset stage (for instance, [25, 26, 34]). However, it is important for a wide variety of applications to be able to understand human emotion quickly, especially for affective computing and human-computer interaction. In actual situations, people often reveal their true emotion in a brief and subtle facial expression that does not progress into an obvious expression. They even momentarily reveal how they truly feel through a subtle facial expression and then try to hide their feeling by switching their expression [9]. Recognizing facial expressions as early as possible is essential for correctly understanding human emotions. On the other hand, natural real-time human computer interaction requires that the robot quickly understand the person's emotion, like another person would. However, there has been little work on early facial expression recognition.

Early facial expression recognition is to recognize facial expressions as early as possible during its onset phase. In other words, it can be considered as subtle facial expression recognition. Subtle facial expression features extracted from early stages are always intermingled and thus cannot be readily discriminated into different categories. This makes the problem more difficult than conventional facial expression recognition or other recognition/evaluation problems [31–33]. It's worth noting if the subtle facial expression features are too sensitive to noise due to their low intensity; in that case, the noise will unavoidably affect the recognition results. To handle this problem more efficiently, in this paper we extend our analysis from 2D to 3D like other recent techniques [12, 13]. We collect 3D facial expression data by using a 3D motion capture system. The noise containing in facial expression features is roughly generated from two aspects. One is the measurement error generated from 3D motion capture system such as calibration error and tracking error. The other one is some non-expressional facial motion accompanying with facial expression displaying such as blinking. The expressionless facial motion can be ignored when the facial expression is obvious (expressional facial motion is prominent). However, because the facial expression in early stage is subtle, the expressionless facial motion may confuse the early recognition results. Therefore, we propose to enhance the subtle extracted facial expression features before classifying them.

Several researchers attempt to develop techniques to analyze subtle facial expressions [7, 17, 19, 20, 22, 24, 28]. However, most of these researchers focus on exactly representing the subtle facial changes in facial expressions and hardly investigate subtle facial expression recognition performance [7, 19, 28] because they rarely consider the noise effect. Song et al. [19] utilize a method based on vector field decomposition to model subtle facial expressions. Few effective approaches are proposed to recognize subtle facial expressions. In [17], Park et al. propose an effective method of recognizing subtle facial expression using motion magnification to transform subtle facial expressions into their corresponding exaggerated facial expressions. Here, motion magnification means magnifying facial motion deformation. However, their magnification method probably magnifies the noise part containing in subtle facial expression features simultaneously, that would inevitably affect the final classification result.

Because the existing of noise would result in data being incorrectly classified to some extent, some feature refinement techniques have been developed to reduce the noise influence on the observed features [3, 5, 6, 8, 23] in other areas. Generally, these feature refinement methods are directly removing the noise from the observed data (features) by analyzing the noise characteristics in frequency domain or spatial domain [1, 3–6, 8, 23, 27]. A suitable filtering algorithm is popularly applied to most of noise reduction work to remove the measurement errors generated from the measurement sensors. In [1], Alexa et al. used Wiener filter as a low-pass filter to remove the noise corresponding to higher frequencies of noisy surface meshes. Rhijn et al. [27] applied a third order smoother Kalman filter to smooth a set of motion capture data. Different from the low-pass filters, the Kalman filter will not make computation errors at the end of a data set. Therefore, it can be applied in real-time processing systems. Existing subtle facial expression works rarely consider the noise effect to subtle features and the resulting unsatisfied recognition rate, whereas some obvious facial expression recognition works consider refining obvious facial features. Most of them enhance their obvious facial expression features based on a hypothesis that, compared to noise, the meaningful facial features compose most of the principle component in feature spatial space. Generally, principal component analysis (PCA) is popularly applied to refine obvious facial features. For example, Calder et al. [4] apply PCA to remove redundant features by extracting the principal component in spatial domain.

The limitation of above feature refinement approaches is that they generally remove or reduce the noise based on a hypothesis that the meaningful facial features compose most of the principle components in spatial domain or correspond to low frequency components of noisy facial features in frequency domain. However, subtle facial features are probably equal to the noise in the principal components and are probably distributed at lower frequencies due to their subtle deformation. And the noise may not correspond to higher frequencies of noisy facial features in frequency domain. Another limitation of existing refining methods is that most of feature refinement methods are proposed to reduce the measurement errors generated from measurement sensors. For facial expression analysis, the noise means unwanted information that is not relevant to the facial expression information being investigated. The noise generated from the unexpected facial motions such as blinking and rigid head motion would definitely influence the subtle facial expression information resulting incorrect classification results. Therefore, the noise generated from the unexpected facial motions is also need to be reduced for subtle facial expression recognition. Therefore, the existing refining methods cannot overcome early facial expression recognition problem.

In this paper, two novel feature refinement methods are proposed to reduce the chance of noise decreasing the discriminability of subtle features and consequently make correct recognitions earlier. One is adaptive wavelet spectral subtraction method, which is developed from our previous work in [21]. This method refines subtle features by using an estimated noise model which is trained from several sets of noise data collected by our 3D motion capture system. Different from existing filtering methods, the proposed adaptive wavelet spectral subtraction method is not limited to reduce the noise distributing at higher frequencies. Moreover, the noise generated from the unexpected facial motions is also considered to estimate the noise model. In other words, this method can reduce the influence of noise from the unexpected facial motions. Last but not least, the adaptive wavelet spectral subtraction method is novelly analyzing the spatial-temporal characteristics of noise. Therefore, the noise that is probably nearly equal in intensity to the subtle facial features in spatial domain can be effectively reduced by the proposed method from spatial-temporal domain. To the best of our knowledge, this is the first effort that refines subtle facial features in the spatial-temporal domain.

The other proposed method is to refine subtle features by using SVM-based LDA without requiring a noise model. For the former proposed refinement method, the recognition performance is largely depends on the accuracy of an available noise model. In the situation that the noise model is not available or accurate enough, the feature refinement method using SVM-based LDA works. This feature refinement method can improve subtle facial expression classification performance by cooperating linear discriminant analysis (LDA) with support vector machine (SVM). The final goal of feature refinement is improving the classification performance. Generally speaking, feature refinements in existing works are independently performed before feature classification. Therefore, the improvement in classification performance from the feature refinements cannot be directly evaluated. Different from existing de-noise methods, the SVM-based LDA integrating together feature refinement and feature classification can improve the classification performance by directly reducing the influence of noise on feature classification. To the best of our knowledge, this is the first effort that refines subtle facial features by using a feature classification-based feature refinement method.

The contributions of this paper can be summarized as follows.

