EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH Open Access

Automatic landmark point detection and tracking for human facial expressions

Yun Tie* and Ling Guan

Abstract

Facial landmarks are a set of salient points, usually located on the corners, tips or mid points of the facial components. Reliable facial landmarks and their associated detection and tracking algorithms can be widely used for representing the important visual features for face registration and expression recognition. In this paper we propose an efficient and robust method for facial landmark detection and tracking from video sequences. We select 26 landmark points on the facial region to facilitate the analysis of human facial expressions. They are detected in the first input frame by the scale invariant feature based detectors. Multiple Differential Evolution-Markov Chain (DE-MC) particle filters are applied for tracking these points through the video sequences. A kernel correlation analysis approach is proposed to find the detection likelihood by maximizing a similarity criterion between the target points and the candidate points. The detection likelihood is then integrated into the tracker's observation likelihood. Sampling efficiency is improved and minimal amount of computation is achieved by using the intermediate results obtained in particle allocations. Three public databases are used for experiments and the results demonstrate the effectiveness of our method.

Keywords: Facial landmark, Kernel correlation analysis, Differential Evolution - Markov Chain

1. Introduction

As computers have become an integral part of our life, the need has arisen for a more natural communication interface between humans and machines. To make human-computer interaction (HCI) more natural and friendly, it would be beneficial to give computers the ability to recognize states of mind of humans the same way a human does. Analyzing facial expression in real time without human intervention will help to understand people's behavior, and thus plays an important role in efficient HCI systems. Automatic facial component localization, such as the eyes, a mouth or nose, is a critical step for expression understanding and emotion recognition [1]. To capture the full range of emotional facial expressions from video sequences, accurate and reliable feature detection and tracking methods are required.

Many researchers have tried to analyze facial expressions by using the distribution of facial features as input of a classification system in order to recognize expressions. However, automatically analyzing facial expressions in video sequences is a challenging task due to the fact that current techniques for the detection and tracking of facial expressions are sensitive to head pose, occlusion, pose, and variations in lighting conditions [2]. In this work, a method based on automatic facial landmark detection and tracking for human expression analysis is proposed. The 26 landmark points shown in Figure 1 display the largest displacements and deformations of the facial components during dynamic changes of the expressions. These points are detected in the facial region by scale invariant feature based detectors, and then tracked through the video sequences using multiple Differential Evolution-Markov Chain (DE-MC) particle filters with kernel correlation techniques. The processing diagram of the proposed method is illustrated in Figure 2.

The rest of this paper is organized as follows. Section 2 presents automatic facial landmark detection. In Section 3, we describe multiple points tracking method with DE-MC particle filters and the kernel correlation technique. The experimental setup and results are presented in Section 4. Finally, Section 5 discuss the results and draw conclusions.

* Correspondence: ytie@ee.ryerson.ca
Ryerson Multimedia Research Laboratory, Ryerson University, Toronto, Ontario, Canada

Figure 1 Illustration of the 26 landmark points selected on facial region.

2. Facial landmark detection

Automatic landmark detection in still image is useful in many computer vision tasks where object recognition or pose determination is needed with high reliability. It aims to facilitate locating point correspondence between images or between images and a known model where natural features, such as the texture shape or location information, are not present in sufficient quantity and uniqueness. Some previous works used shape information for facial feature localization such as template matching [3], graph matching [4], and snakes [5]. These works can detect facial feature well in neutral faces but fail to show good performance in handling large variations such as non-uniform illuminations, change of pose, facial expressions, etc.

Due to the inherent difficulty of detecting the landmark points using a single image, temporal information captured from subsequent frames of a video sequence has been utilized. Detecting and tracking landmark points in video sequences enables computers to recognize affective states of humans, as well as the abilities to interpret and respond appropriately to users' affective feedback [6,7]. We can categorize the landmark detection algorithms in the literature into two groups based on the type of features and anthropometrical information they used, the geometric feature-based methods [8-10] and appearance-based methods [11-13]. The geometric feature-based methods utilize prior knowledge about the face position, and constrain the landmark search by heuristic rules that involve angles, distances, and areas. A number of the existing methods did have success in detecting facial features. For example, [6] used a multi-feature based fusion scheme for facial fiducial point detection and an average of 75% detection rate was achieved, and [8] used Gabor feature based boosted classifier for 20 facial feature point detection, which achieved average recognition rate of 86%. In general, they perform quite well when localizing a small number of facial feature points such as the corners of the eyes and the mouth, however, none of them detects and tracks all the 26 facial landmarks.

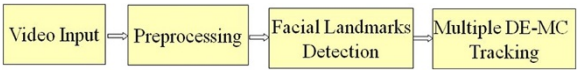

Figure 2 The processing diagram of proposed method.

The appearance-based methods, on the other hand, using image filters such as Gabor wavelets, generate the facial features for either the whole face or specific regions in a face image. The Active Shape Models (ASM) [14] and Active Appearance Models (AAM) [15] are two popular appearance-based methods with statistical face models to prevent locating inappropriate feature points. Cristinacce and Cootes [16] expanded AAM with constrained local models with a set of local feature templates. Milborrow and Nicolls [17] introduced modifications to the ASM with more sophisticated methods. However, these methods were mainly applied to a full face shape model. When the object is small in appearance, cluttered background and occlusion lead to severe ambiguity.

In this section we introduce the scale invariant feature based method for the landmark detection, which includes three steps: *preprocessing, candidate selection and feature vectors extraction*.

2.1 Preprocessing

Since the faces are non-rigid and have a high degree of variability in location, color and pose, it is difficult to detect face automatically in a complex environment. Occlusion and illumination artifacts can also change the overall appearance of a face. We, therefore, propose detecting facial regions in the input video sequence using a face detector with local illumination compensation for normalization and optimal adaptive correlation [18]. Specifically, each frame of the input video sequence is extracted and regularized using an illumination compensation process, including gamma intensity correction (GIC), difference of Gaussian (DoG), local histogram matching (LHM) and local normal distribution (LND). Face candidate regions are then located by the OAC technique with kernel canonical correlation analysis (KCCA). Compare to Viola and Johns' algorithm [19], the local normalization based method is adaptive to the normalized input image and designed to complete the segmentation in a single iteration. With the local normalization based method, the proposed method tends be more robust under different illumination conditions.

Before the raw data sequences can be used for automatic landmark point detection and tracking, it is necessary to normalize the size of the sequence such that they were in the format required by the system. Since the displacement of landmark point in each frame depends on each individual, we use the Inter-ocular Distance (IOD) for size normalization. The distance between left and right eye

pupils is determined in the first input frame. We also manually marked the landmarks for the selected sequences to create the ground truth data.

After the facial region being detected, we propose to use the scale space extrema method to find the locations of candidate points in Section 2.2. The scale invariant feature for each candidate point is extracted and the 26 landmark detectors are constructed as described in Section 2.3.

2.2 Candidates selection

We propose using a scale space extrema method introduced in [20] to detect the locations of interest candidate points in the facial region. The scale space extrema can be detected using the Gaussian kernel function convolved with the input image. The description function $L(x, y)$ of input image in different scale space is expressed as:

$$L(x, y, \sigma) = G(x, y, \sigma) * s(x, y) \tag{1}$$

Where $L(x,y,\sigma)$ is the spatial scale image, $s(x, y)$ indicates input image of facial region, and $G(x,y,\sigma)$ is the Gaussian convolution kernel function defined as:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} exp\left[-\left(x^2 + y^2\right)/2\sigma^2\right] \tag{2}$$

with σ being the scale factor. The image smoothness varies with σ, and a series of scale images is obtained with different σ values. The scale space extrema are computed using the difference of Gaussian (DoG) function of the input image, which calculates the difference of two nearby scales separated by a constant multiplicative factor k, that:

$$\begin{aligned} D(x, y, \sigma) &= [G(x, y, k\sigma) - G(x, y, \sigma)] * s(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \tag{3}$$

where $D(x,y,\sigma)$ is the DoG function of the input image. In this work, we set the interval number n to 3 to form $n+2$ DoG images, and k to $2^{1/3}$. Each pixel in a DoG image is compared to its eight neighbors on the same scale and each of its nine neighbors one scale up and down. If this value is the minimum or maximum among the pixels compared, it is an extremum. These pixels are chosen as interest candidate points, including the adjacent scale, the position and scale of the local extreme point. Since the success of landmark detection depends on the quantity of the selected candidates, we used a larger number of scale samples. (Those points are generally the feature points of the image, located on contours, corners and edges.) DoG extrema are repeatedly assigned in the scale space. They are stable features across all possible scales and are invariant to scale and rotation. These points are highly distinctive and are located on contours, corners and edges in a facial region. Since there are 5 DoG images in our work, all the interest candidates are examined to determine location and scale. The landmarks are detected based on the measurements from these local decisions.

