Intel Serv Robotics (2008) 1:143–157
DOI 10.1007/s11370-007-0014-z

SPECIAL ISSUE

Facial expression recognition and tracking for intelligent human-robot interaction

Y. Yang · S. S. Ge · T. H. Lee · C. Wang

Received: 27 June 2007 / Accepted: 6 December 2007 / Published online: 23 January 2008

Abstract For effective interaction between humans and socially adept, intelligent service robots, a key capability required by this class of sociable robots is the successful interpretation of visual data. In addition to crucial techniques like human face detection and recognition, an important next step for enabling intelligence and empathy within social robots is that of emotion recognition. In this paper, an automated and interactive computer vision system is investigated for human facial expression recognition and tracking based on the facial structure features and movement information. Twenty facial features are adopted since they are more informative and prominent for reducing the ambiguity during classification. An unsupervised learning algorithm, distributed locally linear embedding (DLLE), is introduced to recover the inherent properties of scattered data lying on a manifold embedded in high-dimensional input facial images. The selected person-dependent facial expression images in a video are classified using the DLLE. In addition, facial expression motion energy is introduced to describe the facial muscle's tension during the expressions for person-independent tracking for person-independent recognition. This method takes advantage of the optical flow which tracks the feature points' movement information. Finally, experimental results show that our approach is able to separate different expressions successfully.

Keywords Human–Robot interaction · Facial expression recognition · Affective computing · Distributed locally linear embedding · Facial expression motion energy

Y. Yang · S. S. Ge (✉) · T. H. Lee · C. Wang
Social Robotics Lab, Interactive Digital Media Institute
and Department of Electrical Computer Engineering,
National University of Singapore,
Singapore 117576, Singapore
e-mail: samge@nus.edu.sg

Y. Yang
e-mail: yangyong07@gmail.com

T. H. Lee
e-mail: eleleeth@nus.edu.sg

C. Wang
e-mail: g0600196@nus.edu.sg

1 Introduction

With the rapid advancement of both hardware and software technologies, robots are no longer confined to industry, and are entering and influencing the human social landscape in a big way. This has led to recent efforts by researchers worldwide in the area of social robotics [3,33,35]. Social robotics is the study of robots that are able to interact and communicate between themselves, with humans, and with the environment, within the social and cultural structure attached to its role [9,29,34]. As a class of social robots, intelligent service robots focuses on the provision of personalized services within the entertainment, games, healthcare industries, amongst many others. Unlike industrial robots, these sociable robots are specifically developed to interact with humans socially and evoking emotions through those interactions. It is crucial that social robots understand, perceive, respond appropriately, and even adapt their behavior based on the cues from human partners and augmented with their own understanding of the social environment they are situated within.

Intelligent service robots rely on effective utilization of available sensors – such as sound and vision sensors [11,12]—to gather information for decision making, planning, and ultimately empathetic interaction with humans. This paper deals specifically on the visual capabilities of intelligent social robots. Effective visual localization allows these robots to focus their attention on pertinent objects/

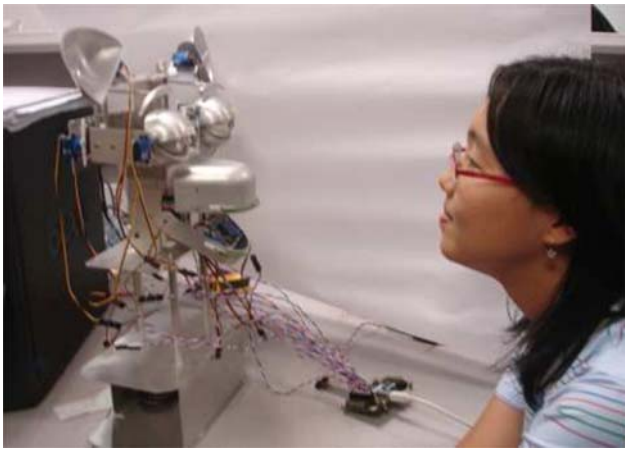


Fig. 1 Human robot interaction

elements within their environment, much like their human counterparts. As a crucial component of a social robot's sensing suite, a large part of research on social robots has focused on visual data analysis. It involves human/face detection and the fusion of stereo and infrared vision on board social robots with greater flexibility and robustness [10,14,20], for the purposes of attention focusing and for synthesizing more complex social interaction concepts, like comfort zones, into the robots.