- We investigate early facial expression recognition problem with a high-frame rate 3D motion capture system, while current facial expression recognition works are based on 2D facial expression or 3D facial expression at a lower frame rate. The 3D facial motion can provide facial deformation along with depth direction which is helpful for analyzing subtle facial motion in early stage. And the motion capture system with high-frame rate can capture the quick facial changes that occur when forming facial expressions in their early stage.
- Two novel feature enhancement methods are proposed to refine the noisy subtle facial expression features considering estimating a noise model or not. One is adaptive wavelet spectral subtraction method. This method refines subtle features by using an estimated noise model which is trained from some sets of noise data collected with our 3D motion capture system. In this paper, a simple review of adaptive wavelet spectral subtraction method is given. The other proposed feature refining method is using SVM-based LDA to enhance subtle features without the limitation of learning a noise model. The feature refining method named SVM-based LDA can improve early facial expression recognition performance by cooperating with Support vector machines (SVM), as the margin of SVM can be enlarged by using LDA to maximize class separability. By comparing with the numerical experiment results of the former proposed refining method, we could reliably investigate the early recognition performance of the SVM-based LDA feature refining method.

This paper is organized as follows. In Section 2, two early facial expression recognition architectures using different feature refining methods are introduced. In Section 3, numerical early expression recognition results on a 3D facial expression dataset are given to investigate the performance of the two proposed methods. Section 4 gives a conclusion of this paper.

2 Early facial expression recognition

In this section, two early facial expression recognition architectures based on different refining methods are explained. Figure 1 shows the overall framework of Architecture I and Fig. 2 shows the overall framework of Architecture II. In both architectures, Support vector machines (SVM) is employed to classify the refined facial features. Moreover, a theoretical analysis is given to explain why the proposed feature refinement method SVM-based LDA can improve early facial expression recognition performance by cooperating with SVM.

2.1 3D facial expression modelling

In this paper, we firstly utilize a high-speed, maker-based optical facial motion capture system to capture subtle facial motion that occurs when forming facial expressions in

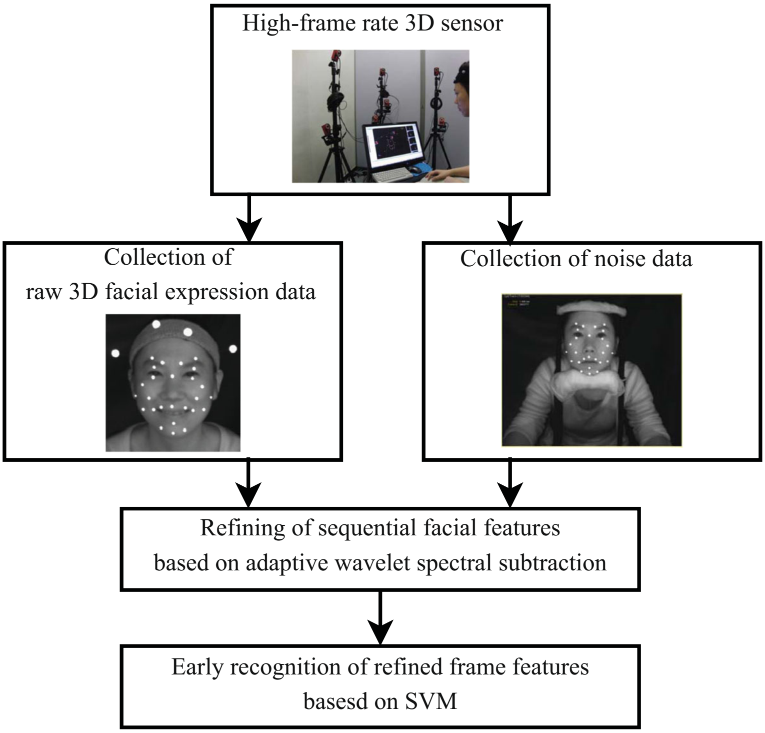
High-frame rate 3D sensor
Collection of raw 3D facial expression data
Collection of noise data
Refining of sequential facial features based on adaptive wavelet spectral subtraction
Early recognition of refined frame features basesd on SVM

Fig. 1 Architecture I of early facial expression recognition

early stage. We collect expression sequence data starting from a neutral expression, and use the concatenated displacement trajectory of M ($M = 26$ in our experiments) 3D facial markers to describe sequential feature of facial expression, which is denoted as $F = [f_1, f_2, \ldots, f_M] \in \mathbf{R}^{N\times 3M}$. The displacement trajectory of the j_{th} marker is specified as $f_j \in \mathbf{R}^{N\times 3}$, where $j = 1, 2, \ldots, M$. N denotes the total number of frames in the expression sequence. The concatenated 3D coordinate value of M 3D facial markers is used to describe frame feature of facial expression at frame t, which is denoted as $d_t = [x_{t,1}, y_{t,1}, z_{t,1}, \ldots, x_{t,M}, y_{t,M}, z_{t,M}] \in \mathbf{R}^{3M}$. The positions of M facial markers pasted on human face are determined by carefully considering muscle movements, as shown in Fig. 1. The four markers rigidly fixed to a user's head is used to align the raw 3D facial feature data from the world coordinate system into a facial coordinate system before feature refining.

2.2 Two early facial expression recognition architectures

Figure 1 shows the overall framework of Architecture I, which consists of four parts: (a) obtain facial expression features (facial motion data) from 3D motion capture system; (b) collect noise data from 3D motion capture system; (c) use the collected noise data to refine the captured facial expression features by using wavelet spectral subtraction; (d) infer the category of facial expression by using SVM.

Figure 2 shows the overall framework of Architecture II, which consists of three parts: (a) obtain facial expression features (facial motion data) from 3D motion capture system; (b)

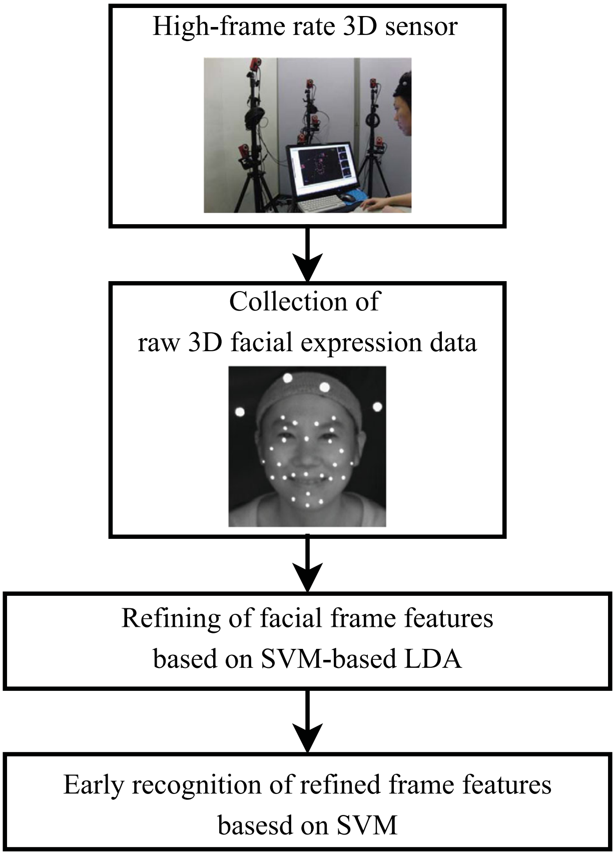
High-frame rate 3D sensor
Collection of
raw 3D facial expression data
Refining of facial frame features
based on SVM-based LDA
Early recognition of refined frame features
basesd on SVM

Fig. 2 Architecture II of early facial expression recognition

refine and the captured facial expression features by using SVM-based linear discriminant analysis (LDA); (c) infer the category of facial expression by using SVM.