2.3 Feature vectors extraction

After the positions of the interest candidate points are determined from the input image, we choose $\sigma = 1.6$ for the scale, a reasonable compromise between stable extrema detection and computational cost. This value is used throughout this work. A gradient orientation histogram is calculated for the direction of each interest point in its neighborhood. The gradient magnitude $m(x, y)$ and orientation $\theta(x, y)$ are computed using pixel differences, that:

$$\begin{aligned} m(x, y) = \sqrt{\left[L(x+1, y) - L(x-1, y)\right]^2} \\ + \left[L(x, y+1) - L(x, y-1)\right]^2 \end{aligned} \tag{4}$$

$$\theta(x, y) = arctan\left[\frac{L(x+1, y) - L(x-1, y)}{L(x, y+1) - L(x, y-1)}\right] \tag{5}$$

where L is the image at scale σ. We choose a neighborhood F centered at the interest point. By calculating the directions of points in F, we obtain the histogram of gradient directions. The orientation has a range of 360 degrees calculated by Eqs. (4) and (5). However, it is complex and computationally expensive to use the original orientation histogram with 360 bins. To reduce the computing cost, we equally divide the histogram into 36 phases each covering a range of 10 degrees of the orientations. As a result, the orientation histogram has 36 bins. The direction of the interest candidate point is the maximal component of the 36 phases in the histogram.

To detect the landmarks from the interest candidate points, a set of landmark detectors with the feature description from the gradient orientation histogram of the input images are constructed. The descriptor is constructed from a vector containing the values of all the orientation histogram entries. At the center of each landmark, a neighborhood window is selected and divided into 16 subregions of 4×4. Using (4) and (5), the directions and amplitudes of all pixels in the subregions are obtained, and then accumulated into orientation histograms summarizing the contents over the 4×4 subregions. Using the orientation histogram, we can calculate the eight direction distributions in the ranges of (0,π/4,π/2,3π/4,π,5π/4,3π/2,7π/4) with the length corresponding to the sum of the gradient magnitudes near that direction within the region. The amplitude and Gaussian function are also applied on the eight direction distributions to create the direction histogram of subregions. The feature description of each landmark point is obtained by connecting the direction descriptions of all subregions. The total number of the direction descriptions is 16 since we have 4×4 subregions of the landmark descriptor. So the length of a landmark point detector is $128 = 16 \times 8$, and should be normalized in order to ensure illumination invariance.

3. Multiple points tracker

Most tracking algorithms impose constraints on the motion and appearance of objects such as the prior knowledge of motion model, the number and size or the shape of objects. Various approaches have been proposed so far

including the mean-shift, the Kalman filtering and particle filter. The mean-shift based tracker iteratively shifts a data point to the average of data points in its neighborhood, which minimizes the distance between a model histogram representing the target and candidate histograms computed on the current frame. However, it ignores the motion information and is difficult to recover from temporary tracking failures. The Kalman filter is the minimum-variance state estimator for linear dynamic systems with Gaussian noise [21]. For the visual object which moves rapidly, it is hard, in general, to implement the optimal state estimator in closed form [22]. Various modifications of the Kalman filter can be used to estimate the state. These modifications include the extended Kalman filter [23], and the unscented Kalman filter [24]. A multi-step tracking framework was also introduced in [25] to track facial landmarks points under head rotations and facial expressions. The Kalman filter was used to predict the locations of landmarks and a better performance was achieved. However, there are some shortcomings for Kalman filter to track the landmarks of facial expressions, such as the nonlinearity of the head motions, the unimodality of the Kalman, the inherent tracking delay, etc.

Over the last few years, there has been immense attention on particle filters for image tracking because of their simplicity, flexibility, and systematic treatment of nonlinearity and non-Gaussianity. Particle filters provide a convenient Bayesian filtering framework of integrating the detector into the tracker. Based on point mass representations of probability densities, particle filters operate by propagating the particle estimation and can be applied to any state-space model [26-29]. However the sampling results from the proposal density are assigned with low weights and a large number of the particles are wasted in areas with small likelihood. To track the state of a temporal event with a set of noisy observations, the main idea is to maintain a set of solutions that are an efficient representation of the conditional probability. However a large amount of particles that result from sampling from the proposal density might be wasted because they are propagated into areas with small likelihood. Some of the existing works ignore the fact that, while a particle might have low likelihood, parts of it might be close to the correct solution. The estimation of the particle weights does not take into account the interdependences between the different parts of the state of a temporal event.

Particle filter can use multi-modal likelihood functions and propagate multi-modal posterior distributions [30,31]. There are two basic schemes: sending the output of the detector into the measurement likelihood [32,33], or applying a mixture proposal distribution by combining the dynamic model with the output of the detector [34]. However, directly applying particle filter on multiple objects tracking is not feasible because the standard particle filter does not define a way to identify individual modes or hypotheses. Some researchers used sequential state estimation techniques to track multiple objects [35]. Patras and Pantic applied auxiliary particle filtering with factorized likelihoods for tracking of facial points [27]. Zhao et al. [36] introduced a method for tracking of facial points with multi cue particle filter. They have incorporated information from both color and edge of facial features and proposed the point distribution model for constraint tracking results and avoid tracking fails during occlusion. The standard particle filter has a common problem that it turns out to be inadequate when the dynamic system has a very low process noise, or if the observation noise has very small variance [34]. The reason is due to its defective sampling strategy with large dimensionality of the state space. After a few iterations, the particle set will collapse to one single point [31]. Therefore, the resampling method is applied to eliminate particles that have small weights and to concentrate on particles with large weights. It has been realized that improving the resampling or global optimization strategy is more decisive to the success of the tracking [30].

In this paper, we use multiple DE-MC particle filters to track the facial landmarks through the video sequence depending on the locations of the current appearance of the spatially sampled features.

3.1 DE-MC particle filter

The particle filter provides a robust Bayesian framework for the visual tracking problem. It maintains a particle based representation of the a posteriori probability $p(X_k|Y_{1:k})$ of the state X_k given all the observations $Y_{1:k} = \{Y_1, Y_2, \ldots, Y_k\}$ up to and including the current time, k, instance, according to:

$$p(X_k|Y_{1:k}) = \lambda_k p(Y_k|X_k) \int p(X_k|X_{k-1}) p(X_{k-1}|Y_{1:k-1}) dX_{k-1} \tag{6}$$

In (6), the state X_k is a 2 M-component vector that represents the location of landmarks, the observation $Y_{1:k}$ is the set of image frames up to the current time instant. The normalization constant λ_k is independent of X_k. The motion model $p(X_k|X_{k-1})$ is conditioned directly on the immediate preceding state and independent of the earlier history if the motion dynamics are assumed to form a temporal Markov chain. The distribution is represented by discrete samples N through particle filtering. The N samples (particles) are drawn from a proposed distribution $p\left(X_k^{(i)} \middle| X_k^{(i-1)}, Y_k\right)$, $i = 1,2,\ldots,N$ and assigned with weights $w\left(X_k^{(i)}\right)$.