For empathetic interaction between humans and social robots, the detection and recognition of faces are insufficient. It is vital to extract more information from visual data to facilitate meaningful interactions reminiscent of that between humans. The ability of emotional exchanges and interaction is one of the most important and necessary factor in the determining the level and type of interactions that occur between a robot and a human. The capacity to emote, and respond to emotes, further improves the robots' ability to engage in fruitful interactions with the human user, and is the first step in synthesizing the aspect of empathy within robots. A face to face human robot interaction is depicted in Fig. 1.

The most expressive way that humans display emotions is through facial expressions. Facial expression includes much information about human emotion. It can provide sensitive and meaningful cues about emotional response and plays a major role in human interaction and nonverbal communication.

Facial expression classification approaches could be divided into two main categories: target oriented and gesture oriented [4]. Target oriented approaches [27,28] attempt to infer the human emotion and classify the facial expression from one single image containing one typical facial expression. Gesture oriented methods [26,42] utilize temporal information from a sequence of facial expression motion images. In particular, transitional approaches attempt to compute the facial expressions from the facial neural condition and expressions at the apex.

Facial expressions are for the most part extremely dynamic. As such, the effective use of this dynamic information can prove invaluable and critical to the recognition and emotion interpretation process [17]. The temporal pattern of different expressions is considered in the recognition feature vector presented in [6]. In the system using a 3D face mesh based on the FACS model, the motion of the head and facial expressions is estimated in model-based facial image coding [31]. An algorithm for recovering rigid and nonrigid motion of the face was derived based on two or more frames. Independent component analysis, optical flow estimation and Gabor wavelet representation methods have also been used to achieve a 95.5% average recognition rate [5]. A method which relies on the overall pattern of a face which is represented by a potential field, and activated by edges in the image, has also been previously used for recognition [23]. Another technique consists of a motion energy template that uses a physics-based model to generate spatio-temporal motion energy templates for each expression [7]. The motion energy is converted from the muscular activations, and purely spatial information is used in the recognition pattern.

This paper investigates the efficient treatment of raw data, which is typically extremely high dimensional, for extracting vital information and facilitate emotion recognition. In addition, real-time response problems of emotion recognition techniques are examined in detail to remove reliance on manual intervention and tuning of any sort. In particular, the main contributions of this paper are:

(i) The introduction of the unsupervised learning method, distributed locally linear embedding (DLLE), to recover the inherent properties of scattered data lying on a manifold embedded in high-dimensional input data. High-dimensional facial expression images are embedded into a low-dimensional space which retains the intrinsic structures and main characteristics of a facial expression motion. Associated with support vector machines (SVM), a high recognition accuracy algorithm has been developed for static facial expression recognition.
(ii) A complete definition of facial expression potential energy and kinetic energy based on the facial features' movements is presented, and a facial expression energy system is constructed to describe the muscles' tension in facial expression for classification. By further considering different expressions' temporal transition characteristics, the actual occurrence of specific expressions can be identified with higher accuracy.

The remainder of this paper is organized as follows: In Sect. 2, an unsupervised learning algorithm is presented to

discover the intrinsic structure of the input data by preserving neighborhood relationship. In Sect. 3, the face detection and facial features extraction methods are discussed. The facial expression motion energy is proposed to describe the facial muscle's tension during the expressions for person independent tracking. In Sects. 4 and 5, we present the experimental results with our system and conclusion respectively.

2 Nonlinear dimension reduction (NDR) methods for person dependent recognition

Images lie in a very high dimensional space, but a class of images generated by latent variables lies on a manifold in this space. For human face images, the latent variables may be the illumination, identity, pose and facial deformations. In this paper, we are interested in embedding the facial deformations of a person in a very low dimensional space, which reflects the intrinsic structure of facial expressions. From training video sequences of different people undergoing different expressions, a low dimensional manifold is learned, with a subsequent probabilistic model used for tracking and recognition. On the manifold of expression, similar expressions are points in the local neighborhood, while different expressions separate apart.

Typical nonlinear dimensionality reduction techniques include Isomap by which the geodesic relationship among the input data, and the calculated low dimension embeddings remain constant [36], and locally linearly embeddings (LLE) by which the local intrinsic structures are maintained in dimensionality reduction [32]. A methodology called neighborhood linear embedding (NLE) [10] has been developed to discover the intrinsic property of the input data which is an adaptive scheme without the trial and error process in LLE. We modify the LLE algorithm and propose a new DLLE to discover the inherent properties of the input data [13].