In Architecture I, some noise datasets are collected to estimate noise containing in facial expression features, as shown in Fig. 1. A feature refining method based on adaptive wavelet spectral subtraction is performed to refine sequential facial features F by using the collected noise data. Finally, classical SVM is applied to classified the refined facial frame features $\hat{d}_t$, which is introduced in Section 2.3.1. In Architecture II, a feature refining method based on SVM-based LDA is performed to indirectly refine facial frame features d_t and classified the refined facial frame features $\hat{d}_t$ without using the collected noise data, as shown in Fig. 2.

2.3 Early facial expression recognition based on subtle features enhancement

In this subsection, two feature refining methods utilized in Architecture I and Architecture II are introduced. It's worth noting that these two refining methods are operated on different types of facial expression features. The adaptive wavelet-based method is employed to refine sequential facial expression features F, while the SVM-based LDA is employed to refine frame features of facial expression d_t.

2.3.1 Early facial expression recognition based on adaptive wavelet spectral subtraction

Actually, the proposed adaptive wavelet spectral subtraction method is an improvement algorithm of classical wavelet thresholding methods, which are popularly applied to enhance noise speech signal. The wavelet transform is an efficient signal analysis method in spatio-temporal terms. The wavelet thresholding methods generally follow three steps. (a) Transform the noisy signal into wavelet coefficients. (b) Apply a soft or hard threshold at each scale. (c) Transform back the resulting coefficients and get the estimated signal.

Different with classical wavelet thresholding methods, the proposed adaptive wavelet spectral subtraction method employ an adaptive threshold at each scale which is estimated from the noise beforehand. As the analysis in the former section, the limitation of existing feature refinement methods is mostly based on a hypothesis that the noise mostly corresponds to higher frequencies of noisy signal (features) in frequency domain. According to this assumption, the wavelet thresholding methods can effectively remove the noise from noisy signal by simply using a constant threshold. However, compared to the subtle facial deformation of subtle expressions, the noise may correspond to low frequencies of noisy facial features in frequency domain. Therefore, an adaptive threshold which can well characterize the noise in subtle expressions is essential for wavelet thresholding method.

Actually, the process of refining sequential facial expression features F is operated on the displacement trajectory f_j of each single facial marker respectively. For the sake of clarity of presentation, we omit the index j of the j_{th} marker unless it is necessary to show it explicitly. Assume that sequential facial expression features f is a noisy observation signal of a clean sequential signal c so that

$$f = c + n \tag{1}$$

where n is the sequential noise. Taking wavelet packet transform of f, c and n, we get

$$\sum_{a,b} \lambda_{a,b}(f)\psi_{a,b}(t) = \sum_{a,b} \lambda_{a,b}(c)\psi_{a,b}(t) + \sum_{a,b} \lambda_{a,b}(n)\psi_{a,b}(t) \tag{2}$$

where $\psi_{a,b}(t)$ is complete orthogonal wavelet basis. The (2) becomes

$$\lambda_{a,b}(f) = \lambda_{a,b}(c) + \lambda_{a,b}(n) \tag{3}$$

where $\lambda_{a,b}(f)$, $\lambda_{a,b}(c)$ and $\lambda_{a,b}(n)$ are the wavelet packet coefficients of noisy feature, clean feature and noise. For the sake of clarity of presentation, we omit the index a, b of wavelet packet coefficients. Based on (3), we can obtain

$$|\lambda\,(f)\,|^2 = |\lambda\,(c)\,|^2 + |\lambda\,(n)\,|^2 + 2\lambda\,(c) \cdot \lambda\,(n) \tag{4}$$

$$|\lambda\,(c)\,|^2 = |\lambda\,(f)\,|^2 - (1 \pm 2\lambda\,(c)\,/\lambda\,(n))|\lambda\,(n)\,|^2 \tag{5}$$

According to normal spectral subtraction, we can roughly estimate the wavelet packet coefficients of clean facial expression features c as follow [29],

$$\hat{\lambda}\,(c) = \begin{cases} sgn\,(\lambda\,(f))\,(|\lambda\,(f)\,|^{\gamma} - \alpha|\hat{\lambda}(n)|^{\gamma})^{1/\gamma} \\ \text{if}\Big(|\lambda\,(f)\,|^{\gamma} - \alpha|\hat{\lambda}(n)|^{\gamma}\Big)^{1/\gamma} > \mu|\hat{\lambda}(n)| \\ \mu|\hat{\lambda}(n)| & \text{otherwise} \end{cases} \tag{6}$$

where $|\hat{\lambda}(n)|$ are the wavelet packet coefficients magnitude estimation of noise n. The estimation algorithm is described in next paragraph. The parameter γ selects the subtraction domain. The parameter α controls the amount of over-subtraction which is linear with signal to noise rate $\rho(\lambda)$.

Before performing wavelet spectral subtraction on the observed facial features, a noise model is estimated by using wavelet packet transform. Compared to the normal wavelet analysis, wavelet packet transform provides more precise analysis because it can choose to decompose not only the lower frequency domains, but also the higher frequency domains of a signal. Therefore, the noise model is characterized by synthesizing the wavelet packet coefficients of the noise.

In order to estimate the wavelet packet coefficients of noise, we collect some noise datasets $\{\Omega_a, \Omega_b, \ldots\}$ from 3D motion capture system. The coefficients estimation of the noise $\hat{\lambda}(n)$ are computed from following relationships [18],

$$|\lambda_t^*(\Omega)| = \delta|\lambda_{t-1}^*(\Omega)| + (1-\delta)|\lambda_t(\Omega)| \tag{7}$$

$$\hat{\lambda}(n) = \max_t |\lambda_t^*(\Omega)| \tag{8}$$

where $|\lambda_t(\Omega)|$ is the short-term wavelet packet coefficients estimation of noise data Ω at frame t and $|\lambda_t^*(\Omega)|$ is the smoothed-out coefficients estimation at frame t. δ is a memory parameter.