Suppose that at a previous time instance $k-1$, we have a particle based representation of the density, that is, we have a collection of N particles and their corresponding weights $\left\{X_{k-1}^{(i)}, w\left(X_{k-1}^{(i)}\right)\right\}_{i=1}^{N}$. At time step k, select a new set of

samples $\left\{\tilde{X}_k^{(i)}\right\}_{i=1}^N$ from $\left\{X_{k-1}^{(i)}\right\}_{i=1}^N$ with the probability proportional to $w\left(X_{k-1}^{(i)}\right)$. The samples with a larger weight should be selected with a higher probability. Then, applying a constant velocity dynamical model to the samples yields:

$$X_k^{(i)-} = \hat{X}_k^{(i)} + V_{k-1} \tag{7}$$

where $\hat{X}_k^{(i)}$ is a new set of samples selected at time k, and V_{k-1} is the velocity vector computed in time step k-1.

The particle set $\left\{X_k^{(i)-}\right\}_{i=1}^N$ acts as the initial N population for a T-iteration DE-MC processing. For any one landmark in the T-iteration processing, two different inte gers, $r_1 r_2$ that $r_1 \neq r_2 \neq k$, are randomly chosen from the population of previous iteration. A new member $\left\{X_k^{*(i)}\right\}$ that $\left\{X_k^{*(i)}\right\} = \left\{X_{k-1}^{(i)}\right\} + \lambda\left(\left\{X_{k-1}^{(r_1)}\right\} - \left\{X_{k-1}^{(r_2)}\right\}\right) + g$ is created, where λ is a scalar whose value is found to be optimal when $\lambda = 2.38/\sqrt{2N}$, g is drawn from a symmetric distribution with small variance compared to that of $\left\{X_k^{(i)}\right\}_{i=1}^N$. A target function is given based on the ratio between the populations of current and previous step until a convergence or a preset end point is reached. Then the weights of particles are subject to update by the DE-MC. At the end of this step, we take the output population as the particle set of current time step $\left\{X_k^{(i)}, w\left(X_k^{(i)}\right)\right\}_{i=1}^N$.

We estimate the state at time step k as:

$$X_k = argmax_{X_k^{(i)}; i=1,\ldots,N} w\left(X_k^{(i)}\right) \tag{8}$$

and update the velocity vector of current time step $V_k = X_k - X_{k-1}$. The step size of random jumping for current DE-MC iteration is reduced if the survival rate of the last DE-MC iteration is high or inflated otherwise [37]. The update scheme for the maximum likelihood decision on the weights w can be summarized as follows:

Starting from the set of particles which are the filtering result of time step $k - 1$: $\left\{X_{k-1}^{(i)}, w\left(X_{k-1}^{(i)}\right)\right\}_{i=1}^N$.

1. Selection: select a set of samples $\left\{\hat{X}_k^{(i)}\right\}_{i=1}^N$ from $\left\{X_{k-1}^{(i)}\right\}_{i=1}^N$ with the probability proportional to $w\left(X_{k-1}^{(i)}\right)$.
2. Prediction and Measurement: Apply a constant velocity dynamical model to the samples using Eq. (7). At the end of this step, we take the output population as the particle set of current .time step that $\left\{X_k^{(i)}, w\left(X_k^{(i)}\right)\right\}_{i=1}^N$.
3. Representation and Velocity Updating: Estimate the state at time step k by Eq. (8) and update the velocity vector of current time step.

While the tracker updates and tracks the X_k vector that represents the coordinates of the 26 landmark points, the samples are already drawn. The DE-MC particle filter is able to make a more reasonable sampling and keeps them from running off into implausible shapes even if they are placed in the positions far away from the solution point or are trapped in the local cost basin of the state space. The observation model can help the sample points for positions close to the solution in regard to their starting points. The measurement module provides necessary feedback to the sampling module, according to which, the hypothesis moves to the regions where it is more likely for the global maximum of the measurement function to be found.

3.2 Kernel correlation-based observation likelihood

The kernel correlation based on Hue Saturation Value (HSV) color histograms is used to estimate the observation likelihood and measure the correctness of particles, since HSV decouples the intensity (value) from color (hue and saturation) and corresponds more naturally to human perception [38]. We set each feature point at the centre of a window as the observation model. The kernel density estimate (KDE) $K(X_k)$ for the color distribution of the object X_k at time step k is given as:

$$K(X_k; r) = \frac{1}{\zeta}\sum_{i=1}^{N} \frac{\left(c\left(X_k^{(i)}\right) - c(r)\right)}{d^{i_x}} \tag{9}$$

where the $c(.)$ function is a three dimensional vector of HSV and $c\left(X_k^{(i)}\right)$ can be generated from the candidate region within a search region R centered at X_k at time step k. It should be sufficiently large to reach the maximum facial point movement without overlapping with any neighboring windows. $c(r)$ can be generated from the target region, which is r position translation in the search region R. The normalizing constant ζ ensures K $(X_k;r)$ to be a probability distribution, $\sum_{k=1}^{N} K(X_k; r) = 1$. The kernel width d^{i_x} is used to scale the KDE $K(X_k;r)$, and the optimal solution for kernel width d^{i_x} that minimizes the Mean Integrated Square Error (MISE) [39] is given by:

$$d_{opt} = \left(\frac{4}{(i_x + 2)N}\right)^{1/(i_x+4)} \tag{10}$$

where i_x is the number of particles in the set at time k and d_{opt} denotes the optimal solution for d^{i_x}. If we denote K^* $(X_k;r)$ as the reference region model and $K(X_k;r)$ as a candidate region model, we can measure the data likelihood to track the facial point movements by considering the maximum value of the correlation coefficient between the color histograms in this region and in a target region. The correlation coefficient $\rho(X_k)$ is calculated as:

$$\rho(X_k) = \left| \frac{\sum_{i=1}^{N} \sum_{r \in R} |K^*(X_k;r) - E(K^*(X_k;r))||K(X_k;r) - E(K(X_k;r))|}{\sqrt{\sum_{i=1}^{N} \sum_{r \in R} |K^*(X_k;r) - E(K^*(X_k;r))|^2} \sqrt{\sum_{i=1}^{N} \sum_{r \in R} |K(X_k;r) - E(K(X_k;r))|^2}} \right| \tag{11}$$

where $E(K(X_k;r))$ is the means of the vectors $K(X_k;r)$ and $K^*(X_k;r)$, and $E(K^*(X_k;r))$ is the average intensities of the color model. Finally, we define the observation likelihood of the color measurement distribution using the correlation coefficient $\rho(X_k)$ that:

$$p\left(Y_k | X_k^{(i)}\right) = e^{\frac{\rho^2\left(X_k^{(i)}\right)}{\tau_i}} \tag{12}$$

where τ_i is a scaling parameter, which helps the result evaluated by (12) be more reasonably distributed in the range of (0,1).

3.3 Landmark point tracking

In this section, we present using multiple DE-MC filters for facial landmarks tracking over time. Once the observation model is defined we need to model the transition density and to specify the scheme for reweighting the particles. The single particle filters weight particles based on a likelihood score and then propagate these weighted particles according to a motion model. Simply running particle filters for multiple landmarks tracking needs a complex motion model for the identity between targets. Such an approach suffers from exponential complexity in the number of tracked targets [40]. In contrast to traditional methods, our approach addresses the multi-target tracking problem using the M-component non-parametric mixture model, where each component (every landmark point) is modeled with an individual particle filter that forms part of the mixture. The landmark states have multi-modal distribution functions and the filters in the mixture interact only through the computation of the importance weights. In particular, we combined color based kernel correlation technique for the observation likelihood with DE-MC particle filtering distribution. A set of weighted particles are used to approximate a density function corresponding to the probability of the location of the target given observations.