2.1 Estimation of distribution density function

In most cases, a prior knowledge of the distribution of the samples in high dimension space is not available. However, we can estimate a density function of the given data. Consider a data set with N elements in m dimensional space, for each sample x_i, the approximated distribution density function $\hat{p}_{x_i}$ around point x_i can be calculated as:

$$\hat{p}_{x_i} = \frac{k_i}{\sum_1^N k_i} \tag{1}$$

where k_i is number of the points within a hypersphere kernel of fixed radius around point x_i.

Let $\hat{P} = \{\hat{p}_{x_1}, \hat{p}_{x_2}, \ldots, \hat{p}_{x_N}\}$ denote the set of estimated distribution density function, $\hat{p}_{\max} = \max(\hat{P})$ and $\hat{p}_{\min} = \min(\hat{P})$.

2.2 Compute the neighbors of each data point

Suppose that a data set $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, $x_i \in \mathbb{R}^m$ is globally mapped to a data set $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$, $y_i \in \mathbb{R}^l$, $m \gg l$. For the given data set, each data point and its neighbors lie on or close to a locally linear patch of the manifold. The neighborhood set of x_i, $\mathbf{S}_i$ $(i = 1, \ldots, N)$ can be constructed by making use of the neighborhood information.

Assumption 1 Suppose that the input data set $\mathcal{X}$ contains sufficient data in $\mathbb{R}^m$ sampled from a smooth parameter space Φ. Each data point x_i and its neighbors e.g. x_j, to lie on or close to a roughly linear patch on the manifold. The range of this linear patch is subject to the estimated sampling density $\hat{p}$ and mean distances $\bar{d}$ from other points in the input space.

Based on above geometry conditions, the local geometry in the neighborhood of each data point can be reconstructed from its neighbors by linear coefficients. At the same time, the mutual reconstruction information depends on the distance between the points. The larger the distance between points, the little the mutual reconstruction information between them.

Assumption 2 The parameter space Φ is a convex subset of $\mathbb{R}^m$. If x_i and x_j is a pair of points in $\mathbb{R}^m$, ϕ_i and ϕ_j is the corresponding points in Φ, then all the points defined by $\{(1-t)\phi_i + t\phi_j : t \in (0, 1)\}$ lies in Φ.

In view of the above observations, the following procedure is conducted making use of the neighbor information to construct the reconstruction data set of x_i, $\mathbf{S}_i$ $(i = 1, \ldots, N)$. To better sample the near neighbor and the outer data points, we propose an algorithm using an exponential format to gradually enlarge the range to find the reconstruction sample (Fig 2).

For a given point x_i, we can compute the distances from all other points around it. According to the distribution density function around x_i estimated before, we introduce α_i to describe the normalized density of the sample point x_i and is used to control the increment of the segment according to the sample points density for neighbor selection. We first give the definition of α_i by normalizing $\hat{p}_{x_i}$ using the estimated distribution density function computed by Eq. (1):

$$\alpha_i = \beta \cdot \frac{\hat{p}_{\max} - \hat{p}_{x_i}}{\hat{p}_{\max} - \hat{p}_{\min}} + \alpha_0 \tag{2}$$

where β is scaling constant, default value is set to 1.0, and α_0 is the constant to be set. The discussion of this definition is given later.

According to the distances values from all other points to x_i, these points are rearranged in ascending order and stored in $\mathbb{R}_i$. Based on the estimated distribution density function, $\mathbb{R}_i$ is separated into several segments, where $\mathbb{R}_i = R_{i1} \cup$

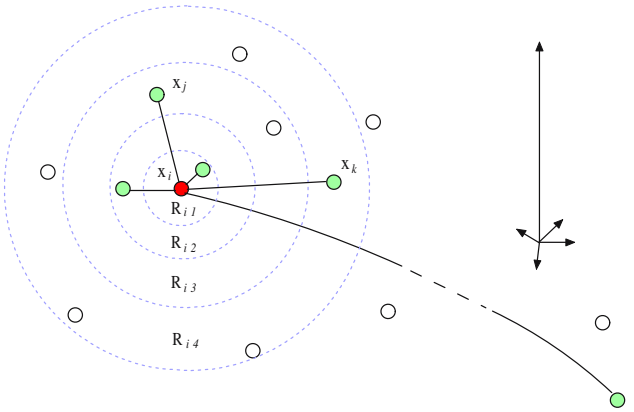


Fig. 2 The neighbor selection process

$R_{i2} \cup R_{i3} \cdots \cup R_{ik} \cdots \cup R_{iK}$. The range of each segment is given following an exponential format:

$$\begin{cases} \min(R_{ik}) = \lceil \alpha_i^k \rceil \\ \max(R_{ik}) = \lceil \alpha_i^{k+1} \rceil \end{cases} \tag{3}$$

where k is the index of segment and $\lceil \alpha_i^k \rceil$ denotes the least upper bound integer when α_i^k is not an integer. A suitable range of α_i is set from 1.0 to 2.0 by setting $\alpha_0 = 1.0$.