We refine the wavelet packet coefficients of the noisy feature f by using (6). The refined coefficients $\hat{\lambda}(c)$ can be used to estimate the refined sequential facial expression features $\hat{c} = W^{-1}(\hat{\lambda}(c))$, where $W^{-1}(\cdot)$ denotes wavelet packet reconstruction function. And the refined facial frame features $\hat{d_t}$ can be picked up from the the refined sequential facial expression features $\hat{c}$. Actually, our proposed wavelet spectral subtraction is an improvement wavelet packet. Different from standard wavelet packet transform, we don't use an empirical value to threshold the wavelet coefficients. We use the wavelet coefficients of the noise to adaptively threshold the wavelet coefficients of the noisy signal, which can be considered as adaptive wavelet spectral subtraction.

Support vector machine (SVM) is a powerful decision machine that provides a decision boundary to maximize the margin. Until recently, SVM was one of the most comprehensive studies with remarkable results on facial expression recognition. To classify facial expression features d_t, $K(K-1)/2$ different 2-class SVMs are firstly trained on all possible pairs of classes, and then classify facial features d_t to which class has the highest number of votes. The classification function of 2-class SVMs is defined as follows,

$$f(d_t) = \omega^T d_t + b \tag{9}$$

The weight vector ω and bias b can be obtained by solving the following optimization problem,

$$\min_{\omega \neq 0, b} \frac{1}{2}\|\omega\|^2 \tag{10}$$

$$s.t.\ \ y_i(\omega^T d_i + b) \geq 1 \ \ \forall i = 1, \ldots n. \tag{11}$$

where $y_i \in -1, +1$ is the label of facial feature d_t. As we all know, the solution of ω is a weighted sum of a set of support vectors d^{sv} as follows,

$$\omega = \sum_{i=1}^{N_{sv}} \alpha_i y_i^{sv} d_i^{sv} \ \ \alpha_i > 0 \tag{12}$$

where N_{sv} is the total number of support vectors.

2.3.2 Early facial expression recognition based on SVM-based LDA

This section introduces another feature refinement method with SVM-based linear discriminant analysis. Different from the former refining method, SVM-based LDA indirectly refines subtle features without requiring noise estimation. Moreover, the SVM-based LDA method integrates the feature refinement and feature classification together. Therefore, this method can more carefully consider the noise influence on the final classification results and effectively improve the classification performance.

The proposed SVM-based LDA can be considered as a classifier which also has the function of feature refinement by integrating LDA into the classical SVM. For subtle facial expression recognition, subtle facial features in their early stage are very sensitive to the noise, because they probably dominate a nearly equal component compared to the noise. The distribution of the support vectors which completely describes the decision surface are polluted by the noise that cannot be ignored. The optimal solution of SVM formulation, as shown in (10) is not the optimal margin classifier for subtle facial expression classification. In this section, LDA is proposed to enhance the distribution of noisy subtle facial features, as it can compact the noise and well preserve the class separability of subtle features simultaneously. The classification function of SVM-based LDA is defined as follows,

$$f(d_t) = \omega^T \phi^T d_t + b \tag{13}$$

where enhancement transformation matrix ϕ is obtained by LDA. The weight vector ω and bias b can be obtained by solving the following optimization problem,

$$\min_{\omega \neq 0, b} \frac{1}{2} \|\omega^T \phi^T\|^2 \tag{14}$$

$$s.t.\ \ y_i(\omega^T \phi^T d_i + b) \geq 1\ \ \forall i = 1, \ldots n. \tag{15}$$

Linear discriminant analysis (LDA) is one of the well-known dimensionality reduction methods which aim to find optimal transformation by minimizing the within-class distance and preserving the between-class distance simultaneously [2, 11]. In other words, LDA searches the best subspace which preserved class separability as much as possible in a lower dimensional space. In this section, the optimal transformation ϕ is applied to enhance the distribution of the observed subtle facial features d_t. The enhancement transformation matrix ϕ is obtained by maximizing the following optimization problem,

$$\phi = \arg\max_{\phi} Tr \left\{ (\phi S_W \phi^T)^{-1} (\phi S_B \phi^T) \right\} \tag{16}$$

where the within-class and between-class covariance matrix S_W and S_B are defined as,

$$S_W = \sum_{k=1}^{K} \sum_{j \in C_k} (d_j - m_k)(d_j - m_k)^T \tag{17}$$

$$S_B = \sum_{k=1}^{K} N_k (m_k - m)(m_k - m)^T \tag{18}$$

where K is the number of facial expression classes, m_k and N_k is the mean and number of features in class k, and m is the mean of total features.

2.4 Analysis of SVM-based LDA refinement method

It’s necessary to explain why SVM-based LDA could work to refine subtle features without the limitation of learning a noise model. In this subsection, a theoretically analysis is given to explain why our proposed feature refinement method SVM-based LDA can provide a superior margin classifier for subtle facial features classification without suffering the noise disturbance and improve early facial expression recognition performance by cooperating LDA with SVM [10, 30].

Assume that captured facial expression features d_t is a noisy observation signal of a clean signal c_t at frame t so that

$$d_t = c_t + n_t \tag{19}$$

where n_t is the noise at frame t. We also assume that the noise satisfies a Gaussian distribution with mean μ_n and covariance matrix Σ_n. For a two-class SVM, the weight vector ω_{SVM} can be rewritten as,

$$\begin{aligned}\omega_{SVM} &= \omega_{SVM}^{+} + \omega_{SVM}^{-}\\ &= \sum_{i=1}^{N_{sv}^{+}} \alpha_i d_i^{sv} - \sum_{j=1}^{N_{sv}^{-}} \alpha_j d_j^{sv}\end{aligned} \tag{20}$$

where ω_{SVM}^{+} and ω_{SVM}^{-} corresponds to the weighted sum of support vector set in class $y_i = +1$ and class $y_j = -1$, N_{sv}^{+} and N_{sv}^{-} is the total number of support vectors in class $y_i = +1$ and class $y_j = -1$. By substituting (19) into (20), we can obtain

$$\omega = D_c + D_n \tag{21}$$

where $D_c = \sum_{i=1}^{N_{sv}^{+}} \alpha_i c_i - \sum_{j=1}^{N_{sv}^{-}} \alpha_j c_j$ and $D_n = \sum_{i=1}^{N_{sv}^{+}} \alpha_i n_i - \sum_{j=1}^{N_{sv}^{-}} \alpha_j n_j$.