To avoid sampling from a complicated distribution, the M-component model is adopted for the posterior distribution over the state X_k of all targets M according to:

$$p(X_k | Y_{1:k}) = \sum_{j=1}^{M} Pi_{j,k} p_j(X_k | Y_{1:k}) \tag{13}$$

where $M = 26$, $p_j(X_k | Y_{1:k})$ is the posteriori probability of the facial landmarks with the M-component non-parametric mixture model, and Pi is the mixture weights satisfy $\sum_{m=1}^{M} Pi_{m,k} = 1$. We utilize training data to learn the interdependencies between the positions of the facial landmarks for the reweighting scheme. It is clear that the performance can be improved if we consider the motion models of the landmark points. The motion model $p(X_k | X_{k-1})$ predicts the state X_k given the previous state X_{k-1}. Using the filtering distribution computed from (13), the predictive distribution becomes:

$$p(X_k | Y_{1:k-1}) = \sum_{j=1}^{M} Pi_{j,k-1} p_j(X_k | Y_{1:k-1}) \tag{14}$$

where $p_m(X_k | Y_{1:k-1}) = \int p(X_k | X_{k-1}) p_m(X_{k-1} | Y_{1:k-1}) dX_{k-1}$. The likelihood $p(Y_k | X_k)$ is the measurement model and expresses the probability of observation Y_k. We approximate the posterior from an appropriate proposal distribution to maintain a particle based representation for the a posteriori probability of the state. It provides a consistent way to resolve the ambiguities that arise in associating multiple objects with measurements of the similarity criterion between the target points and the candidate points. The updated posterior mixture takes the form that:

$$p(X_k | Y_{1:k}) = \sum_{j=1}^{M} Pi_{j,k} p_j(X_k | Y_{1:k}) = \lambda_k \sum_{j=1}^{M} Pi_{j,k} p_j(Y_k | X_k) \int p_j(X_k | X_{k-1}) p_j(X_{k-1} | Y_{1:k-1}) dX_{k-1} \tag{15}$$

The new weights can be approximated with a prior on the relative positions of the facial features as:

$$Pi_{j,k} = \frac{Pi_{j,k-1} \int p_j(Y_k | X_k) p_j(X_k | Y_{0:k-1}) dX_k}{\sum_{l=1}^{M} Pi_{l,k-1} \int p_l(Y_k | X_k) p_l(X_k | Y_{0:k-1}) dX_k} \tag{16}$$

The particles are sampled from the training data to obtain the appropriate distribution in the M-mixture model. The prediction step and the measurement step are integrated together instead of functioning separately. The use of the priors provides sufficient constrains for reliable tracking at the presence of appearance changes due to facial expressions. The measurement function evaluates the resemblance between image features generated by hypothesis and those generated by ground truth positions, as the criterion for judging the correctness of hypothesis.

When tracking the multiple modalities, multiple trackers start with mode-seeking procedure, the posterior modes are subsequently detected through the HSV color histograms based kernel correlation analysis. Using a trained color-based observation model allows us to track different landmark points. Here, we have M different likelihood distributions. At time k we sample candidate particles from an appropriate proposal distribution $\left\{\widehat{\mathrm{X}}_{k-1}^{(i)}\right\}_{i=1}^{N}$ from $\left\{X_{k-1}^{(i)}\right\}_{i=1}^{N}$ and weight these particles according to the probability proportional:

$$w_k^{(i)} = w_{k-1}^{(i)} \frac{p\left(Y_k|\widehat{\mathrm{X}}_k^{(i)}\right) p\left(\widehat{\mathrm{X}}_k^{(i)}|X_{k-1}^{(i)}\right)}{p\left(\widehat{\mathrm{X}}_k^{(i)}\middle|X_{0:k-1}^{(i)}, Y_{1:k}\right)} \tag{17}$$

In our work, scaling is normalized by person-related scaling factors that are estimated from the positions of the facial features at the first frame, such as the dimensions of the mouth. This scheme simply processes with the prior knowledge by sampling from the transition priors and updating the particles using importance weights derived from (17).

4. Experiments and results

To evaluate the system performance of the proposed detection and tracking method for facial expression, we construct an experimental dataset from three publicly available databases: RML Emotion database [9], Cohn-Kanade (CK) database [41] and Mind Reading (MR) database [42]. The RML Emotion database was originally recorded for language and context independent emotional recognition with the six fundamental emotional states: happiness, sadness, anger, disgust, fear and surprise. It includes eight subjects in nearly frontal view (2 Italian, 2 Chinese, 2 Pakistani, 1 Persian, and 1 Canadian) and 520 video sequences in total. Each video pictures a single emotional expression and ends at the apex of that expression while the first frame of every video sequence shows a neutral face. Video sequences from neutral to target display are digitized into 320 × 340 pixel arrays with 24-bit color values. The CK database consists of approximately 2000 image sequences in nearly frontal view from over 200 subjects. Each video pictures a single facial expression and ends at the apex of that expression while the first frame of every video sequence shows a neutral face. The MR database is an interactive computer-based resource for face emotional expressions, developed by Cohen and his psychologist team. It consists of 2472 faces, 2472 voices and 2472 stories. Each video pictures the frontal face with a single facial expression of one actor (30 actors in total) of varying age ranges and ethnic origins.

We select 320 videos of eight subjects from the RML Emotion database, 360 image sequences of 90 subjects from CK database and 360 videos of 30 subjects from MR database for the experiments. As a result, the experimental dataset includes 1040 image sequences of 128 subjects in total. The experiments are implemented on a Quad CPU 2.4 GHz PC with 3.25 GB memory, under the Windows XP operating system.

We compare the automatically located facial landmarks with the ground truth points to evaluate the performance of the detection and tracking method. In general, the detecting and tracking methods are usually regarded as a SUCCESS if the bias of the automatic labeling result to the manual labeling result is less than 30% of the true inter-ocular distance [43]. However, this is unacceptable in the case of facial expression analysis. To follow the subtle changes in the facial feature appearance, we define a SUCCESS case if the bias of a detected point to the true facial point is less than 10% of inter-ocular distance in the test image. The one-against-all (OAA) and leave-one-subject-out (LOSO) cross validation strategies are utilized to perform the experiments. The OAA strategy works as follows: for each time, one sample is held out as the testing data, while the rest of the data in the entire dataset is used as the training data. This procedure continues until all the individual samples in the entire dataset have been held out once. In the LOSO strategy, the samples belonging to one subject are used as the testing data and the remainders as the training data. This is also repeated for all of the possible trials until all the subjects are used as the testing data. There is no overlap between the training and testing subjects. The experimental results are averaged as the final accuracy.

4.1 Facial landmark detection

In this section, we present the experimental results using the proposed facial landmarks detection method. Adaboost algorithm is applied for training the 26 facial landmark detectors. We use ten frames from each training sequence with the manually labeled ground truth points. The surrounding eight positions of the true point are also selected as the positive examples in a training image. Another five arbitrary points in the same frame are chosen as the set of negative examples. The prototypical 128-dimensional feature vector is used for each sample point. In the testing images, candidate points are first extracted from facial region using scale invariant feature. For a certain facial landmark, Adaboost classifier outputs a response depicting the similarity between the representations of the candidate points compared to the learned training model. After checking the entire facial region, the position with the highest response reveals the landmark point.

Boost algorithm has been proposed to reduce the redundancies of the high dimensional feature space and computational cost. The Adaboost algorithm by Viola and Jones [19] for face detection is a typically successful

example as it has a very low false positive rate and can detect faces in real time. It can be trained for different levels of computational complexity, speed and detection rate which are suitable for specific applications. The performances of RealAdaboost [44], GentleAdaboost [45] and ModestAdaboost [46] for fiducail point detectors are compared in our work using GML AdaBoost Matlab Toolbox [47] and shown in Figure 3. GentleAdaboost returns the best detection rates from the results. In contrast to other Adaboost algorithms, GentleAdaboost uses real valued features and converges faster. It gives less emphasis to misclassified examples since the increase in the weight of the example is quadratic in the negative margin, rather than exponential. Thus, GentleAdaboost is selected as the classification algorithm in our system.