For each segment R_{ik}, the mean distance from all points in this segment to x_i is calculated by:

$$\overline{d}_{ik} = \frac{\sum_j \|x_i - x_j\|^2}{\max(R_{ik}) - \min(R_{ik})}, \quad \forall\, j \in R_{ik} \tag{4}$$

To overcome the information redundancy problem, using the mean distance computed by Eq. (4), we find the most suitable point in R_{ik} to represent the contribution of all points in R_{ik} by minimizing the following cost equation:

$$\varepsilon(d) = \min \|\overline{d}_{ik} - x_j\|^2, \quad \forall\, j \in R_{ik} \tag{5}$$

To determine the number of neighbors to be used for further reconstruction and achieve adaptive neighbor selection, we can compute the mean distance from all other samples to x_i

$$\overline{d}_i = \frac{1}{N} \sum_{j=1}^{N} \|x_i - x_j\|^2, \quad i \neq j \tag{6}$$

Starting with the $\mathbf{S}_i$ computed above at given point x_i, from the largest element in $\mathbf{S}_i$, remove the element one by one until all elements in $\mathbf{S}_i$ is less than the mean distance $\overline{d}_i$ computed by Eq. (6). Then the neighbor set $\mathbf{S}_i$ for point x_i is fixed.

2.3 Calculate the reconstruction weights

The reconstruction weight W is used to rebuild the given point. To store the neighborhood relationship and reciprocal contributions to each other, the sets $\mathbf{S}_i$ $(i = 1, 2, \ldots, N)$ are converted to a weight matrix $W = \{w_{ij}\}(i, j = 1, 2, \ldots, N)$. The construction weight W that best represents the given point x_i from its neighbor x_j is computed by minimizing the cost function given below:

$$\varepsilon(W) = \sum_i^N \left\| x_i - \sum_{j=\mathbf{S}_{i(1)}}^{\mathbf{S}_{i(n_i)}} w_{ij} x_j \right\|^2, \quad i \neq j \tag{7}$$

where the reconstruction weight w_{ij} represents the contribution of the jth data point to the ith point's reconstruction.

2.4 Computative embedding of coordinates

Finally, we find the embedding of the original data set in the low-dimensional space, e.g. l dimension. Because of the invariance property of reconstruction weights w_{ij}, the weights reconstructing the ith data point in m dimensional space should also reconstruct the ith data point in l dimensional space. Similarly, this is done by trying to preserve the geometric properties of the original space by selecting l dimensional coordinates y_i to minimize the embedding function given below:

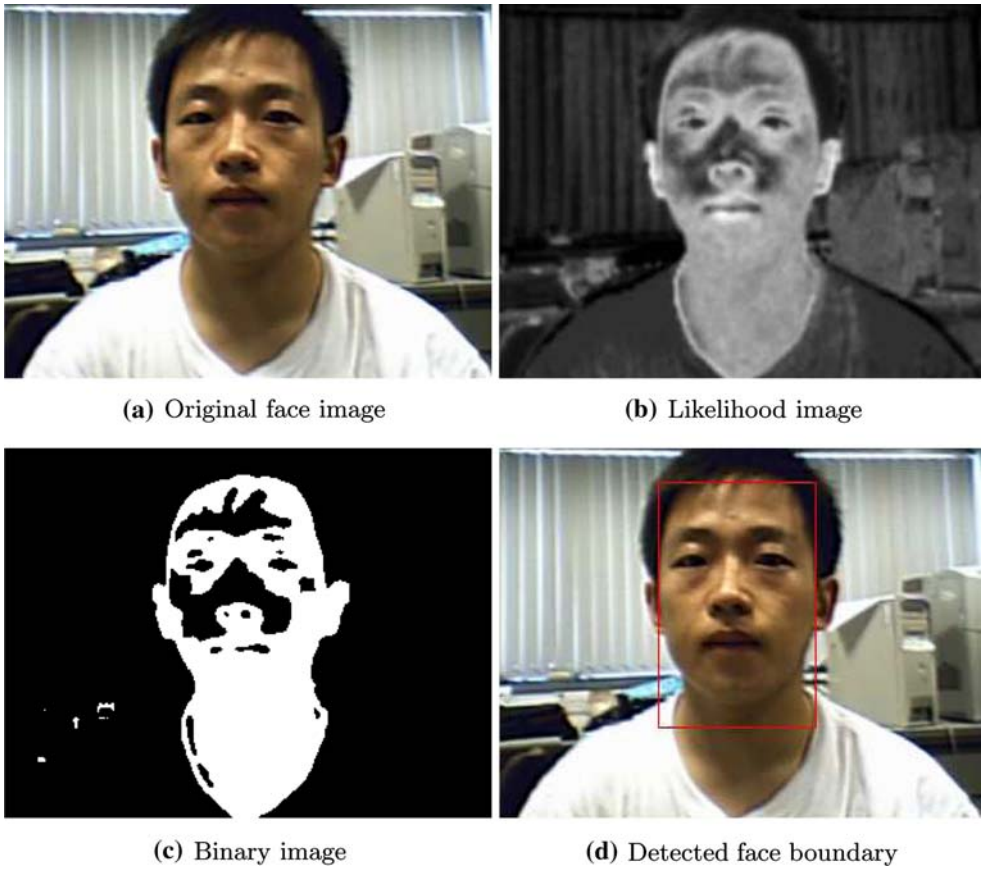
$$\Phi(Y) = \sum_i^N \left\| y_i - \sum_{j=\mathbf{S}_{i(1)}}^{\mathbf{S}_{i(n_i)}} w_{ij} y_j \right\|^2 \tag{8}$$

where w_{ij} are the reconstruction weights computed in Sect. 2.3, y_i and y_j are the coordinates of the point x_i and its neighbor x_j in the embedded space.

Based on the distances computed in low-dimensional space, support vector machines (SVM) is selected in our system as the classifier because of its rapid training speed and good accuracy [39]. SVM, which is proposed by Vapnik, is particularly a good tool to classify a set of points which belong to two or more classes. It is based on statistical learning theory and attempts to maximize the margin to separate different classes.

3 Facial expression energy for person independent recognition

Human face detection is the first task performed in the face recognition system which can ensure good results in the recognition phase. For example, it can fix a range of interests, decrease the searching range and initial approximation area for the feature selection [30]. However, face detection from a single image is a challenging task because of the high degree of spatial variability in scale, location and pose. In our system, we assume and only consider the



(a) Original face image **(b)** Likelihood image

(c) Binary image **(d)** Detected face boundary

Fig. 3 Face detection

situation that there is only one face contained in one image. The face should take up a significant area in the image. Once the face is detected, facial feature extraction is conducted which include: locate the position and shape of the eyebrows, eyes, nose, mouth, and extract features related to them in a still image of human face. Image analysis techniques are utilized which can automatically extract meaningful information from facial expression motion without manual operation to construct feature vectors for recognition.

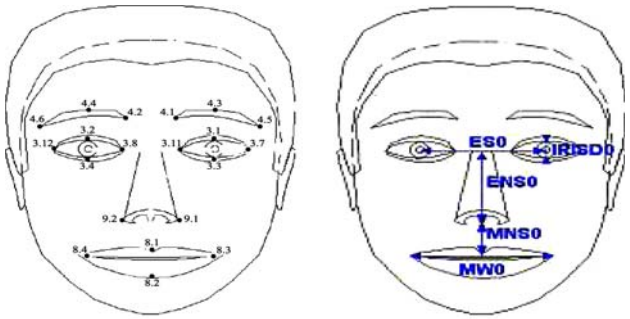
As we know, there is a maximal intensity of display a particular expression for each person [30]. There is also a maximal energy pattern for each person for their each facial expression. Therefore, facial expression energy can be used for classification by adjusting the general expression pattern to a particular individual according to the individual's successful expression recognition results. In this paper, we firstly give out a complete definition of facial expression potential energy and kinetic energy based on the facial features' movements information. A facial expression energy system is built up to describe the muscles' tension in facial expression for classification.

3.1 Face detection and feature extraction

3.1.1 Face detection

As indicated in many literatures, many different approaches make use of the skin color as an important cue for reducing the searching space [19,30]. We know that although the images are from different ethnicities, the skin distribution is relatively clustered in a small particular area [24]. On this 2D plane, the skin color area is comparatively more centralized which could be described by a Gauss distribution.