By applying the enhancement transformation matrix ϕ computed by LDA, a new weight vector ω_{LSVM} is obtained as follow,

$$\omega_{LSVM} = \phi^T D_c + \phi^T D_n \tag{22}$$

The weight vector ω_{LSVM} is a solution in the solution set of LDA-based SVM. For better understanding, we restrict ourselves to the 2-class problem, an extension analysis of the multi-class case can be given by an induction way. The enhancement transformation matrix ϕ becomes

$$\phi = S_W^{-1}(m_{c1} - m_{c2}) \tag{23}$$

where m_{c1} and m_{c2} is the mean of facial features in class $y = +1$ and $y = -1$.

In order to compare the classification performance of the ω_{SVM} from SVM and the ω_{LSVM} from LDA-based SVM, we divide the ω_{SVM} into two complementary subspace $V = [\phi] \in \mathbf{R}^{n\times 1}$ and $\bar{V} = [\bar{\phi_1}, \bar{\phi_2}, \ldots, \bar{\phi}_{n-1}] \in \mathbf{R}^{n\times(n-1)}$.

Then the ω_{SVM} from SVM can be divided into four parts as follows,

$$\begin{aligned}\omega_{SVM} &= (VV^T + \bar{V}\bar{V}^T)D_c + (VV^T + \bar{V}\bar{V}^T)D_n\\ &= \phi\phi^T D_c + \phi\phi^T D_n + \bar{V}\bar{V}^T D_c + \bar{V}\bar{V}^T D_n\end{aligned} \tag{24}$$

- The former two parts $\phi\phi^T D_c + \phi\phi^T D_n$ in ω_{SVM} is corresponding to the ω_{LSVM} for LDA-based SVM.
- Because D_c approximates to $\phi = S_W^{-1}(m_{c1} - m_{c2})$ and $\phi \perp \bar{V}$, the third part $\bar{V}\bar{V}^T D_c$ in ω_{SVM} approximates to 0.
- The noise assuming in our study is Gaussian noise. That means the energy of noise would be relatively averagely distributes on each eigenvectors. Therefore, compared to the second part $\phi\phi^T D_n$, the forth part $\bar{V}\bar{V}^T D_n$ contains a much larger component in the covariance information of noise D_n.

According to the above three points, it can be found that $\frac{1}{2}\|\omega_{LSVM}\|^2$ is smaller than $\frac{1}{2}\|\omega_{SVM}\|^2$ in terms of information entropy, as $\frac{1}{2}\|\omega_{LSVM}\|^2$ contains much less noise. And the idea of SVM is to maximize the margin $\|\omega\|^{-1}$. Moreover, D_c which is strongly dependent on class separability of subtle facial features is well preserved in $\frac{1}{2}\|\omega_{LDA}\|^2$. Therefore, it can be concluded that a superior and more reliable margin classifier is obtained by using LDA-based SVM.

3 Experimental results and discussions

In this section, we investigate the early facial expression recognition performance of two proposed facial expression recognition architectures by making numerical experiments. Two comparison results will be given and we also discuss the early recognition improvement of the proposed feature refining methods.

3.1 The collection of experiment dataset

3.1.1 Facial expression dataset

We ask five subjects (five females) to make six universal facial expressions (Happiness, Anger, Surprise, Sadness, Fear, Disgust) as realistically as possible multiple times in front of a high-frame rate 3D motion capture system as shown in Fig. 1. The speed of the motion capture system reaches 100 frames per second, which enables it to capture subtle facial motion. We collected 150 sequences and each expression category contained 25 sequences. The expression sequence varies from a neutral expression to an apex expression. For each sequence, we manually label the facial expression category of every single frame. Especially, we manually judge the earliest frame from that the subject begins to display facial expression (not neutral expression) by watching the facial expression videos several times.

Figure 3 shows a happy sequence in our dataset. We manually label the facial expression category of each frame (Neutral or Happy) and judge the first frame that the subject begins to display happy expression and the frames at which the displaying happiness reach apex intensity. We denote the first frame as T_{exp_begin} and the first frame reaching apex expression

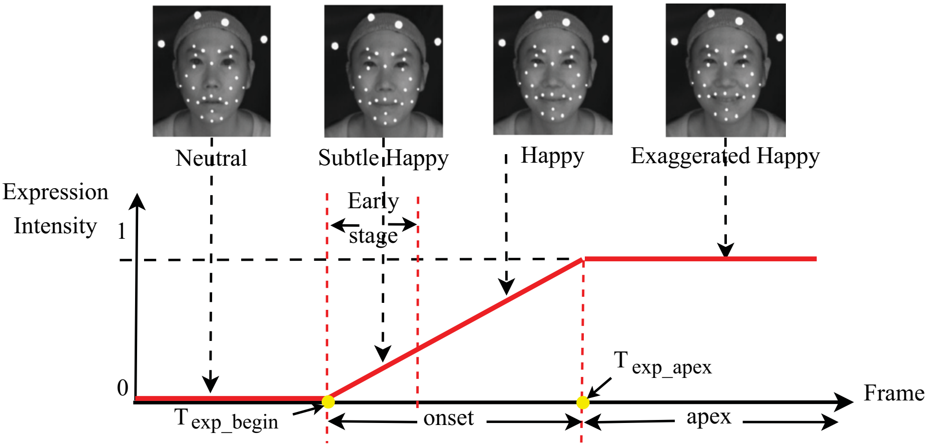

Fig. 3 A happy expression sequence. The expression-begin frame and expression-apex frame are manually labeled

as T_{exp_apex}. In the early expression recognition test, the best results are neutral before T_{exp_begin} and happy after the time T_{exp_begin}.

3.1.2 Noise data collection

In order to refine facial expression features by using adaptive wavelet spectral subtraction, we need to estimate the noise generated from the 3D motion capture system before performing feature refining. We collect four types of noise datasets $\{\Omega_1, \Omega_2, \Omega_3, \Omega_4\}$ from 3D motion capture system. They are static objects, free falling objects, static face fixing on a chinrest with nature expressionless facial motion and rigid facial motion without displaying expressions respectively. The static type dataset can be used to estimate the calibration error and moving type dataset can be used to estimate tracking error. Besides that, the face type dataset can be used to estimate the noise generated from expressionless facial motion. In [21], we have analyzed the refining performance based on these four types of noise datasets and concluded that the noise dataset collected by capturing static face is the best choice to estimate the noise model used in wavelet spectral subtraction, as shown in Fig. 1.

3.2 Comparison results of early facial expression recognition

We evaluate the proposed early facial expression recognition architectures based on the leave-one-subject-out cross validation methodology. In order to compare early recognition results, we temporally normalize the length of each facial expression sequence. The normalized value of the frame T_{exp_begin} is $N(T_{exp\_begin}) = 0$. The frame after T_{exp_begin} is given a positive value and the frame before T_{exp_begin} is given a negative value. The value of the frame T_{exp_apex} is set as $N(T_{exp\_apex}) = 1$. Then, the normalized value of other frames can be linearly computed.