The overall detection rates for each point are shown in Figure 4, and the proposed method achieves 91% average detection rate of the facial landmarks. We illustrate some representative cases in Figure 5. The proposed method is applied on each frame of the input video sequences, and the 26 facial landmarks are automatically detected.

4.2 Tracking results

In this section, we present the experimental results using the proposed multiple DE-MC filters. The positions of the facial landmarks in the first frame of an input sequence are automatically found using the detection method. The positions in all subsequent frames are then determined by the multiple particle filters with the color based observation likelihood. The observation model is built from the training data of manually labeled sequences using a finite set of particles within the feature point centered window. We approximate the posterior $p(X_k|Y_{1:k})$ from an appropriate proposal distribution to maintain a particle based representation for the a posteriori probability of the state.

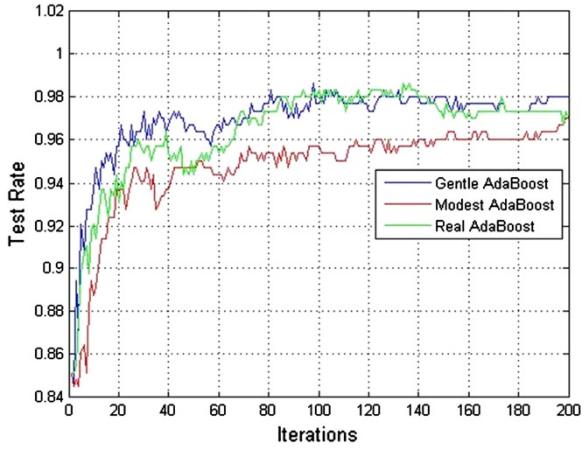

Figure 3 Test rates from Adaboost algorithms with 200 boosting iterations.

Since the calculation of the weights of the particles is a critical step of multiple points tracking, in the proposed M-mixture model, we sample the particles from the training data to obtain the appropriate proposal distribution. The proposed method simply proceeds by sampling from the transition priors and updating the particles using importance weights derived from Eq. (17). In the DE-MC iterations, the measurement module provides necessary feedbacks to the sampling module. According to them the sampling moves to regions in the state space where it is more possible to find the global maximum of the measurement function. Since we are interested in the global optimal state, we place denser sampling grids in the region of interest. This approach yields a result reasonably close to that obtained by sampling strictly according to the ground truth posterior distribution.

We present some representative cases using the proposed method, exploring various practical aspects for the facial landmark detection and tracking. Figures 6 and 7 summarize the experimental results for two different emotional expressions. The facial landmarks are first detected by the point detectors in the first frame and then tracked by the kernel correlation based multiple DE-MC filters. For all figures, the white dots represent the positions of facial landmarks to be detected and tracked, which are all labeled with the associate numbers. In Figure 6, the subject exhibits a set of sadness expressions from a neutral face at the beginning and ends at the apex of that expression. Figure 7 shows the anger expression with talking at the same time. As expected, all the points are tracked reliably for the whole sequence. Since the motions of the faces are not intensive and the facial appearances are not heavily changed, the features extracted from consecutive frames are highly correlated and the results achieve a very impressive tracking rate.

We apply the proposed method to the zoomed case, as shown in Figure 8. When the camera zooms, the factors assigned with the color based kernel correlation keep changing and are going to descend, as a result of (9) and (10) which can be seen from frame 320. However, the facial landmark can still be tracked with the updating weights using (16), as we keep track of the points from the previous frame. It shows that the use of the priors for the multiple filters provides constraints that are sufficient for the reliable tracking of the points at the presence of the facial appearances.

While performing the experiments, we also consider the cases with head in rotation or point occlusion, as shown in Figure 9. In this case we can see the points 3, 7, 15, 20 and 23 are lost after frame 78 when a frontal face is rotating to a profile view. So far, the multiple detectors and trackers are based on different configurations of color intensity regions. If both detectors and particle trackers fail

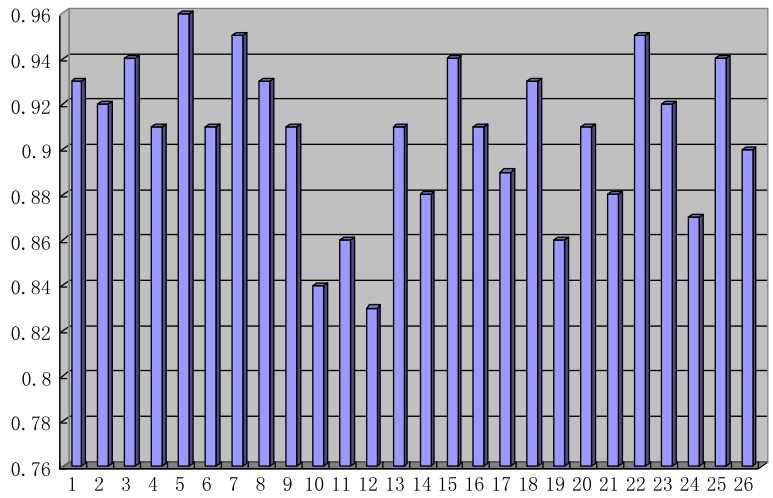

Figure 4 The final facial landmark detection rate.

for several consecutive frames, the proposed approach will eventually lose the landmark points.

To solve this problem, we execute a conservative way to update the trackers temporally with the response distribution [48] for the next n frames when the missing points first occurred. This step length n can be changed by the user and should not be crucial to the system. If the trackers respond correctly after a few frames, the trackers are able to recover due to the accumulation of probabilities. However, when the step length n continues to grow, due to incorrect responses of the detector, the color correlation of the observation likelihood drops and the trackers will begin to lose points. After that, "point lost" will be declared. We then stop estimating its motion V_k and discard the motion likelihood term. The trackers will be reinitialized by the point detectors in the following frames. All the 26 points can be detected with a new set of parameters if the facial region appears again in the scene. The improved

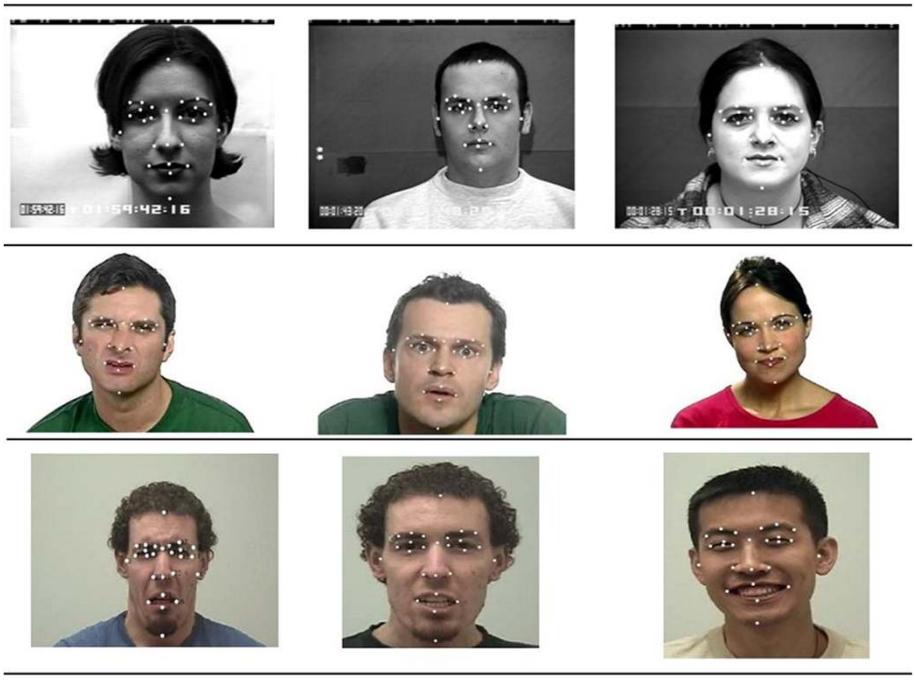

Figure 5 Sample sequences from the test videos for facial point detection.

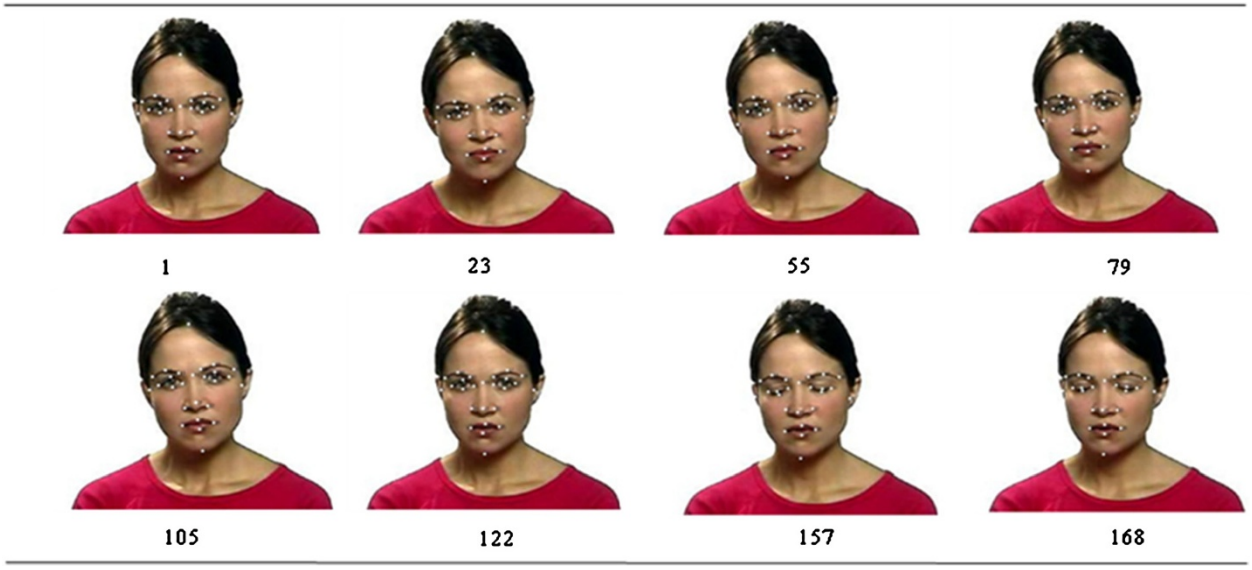

Figure 6 Sample sequences for sadness facial expression. The frame numbers are marked below.

result is shown in Figure 10 that reinitialization executes and all facial landmarks are found again after frame 183.

4.3 Performance evaluate

To evaluate the performance of the detection and tracking method for emotional expressions, we use recall and precision as the performance measures. The missing rates and false alarms are conducted by comparison between the output and the SUCCESS point, which is defined as:

$$\begin{aligned} recall &= \frac{N_{\text{SUCCESS}}}{N_{\text{SUCCESS}} + N_{miss}} \times 100\% \\ precision &= \frac{N_{\text{SUCCESS}}}{N_{\text{SUCCESS}} + N_{false}} \times 100\% \end{aligned} \tag{18}$$

where N_{SUCCESS} stands for the number of SUCCESS point from the detection and tracking, N_{miss} stands for the number of missed points, and N_{false} stands for the number of

Figure 7 Sample sequences for anger facial expression with talking simultaneously. The frame numbers are marked below or sadness facial expression. The frame numbers are marked below.

Figure 8 Sample sequences for the zoomed case. The frame numbers are marked below.

false alarms. The sum $N_{\text{SUCCESS}} + N_{miss}$ is the total number of manually labeled facial landmarks in the entire video sequence.

The overall performance of the system in term of false alarm rate using the aforementioned datasets is illustrated in Figure 11. From this figure, we can see that the precision is decreasing and recall is increasing with the increment of false alarms. Note, in the graph, a system performance of recall 94.15% and precision 92.86% is achieved simultaneously.

We also checked the displacement accuracy for the proposed methods. The Euclidean distances between each individual landmark point are used for the measurement. Since we use the scale normalization for the variation in size of each individual face, therefore the distance measurement is invariant for the experimental datasets. We calculate the average accuracy for the displacement of the proposed automatic method compared to the manually labeled ground truth. The distributions of the displacement accuracy are shown in Figure 12. From the figure we can see that given a 10% normalized distance, the proposed method achieves a 93% average accuracy from the ground truth.

4.4 Comparison with state-of-the-art

To distinguish person-independent affective states, subtle changes of facial expressions should be extracted for feature construction. Automatic facial landmark

Figure 9 Sample sequences for the head's rotation case. The frame numbers are marked below.

Figure 10 The improved sample sequence for the head's rotation case. The frame numbers are marked below.

detection and tracking are crucial for analyzing the current facial appearance since it will facilitate the examination of the fine structural changes inherent in the spontaneous expressions. A key motivation for developing landmark point techniques is that they lay the foundation for developing 3D models and associated dynamic feature extraction and recognition techniques which are highly likely superior to 2D-based and static 3D-based techniques. We therefore first compare with the result reported in [9] which also used the RML Emotion database, but with static visual features extracted by 2D Gobar filters. The comparison shows that working on the same database, facial landmark based 3D dynamic features [49] (90% recognition rate) substantially outperforms the recognition rate by the 2D Gabor features (approximately 50%), and also the bimodal features (approximately 82%).

In general, some existing methods perform quite well when localizing a small number of facial feature points such as the corners of the eyes and the mouth, however, none of them detects and tracks all the 26 facial landmarks illustrated in Table 1. To present a straightforward comparison with state-of-the-art, we conduct extensive experiments using two publicly available face databases, BIOID database [50] and BUHMAP database [11], with manually marked ground truth positions. Table 1 summarizes the comparative experimental results along with that from some state-of-the-art methods on the same test sets. The results from other methods are taken from expanded AAM [14], factorized PF [29], SIR PF [51] and Gabor feature PF [52].

As is evident from these results, our method achieves the best overall performance of 90.8% average rate. In

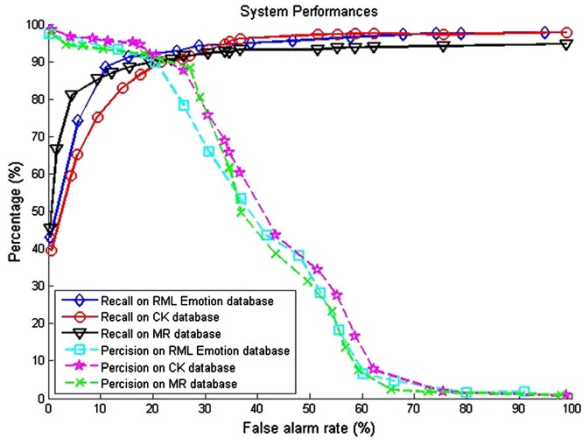

Figure 11 Recall and precision against false alarm rate for the test databases.

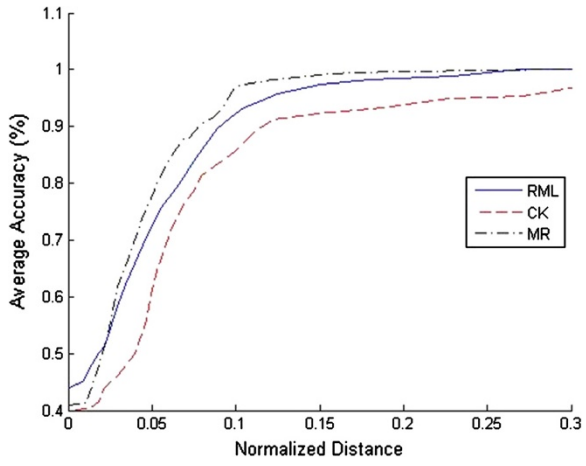

Figure 12 Displacement accuracy based on the normalized distance.