Figure 4 plots the early facial expression recognition result of one raw test sequence displaying "*Fear*". Figures 5 and 6 plots the early facial expression recognition result of its

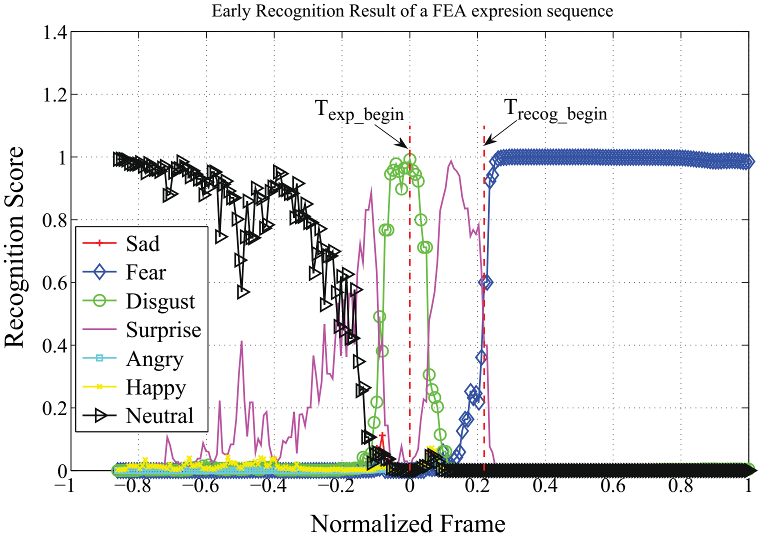

Fig. 4 Early recognition result of one selected original Fear expression sequence F_{Fear}.

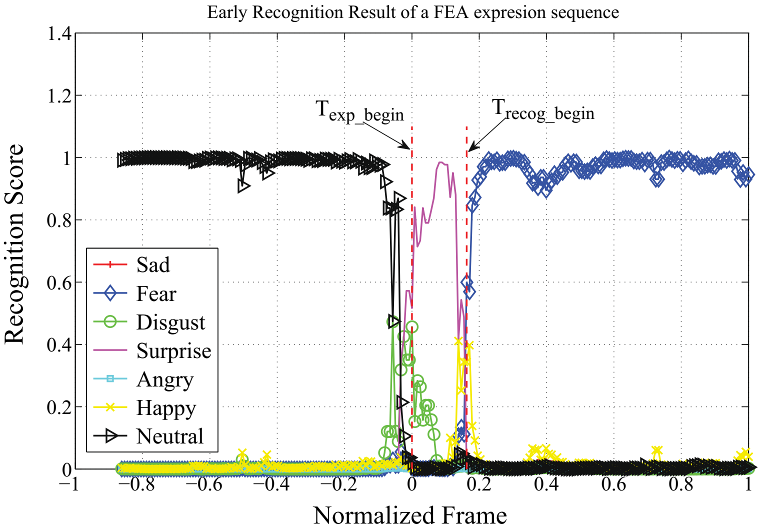

Fig. 5 Early recognition result of one selected refined Fear expression sequence $\hat{F}_{Fear}$ based on adaptive wavelet spectral subtraction

refined test sequence by using adaptive wavelet spectral subtraction and SVM-based LDA. The different color (symbol) line represents the recognition score of different facial expression classifier. As you can see from Fig. 4, the frames around T_{exp_begin} are mis-recognized as "*Disgust*" or "*Surprise*", as the green (with circle symbol) or magenta line obtains the highest recognition score. That is because the noisy facial motion of "*Fear*" in early stage is very similar to the noisy facial motion of "*Disgust*"and "*Surprise*". People usually pull together their brows in the early stage of displaying "*Fear*" and "*Disgust*"and

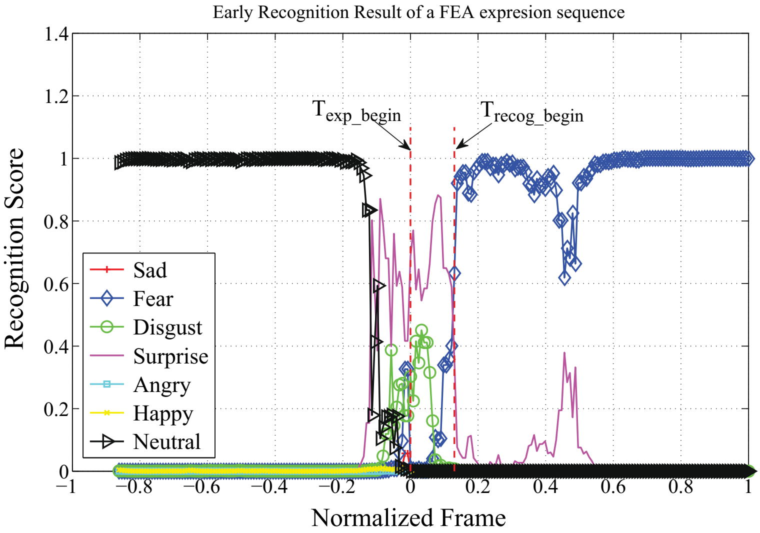

Fig. 6 Early recognition result of one selected refined Fear expression sequence $\hat{F}_{Fear}$ based on LDA

begin to open mouth in the early stage of displaying "$Fear$" and "$Surprise$". As shown in Fig. 5, the recognition results of the frames on the left of T_{exp_begin} frame are obviously improved by using adaptive wavelet spectral subtraction, as the noise model used in adaptive wavelet spectral subtraction successfully estimate the noise which disturb the discriminability between neutral expression and "$Fear$"expression. The early recognition rate of "$Fear$" expression in this sequence is improved. As shown in Fig. 6, the recognition results of the frames on the right of T_{exp_begin} frame are well improved by using LDA, as LDA effectively improve the discriminability between "$Fear$" expression and "$Surprise$"expression. The "$Fear$" expression is earlier recognized in this sequence consequently.

We investigate the performance of early facial expression recognition from two aspects.

- One is how early that the facial expression can be recognized in their early stage. We use the frame T_{exp_begin} as the ground truth of the facial expression occurring frame. The distance between the earliest frame T_{recog_begin} correctly recognized by SVM and the frame T_{exp_begin} is computed to evaluate the "*early*" performance. We denote this distance as early recognition error $error = T_{recog\_begin} - T_{exp\_begin}$.
- The other one is miss recognition rate in the early stage of displaying facial expression. For a real-time human interaction system, it's acceptable to recognize these frames as neutral or the ground truth expression category. However, it should be avoided that these frames are wrongly recognized as other expression categories. The miss recognition rate can be obtained by calculating the amount of miss recognized frames in early stage.