Table 1 Comparisons based on different public databases

Fiducial points	Proposed		Expanded AAM [14]		Factorized PF [29]		SIR PF [51]		Gabor PF [52]	
	BIOID	BUHM AP	BIOID	BUHM AP	BIOID	BUHM AP	BIOID	BUHM AP	BIOID	BUHM AP
P1	92.89	91.97	85.45	86.13	83.66	83.27	81.35	79.19	87.42	89.73
P 2	94.68	93.06	87.15	88.56	84.81	82.99	79.64	80.61	86.25	86.44
P 3	93.33	89.56	84.21	83.68	79.40	76.72	74.39	75.65	82.91	80.19
P 4	90.94	91.76	84.48	81.95	78.38	78.49	76.34	73.71	82.03	84.97
P 5	95.31	94.28	90.33	89.95	82.67	82.01	79.08	80.14	89.50	90.32
P 6	88.86	89.59	80.94	81.38	77.78	74.91	75.12	72.92	79.74	80.63
P 7	94.99	93.47	88.34	87.45	82.40	82.74	82.94	81.68	80.42	81.06
P 8	89.33	88.45	83.47	81.97	79.61	74.96	76.27	75.54	79.24	78.62
P 9	96.01	94.73	91.04	90.15	81.92	82.59	79.35	76.54	81.48	79.20
P 10	86.31	87.14	79.69	80.41	74.02	79.14	76.05	79.13	79369	79.06
P 11	89.03	90.02	86.63	87.37	82.33	81.86	85.02	82.46	85.24	83.15
P 12	85.12	86.24	80.31	81.06	75.21	75.83	73.66	76.98	79.45	79.13
P 13	91.92	93.10	86.69	87.81	82.70	83.51	81.12	83.06	81.28	83.67
P 14	84.97	83.15	79.62	78.38	78.26	77.44	78.45	77.96	79.66	81.71
P 15	91.24	92.45	84.71	84.55	82.83	81.34	76.52	75.43	86.01	86.93
P 16	89.56	88.74	85.35	86.16	78.25	79.59	78.05	79.29	86.36	82.98
P 17	82.49	86.35	79.22	81.74	76.17	78.79	74.62	73.64	81.31	84.52
P 18	89.45	90.12	86.08	85.15	81.93	83.58	80.23	80.71	85.51	85.78
P 19	88.94	90.84	87.11	87.84	80.81	82.14	76.03	75.34	84.18	86.19
P 20	91.03	93.21	82.21	81.16	79.85	79.62	76.82	78.02	84.45	85.56
P 21	89.62	88.82	81.74	81.06	81.48	84.44	82.82	79.12	79.82	80.43
P 22	93.87	92.53	83.94	80.09	85.30	86.52	79.99	79.21	85.28	86.58
P 23	96.40	94.77	86.37	85.52	86.23	85.17	84.80	82.95	86.34	84.42
P 24	90.85	91.06	85.81	81.03	79.41	78.30	79.27	78.13	84.73	83.96
P 25	94.97	95.43	89.24	86.11	83.22	83.80	80.59	80.62	88.27	86.48
P 26	91.67	90.28	83.67	84.84	76.53	79.43	76.27	78.74	77.75	78.58
Ave.	90.86%		84.51%		80.66%		78.58%		83.35%	

contrast with other approaches, the most evident improvement of the proposed method is that the prediction step and the measurement step are integrated together instead of functioning separately. The use of the priors provides sufficient constrains for reliable tracking at the presence of appearance changes due to facial expressions. The measurement function evaluates the resemblance between image features generated by hypothesis and those generated by ground truth positions, as the criterion for judging the correctness of hypothesis.

The proposed method has demonstrated its ability to handle pose variations problems and can be used for both image and video based facial expression recognition. Computationally, the proposed method has the advantages of automatic initialization by using the scale invariant features extraction over the other methods that examine pixels one by one. Note that the method proposed in [27] achieved a better overall detection rate. However, this method is only tested on perfect manually aligned image sequences and no experiments in fully automatic conditions were reported. In addition, only 13 sequences were experimented on in [27]. Therefore, the result is far from conclusive.

5. Discussions and conclusions

Automatic facial landmark detecting and tracking is a challenging task in facial expression analysis. In this paper, we proposed an automatic approach to detect and track facial landmarks for varying facial expressions. We first construct a set of facial landmark detectors with scale invariant feature. Locating feature points automatically on a single frame makes it possible to eliminate the manual initiation step for the tracking algorithm.

We also adopt the multiple DE-MC filters for facial landmarks tracking. Compared with the existing multi-target tracking methods, such as the joint probabilistic data association filter (JPDAF) [53], moving horizon estimation [54], various modifications of the Kalman filter [55], or the interior point approaches [56], the DE-MC particle filter leads to a more reasonable approximation to

the proposal distribution. It incorporates the advantage of the Differential Evolution algorithm in global optimization and the ability of the Monte Carlo Markov Chain in reasonably sampling a high-dimensional state space. It evidently boosts the performance of the traditional tracking method in terms of more accurate motion vector prediction. Based on the fact that the posterior depends on both the previous state and the current observation in a visual tracking application, the DE-MC particle filter can also considerably improve the accuracy for tracking by building a path connecting a sampling with measurement. Taking the advantage of the DE-MC algorithm's ability, we can obtain reasonably distributed samples that are concentrated on important regions of the state space. A novel Kernel correlation with robust color histograms is proposed for the observation likelihood to deal with changes in the facial appearance of different expressions.

Furthermore, the facial landmarks are tracked by utilizing prior knowledge on the facial feature configurations. It provides a consistent way to resolve the ambiguities that arise in associating multiple objects with measurements of the similarity criterion between the target points and the candidate points. Instead of simply applying the single DE-MC filter for multiple point tracking, we utilize the M-component non-parametric mixture model for the multiple DE-MC filters' posterior distribution over the states of all target points. This approach yields a result reasonably close to that obtained by sampling strictly according to the ground truth posterior distribution.

For future work, we plan to improve the detection and tracking performance and extend our real-time algorithm to cope with both self and other forms of occlusions.

Competing interests

The authors declare that they have no competing interests.