Figures 7 and 8 compares the average early recognition error and miss recognition rate of our proposed methods with the results on raw features and its refined features based on PCA. The "Ori" in Figs. 7 and 8 denotes the recognition results of raw facial features. The experiment results agree with the theoretical analysis. It is clear that the proposed refining method based on wavelet subtraction and SVM-based LDA both improve the early recognition performance. We obtain better early recognition results than using PCA. The proposed wavelets spectral subtraction method well removes the noise in neutral expression which

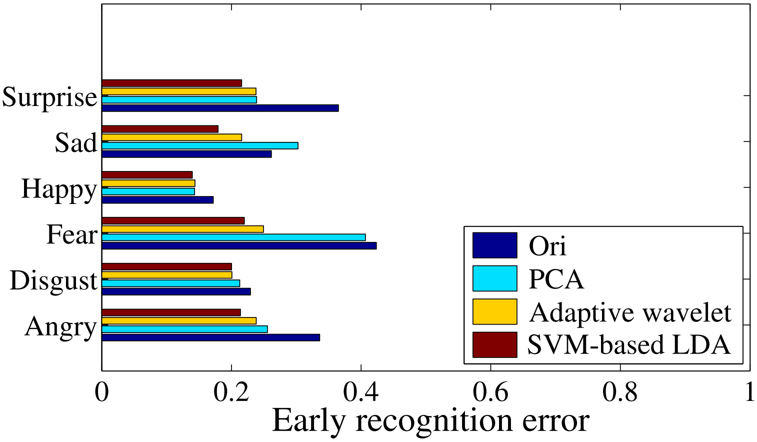

Fig. 7 Early facial expression recognition error. Ori denotes the results of raw facial expression features, PCA denotes the results of the refined features based on PCA, Adaptive wavelet denotes the results of the refined features based on the proposed method adaptive wavelet spectral subtraction and LDA denotes the results of the refined features based on SVM-based LDA

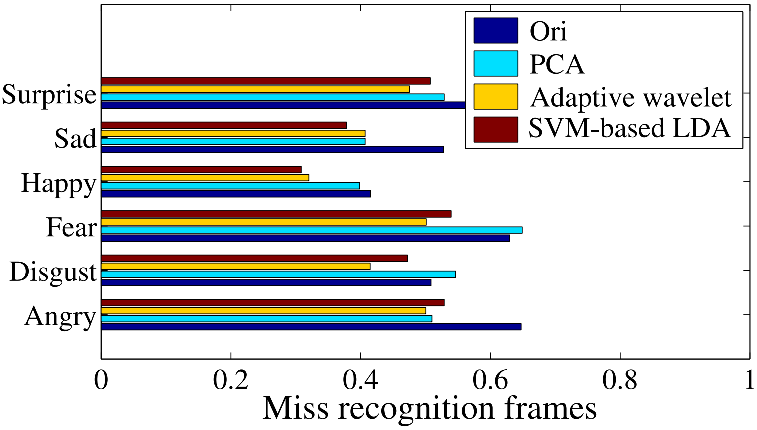

Fig. 8 Miss recognition rate. Ori denotes the results of raw facial expression features, PCA denotes the results of the refined features based on PCA, Adaptive wavelet denotes the results of the refined features based on the proposed method adaptive wavelet spectral subtraction and LDA denotes the results of the refined features based on SVM-based LDA

makes the recognition result of neutral expression more stable. It is helpful to reduce the risk of being recognized as a wrong expression category, as shown in Fig. 8. In Fig. 7, the early recognition error of SVM-based LDA is better than other methods' results. SVM-based LDA refines subtle features by reducing the noise influence and maximize the separability of subtle facial expression features simultaneously which makes the early recognition become earlier.

The experimental results corroborated that both feature refinement methods can successfully reduce the effect of noise for subtle facial expression features and consequently make correct recognitions earlier. The former feature refinement method, adaptive wavelet spectral subtraction, has a prior performance on enhancing the discriminability between neutral expression and other facial expression categories, and consequently reduces the recognition errors. The latter refinement method, SVM-based LDA, has a prior performance on enhancing the discriminability between different subtle facial expression categories, contributing to an earlier recognition of facial expressions.

4 Conclusion and discussion

This paper investigates subtle facial expression recognition problems with a high-frame rate 3D motion capture system, while current facial expression recognition works are based on 2D facial expression or 3D facial expression at a lower frame rate. The 3D facial motion can provide facial deformation along with depth direction which is helpful for analyzing subtle facial motion in early stage. And the motion capture system with high-frame rate can capture the quick facial changes that occur when forming facial expressions in their early stage.

The low intensity of subtle facial expression features (deformations) makes them very sensitive to noise and the noise can easily affect the early recognition result. Conventional facial expression recognition mainly focuses on recognizing obvious facial expressions and often ignores the influence of noise on feature classification. On the other hand, existing feature refinement approaches (such as principal component analysis and filtering) cannot successfully reduce the influence of noise on subtle facial features. This is because they only work when the facial features compose most of the principle components in feature spatial space or the noise is distributed at higher frequencies. In the case of subtle facial expression recognition, the noise is probably nearly equal in intensity to the subtle facial features in spatial domain and is probably distributed at lower frequencies.

Therefore, to alleviate the influence of noise on early facial expression recognition, two feature refinement methods were devised to enhance subtle facial features. One is adaptive wavelet spectral subtraction, which spatial-temporally refines subtle facial expression deformation with an estimated noise model. In particular, a wavelet packet method is used to analyze the spatial-temporal characteristics of the noise in subtle features. To the best of our knowledge, this is the first effort that refines subtle facial features in the spatial-temporal domain. The estimated noise model is then used to adaptively reduce the noise not only at high frequencies but also at low frequencies.

The other subtle feature refinement method is LDA-based support vector machine, which combines the idea of linear discriminant analysis (LDA) with support vector machine (SVM). The SVM-based LDA method refines subtle features by compacting noise and maximizing the class separability of subtle features without requiring a noise model. The margin of the LDA-based SVM can be enlarged, and consequently, the classification performance improves. The final goal of feature refinement is improving the classification performance. Generally speaking, feature refinements are independently performed before feature classification. Therefore, the improvement in classification performance from the feature refinements cannot be directly evaluated. The SVM-based LDA integrating together feature refinement and feature classification can improve the classification performance by directly reducing the influence of noise on feature classification.

Experiments corroborated that the above described feature refinement methods outperform other feature refinement methods by enhancing the discriminability of subtle facial expression features and consequently make correct recognitions earlier. Our current research mainly focuses on processing subtle facial expression feature captured in their early stage. In future work, we will make effort in learning a suitable classifier to completely improve the early recognition performance.

Acknowledgments This research is supported by the Scientific Research Foundation of Xiamen University of Technology (Grant No. YKJ13013R).

References

1. Alexa M (2002) Wiener filtering of meshes. In: Proceedings of international conference on shape modeling and applications, pp 51–60
2. Bishop CM (2006) Pattern recognition and machine learning (Information science and statistics). Springer-Verlag, New York
3. Boehnen C, Flynn P (2005) Accuracy of 3d scanning technologies in a face scanning context. In: Proceedings of the 5th international conference on 3D digital imaging and modeling, pp 310–317
4. Calder AJ, Burton AM, Miller P, Young AW, Akamatsu S (2001) A principal component analysis of facial expressions. Vis Res 41(9):1179–1208
5. Chen CY, Cheng KY (2005) A sharpness dependent filter for mesh smoothing. Comput Aided Geom Des 22(5):376–391
6. Clarenz U, Diewald U, Rumpf M (2000) Anisotropic geometric diffusion in surface processing. In: Proceedings of IEEE visualization, pp 397–405

7. Cohn JF, Zlochower AJ, Lien JJ, Kanade T (1998) Feature-point tracking by optical flow discriminates subtle differences in facial expression. In: Proceedings of IEEE international conference on automatic face and gesture recognition, p 396
8. Desbrun M, Meyer M, Schröder P, Barr AH (1999) Implicit fairing of irregular meshes using diffusion and curvature flow. In: Proceedings of the 26th annual conference on computer graphics and interactive techniques, pp 317–324
9. Ekman P, Rosenberg EL (2005) What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system. Oxford University Press
10. Fransens R, Prins JD, Gool LV (2003) Svm-based nonparametric discriminant analysis, an application to face detection. In: Proceedings of IEEE conference on computer vision, vol 2, p 1289
11. Fukunaga K (1990) Introduction to statistical pattern recognition, 2nd edn. Academic Press Professional, Inc
12. Gao Y, Wang M, Ji R, Wu X, Dai Q (2013) 3d object retrieval with hausdorff distance learning. IEEE Trans Ind Electron 21(4):2088–2098
13. Gao Y, Wang M, Tao D, Ji R, Dai Q, Zhang N (2012) 3d object retrieval and recognition with constructive hypergraph analysis. IEEE Trans Image Process 21(9):4290–4303
14. Lu F, Okabe T, Sugano Y, Sato Y (2014) Learning gaze biases with head motion for head pose-free gaze estimation. Image Vision Comput 32(3):169–179
15. Lu F, Sugano Y, Okabe T, Sato Y (2012) Head pose-free appearance-based gaze sensing via eye image synthesis. In: Proceedings of international conference pattern recognition, pp 2088–2098
16. Lu F, Sugano Y, Okabe T, Sato Y (2014) Adaptive linear regression for appearance-based gaze estimation. IEEE Trans Pattern Analysis and Machine Intelligence 36(10):2033–2046
17. Park S, Kim D (2009) Subtle facial expression recognition using motion magnification. Pattern Recogn Lett 30(7):708–716
18. Poruba J (2002) Speech enhancement based on nonlinear spectral subtraction. In: Proceedings of IEEE conference on devices, circuits and systems, pp T031—1
19. Song M, Wang H, Bu J, Chen C, Liu Z (2006) Subtle facial expression modeling with vector field decomposition. In: Proceedings international conference on image processing, pp 2101–2104
20. Su L, Kumano S, Otsuka K, Mikami D, Yamato J, Sato Y (2010) Subtle facial expression recognition based on expression category – dependent motion magnification. Forum on information technology
21. Su L, Kumano S, Otsuka K, Mikami D, Yamato J, Sato Y (2011) Early facial expression recognition with high-frame rate 3d sensing. In: Proceedings of IEEE conference on systems, man and cybernetics, pp 3304–3310
22. Su L, Sato Y (2013) Early facial expression recognition with early rankboost. In: Proceedings IEEE automatic face and gesture recognition
23. Sun X, Rosin PL, Martin RR, Langbein FC (2008) Noise in 3d laser range scanner data. Shape Modeling International, pp 37–45
24. Sun Y, Yin L (2008) Facial expression recognition based on 3d dynamic range model sequences. In: Proceedings of the 10th European conference on computer vision: part II, pp 58–71
25. li Tian Y, Kanade T, Cohn JF (2005) Recognizing action units for facial expression analysis. Proc IEEE Conf Comput Vis Pattern Recognit 2:568–573
26. Tong Y, Member S, Liao W, Ji Q, Member S (2007) Facial action unit recognition by exploiting their dynamic and semantic relationships. IEEE Trans Pattern Anal Mach Intell 29:1683–1699
27. Van Rhijn A, Mulder JD (2006) Optical tracking and automatic model estimation of composite interaction devices. In: Proceedings of IEEE conference on virtual reality, pp 135–142
28. Wen Z, Huang TS (2003) Capturing subtle facial motions in 3d face tracking. In: Proceedings of IEEE Conference on Computer Vision, pp 1343–1350
29. WLojcicki KK, Shannon BJ, Paliwal KK (2006) Spectral subtraction with variance reduced noise spectrum estimates. In Proceedings of the 11th Australian international conference on speech science and technology, pp 76–81
30. Xiong T, Cherkassky V (2005) A combined svm and lda approach for classification. In: Proceedings of IEEE international joint conference on neural networks, pp 1455–1459
31. Zhang L, Gao Y, Zimmermann R, Tian Q, Li X (2014) Fusion of multichannel local and global structural cues for photo aesthetics evaluation, pp 1419–1429
32. Zhang L, Han Y, Yang Y, Song M, Yan S, Tian Q (2013) Discovering discriminative graphlets for aerial image categories recognition. IEEE Trans Image Process 22(12):5071–5084
33. Zhang L, Song M, Zhao Q, Liu X, Bu J, Chen C (2013) Probabilistic graphlet transfer for photo cropping. IEEE Trans Image Process 22(2):802–815
34. Zhou F, De la Torre F, Cohn J (2010) Unsupervised discovery of facial events. In: Proceedings of IEEE Conf Comput Vis Pattern Recognit:2574–2581

Lumei Su received the B.S and M.S. degrees in control science and engineering from Beijing Institute of Technology University, China, in 2005 and 2007, and the Ph.D. degree in the School of Interdisciplinary Information at the University of Tokyo, Japan, in 2013. She is currently a lecturer in the School of Electrical Engineering and Automation, Xiamen University of Technology. Her current research interests include computer vision, patter recognition, and human computer interaction.

Feng Lu received the BS and MS degrees in Automation from Tsinghua University in 2007 and 2010, and the PhD degree in information science and technology from the University of Tokyo in 2013 respectively. He is currently a project researcher in the Institute of Industrial Science, the University of Tokyo. His research interests include computer vision and patter recognition, especially on human action analysis and HCI.