Received: 3 November 2011 Accepted: 8 January 2013
Published: 4 February 2013

References

1. AA Salah, H Cinar, L Akarun, B Sankur, Robust facial landmarking for registration. Annals of Telecommunications **62**(1), 1608–1633 (2007)
2. Z Zeng, M Pantic, GI Roisman, TS Huang, A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. IEEE Transactions on Pattern Analysis and Machine Intelligent **31**(1), 39–58 (2009)
3. R Brunelli, T Poggio, Face Recognition: Features Versus Templates". IEEE Trans. Pattern Anal. Mach. Intell. **15**(10), 1042–1062 (1993)
4. R Herpers, G Sommer, *An Attentive Processing Strategy for the Analysis of Facial Features* (Springer-Verlag, Berlin Heidelberg, New York, 1998)
5. M Pardas, M Losada, Facial Parameter Extraction System Based on Active Contours. International Conference on Image Processing, Thessaloniki **1**, 1058–1061 (2001)
6. HC Akakin, B Sankur, Multi-attribute robust facial feature localization. Automatic Face & Gesture Recognition, 1–6 (2008)
7. I Cohen, N Sebe, A Garg, MS Lew, TS Huang, Facial expression recognition from video sequences. Proceeding of international conference on Multimedia and Expo **2**, 121–124 (2002)
8. M Pantic, LJM Rothkrantz, Facial action recognition for facial expression analysis from static face image. IEEE Transactions on Systems, Man, and Cybernetics-Part B **34**, 1449–1461 (2004)
9. LCD Silva, SC Hui, Real-time facial feature extraction and emotion recognition. Proceedings of the 4th International Conference on Information Communications and Signal Processing **3**, 1310–1314 (2003)
10. K Anderson, PW McOwan, A real-time automated system for the recognition of human facial expressions. IEEE Transactions on Systems, Man, and Cybernetics Part B **36(1)**, 96–105 (2006)
11. MJ Lyons, J Budynek, A Plante, S Akamatsu, Classifying facial attributes using a 2-D Gabor wavelet representation and discriminant analysis, in *Proceedings of the 4th International Conference on Automatic Face and Gesture Recognition*, ed. by, 2000, pp. 202–207
12. I Cohen, N Sebe, Y Sun, MS Lew, TS Huang, Evaluation of expression recognition techniques, in *Proceedings of International Conference on Image and Video Retrieval*, ed. by, 2003, pp. 184–195
13. Y Wang, L Guan, Recognizing Human Emotional State from Audiovisual Signals. IEEE Transactions on Multimedia **10**(5), 659–668 (2008)
14. TF Cootes, C Taylor, D Cooper, J Graham, Active shape models: their training and their applications. Computer Vision and Image Understanding **61**(1), 38–59 (1995)
15. TF Cootes, GJ Edwards, CJ Taylor, Active appearance models. European Conference on Computer Vision **2**, 484–498 (1998)
16. D Cristinacce, T Cootes, Automatic feature localization with constrained local models. Pattern Recognition **41**, 3054–3067 (2008)
17. S Milborrow, F Nicolls, Locating facial features with an extended active shape model. European Conference on Computer Vision **5305**, 504–513 (2008)
18. T Yun, L Guan, Automatic face detection in video sequences using local normalization and optimal adaptive correlation techniques. Pattern Recognition **42**(9), 1859–1868 (2009)
19. P Viola, M Jones, P Viola, M Jones, Robust Real Time Object Detection. Proceedings of the 2nd International Workshop on Statistical and Computational Theories of Vision (2001)
20. DG Lowe, Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
21. I Rhodes, A tutorial introduction to estimation and Filtering. IEEE Autom. Control **16**(6), 688–706 (1971)
22. D Simon, *Optimal state estimation* (John Wiley & Sons, New Jersey, 2006)
23. S Dan, *"Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*, 1st edn. (Wiley-Interscience, Toulouse, France, 2006)
24. S Julier, J Uhlmann, Unscented filtering and nonlinear estimation. Proceedings of the IEEE **92**(3), 401–422 (2004)
25. HC Akakin, B Sankur, Robust Classification of Face and Head Gestures in Video. Image and Video Computing **29**, 470–483 (2011)
26. MK Pitt, N Shephard, Filtering via simulation: auxiliary particle filtering. J. American Statistical Association **94**, 590–599 (1999)
27. I Patras, M Pantic, Particle filtering with factorized likelihoods for tracking facial features, in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition* (, Seoul, Korea, 2004), pp. 97–102
28. Y Rui, Y Chen, Better Proposal Distributions: Object Tracking using Unscented Particle Filter. Proceedings of the Interntional Conference on Computer Vision and Pattern Recognition **2**, 786–793 (2001)
29. J Deutscher, A Blake, I Reid, Automatic partitioning of high dimensional search spaces associated with articulated body motion capture. Proceedings of the International Conference on Computer Vision and Pattern Recognition **2**, 669–676 (2001)
30. M Du, L Guan, Monocular Human Motion Tracking with the DE-MC Particle Filter. IEEE International Conference on Acoustics, Speech and Signal Processing **2**, 14–19 (2006)
31. M Maghami, RA Zoroofi, BN Araabi, M Shiva, E Vahedi, *Kalman Filter Tracking for Facial Expression Recognition using Noticeable Feature Selection* (International Conference on Intelligent and Advanced Systems, 2007)
32. C Hue, L Cadre, P P´erez, Tracking Multiple Objects with Particle Filtering. IEEE Transactions on Aerospace and Electronic Systems **38**, 791–812 (2002)
33. M Isard, J MacCormick, BraMBLe: A Bayesian multiple-blob tracker. Proceedings of the IEEE International Conference on Computer Vision **2**, 34–41 (2001)
34. J Vermaak, A Doucet, P P´erez, Maintaining Multi-Modality through Mixture Tracking. Ninth IEEE International Conference on Computer Vision **2**, 1110–1116 (2003)
35. T Yu, Y Wu, *Collaborative tracking of multiple targets* (IEEE CVPR, Washington, D.C., 2004)
36. Z Liyue, T Jianhua, *Fast Facial Feature Tracking with Multi-Cue Particle Filter* (Image and Vision Computing New Zealand, IVCNZ2007, Hamilton, New Zealand, 2000)

37. J MacComick, A Blake, Partitioned Sampling, Articulated Objects and Interface-Quality Hand Tracking. Proceedings of the European Conference on Computer Vision **2**, 3–19 (2000)
38. P Perez, C Hue, J Vermaak, M Gangnet, Color-Based Probabilistic Tracking. European Conference on Computer Vision (2002)
39. B Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, Las Vegas, 1986), pp. 254–259
40. Z Khan, T Balch, F Dellaert, Efficient particle filter-based tracking of multiple interacting targets using an MRF-based motion model. IEEE Intl. Conf. on Intelligent Robots and Systems (2003)
41. T Kanade, J Cohn, Y Tian, Comprehensive database for facial expression analysis. IEEE Int. Conf. Automatic Face and Gesture Recognition, 46–53 (2000)
42. SB Cohen, O Golan, S Wheelwright, J Hill, *Mind Reading: The Interactive Guide to Emotions* (Jessica Kingsley, London, 2004)
43. D Vukadinovic, M Pantic, *Fully Automatic Facial Feature Point Detection Using Gabor Feature Based Boosted Classifiers* (IEEE International Conference on Systems, Man and Cybernetics, Waikoloa, 2005)
44. B Wu, H Ai, C Huang, S Lao, Fast rotation invariant multi-view face detection based on RealAdaboost. Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (2004)
45. J Friedman, T Hastie, R Tibshirani, Additive logistic regression: A statistical view of boosting. The Annals of Statistics **28**, 337–374 (2000)
46. A Vezhnevets, V Vezhnevets, *Modest AdaBoost-teaching AdaBoost to generalize better* (Graphicon-2005, Novosibirsk Akademgorodok, Russia, 2005)
47. *GML AdaBoost Matlab Toolbox*. http://research.graphicon.ru/machine-learning/gml-adaboost-matlab-toolbox.html
48. I Matthews, S Baker, T Ishikawa, The Template Update Problem. IEEE T-PAMI **26**(6), 1115–1118 (2004)
49. Y Tie, L Guan, *Human Emotion Recognition Using a Deformable 3D Facial Expression Model* (IEEE International Symposium on Circuits and Systems, ISCAS, Seoul, Korea, 2012)
50. O Jesorsky, K Kirchberg, R Frischholz, Robust Face Detection Using the Hausdorff Distance, in *International Conference on Audio- and Video-Based Biometric Person Authentication*, ed. by (Springer, 2001), pp. 90–95
51. S Fazli, R Afrouzian, H Seyedarabi, Fiducial facial points tracking using particle filter and geometric features. Ultra Modern Telecommunications and Control Systems and Workshops, 396–400 (2010)
52. M Valstar, M Pantic, *Fully Automatic Facial Action Unit Detection and Temporal Analysis* (Computer Vision and Pattern Recognition Workshop, New Jersey, 2006)
53. C Rasmussen, G Hager, Probabilistic data association methods for tracking complex visual objects. IEEE Transactions on Pattern Analysis and Machine Intelligent, 560–576 (2001)
54. G Goodwin, M Seron, JD Dona, *Constrained control and estimation* (Springer, Verlag, 2005)
55. C Yang, E Blasch, Kalman filtering with nonlinear state Constraints. IEEE Trans. Aeros. Electron. Syst. **45**(1), 70–84 (2008)
56. B Bell, J Burke, G Pillonetto, An inequality constrained nonlinear Kalman–Bucy smoother by interior point likelihood maximization. Automatica **45**(1), 25–33 (2009)

doi:10.1186/1687-5281-2013-8
Cite this article as: Tie and Guan: **Automatic landmark point detection and tracking for human facial expressions.** *EURASIP Journal on Image and Video Processing* 2013 **2013**:8.