Through the distance between two pixels and the center we can obtain the information on how similar it is to skin and calculate a distribution histogram similar to the original image. The probabilities should be between 0 and 1, because we normalize the three components (R, G, B) of each pixels color at the beginning. The probability of each pixel is multiplied by 255 in order to create a gray-level image. This image is also called a likelihood image. The computed likelihood image is shown in Fig. 3b. After obtaining the likelihood of skin, a binary image can be obtained by thresholding each pixel as shown in Fig. 3c.



(a) FPs used in system. **(b)** FAPUs

Fig. 4 Feature points (FPs) and facial animation parameters units (FAPUs) (from ISO/IEC IS 14496-2 Visual, 1999 [1])

3.1.2 Facial feature extraction

The positions of eyebrow, eyes, nose and mouth are determined by searching for minima in the topographic grey level relief. The next step is to precisely find contour of the eyes and mouth. Because the real images are always affected by the lighting and noises, using some general local detection method such as corner detection is not robust and often require expert supervision [15]. We make fully use of the priority knowledge of human face, describe the eyes and mouth as piecewise polynomial, then use the deformable template to obtain a more precise contour. Because the eyes's color are not accordant and the edge information is abundant, we do edge detection first and followed by a closed operation. The inner part of the eye become high-luminance while the outer part of the eye become low-luminance. Figure 5 illustrates the feature extraction results on different testers.

To efficiently analyzing and correctly classify different facial expressions, it is crucial to properly determine which feature points are used. The MPEG-4 defines a standard face model using facial definition parameters (FDP) [1]. These proposed parameters can be used directly to deform the face model. The combinational manipulation of these parameters can result in a set of possible facial expressions. The proposed system uses a subset of facial animation parameters (FAPs) for describing the facial expressions which is supported by the MPEG-4 standard (Fig. 4a). The 21 visual features used in our system are carefully selected from the FAPs. These features are more prominent compared to other points defined by FAPs. At the same time, the movements of these feature points are significant, while an expression occur, which could be detected for further recognition.

In order to define FAPs for arbitrary face models, MPEG-4 defines FAP units (FAPUs) that serve to scale FAPs for any face model. FAPUs are defined as fractions of distances between marked key facial features (Fig. 4b). These features, such as eye separation and mouth width, are defined on a neutral face model. We choose the feature displacement and velocity approach due to its suitability for a real time video system, in which motion is inherent and which places a strict upper bound on the computational complexity of methods used in order to meet time constraints.

In order to measure facial related FAPs in real images and video sequences, quantitive modeling of FAPs is implemented using the features labeled as f_i. The features set employs FDP points that lie in the facial area and under some constraints, can be automatically detected and tracked. It consists of distance, $d(p_i, p_j)$, where p_i and p_j correspond to FDP points, between these protuberant points. Some of the points are constant during expressions and can be used as the reference points. Distances between reference points are used for normalization [38].

3.2 Facial expression energy

The facial expression energy is generated by computing the detailed facial feature physical movements data to a set of biologically motion energy. This method takes advantage of the optical flow which tracks the feature points' movements information [7]. For each expression, we use the facial feature movements information to compute the typical pattern of motion energy. These patterns are subsequently used for expression recognition.

3.2.1 Physical model of facial muscle

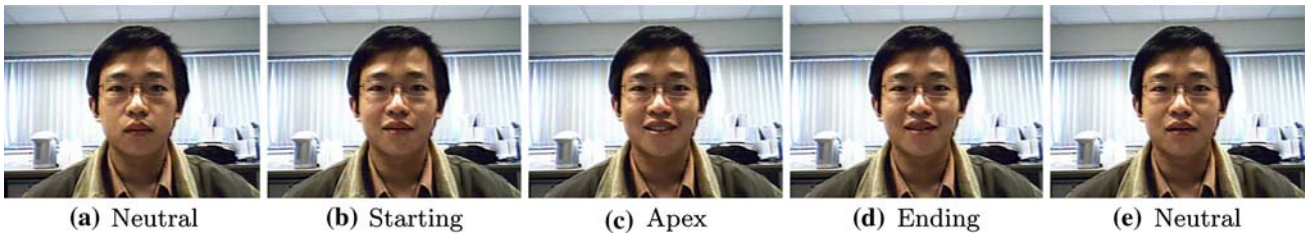
Muscles are a kind of soft tissues that possess contractile properties. Facial surface deformation during an expression is triggered by the contractions of the synthetic facial muscles. Muscle generates maximal concentric tension beyond its physiological range-at a length 1.2 times its resting length. Beyond this length, active tension decreases due to insufficient sarcomere overlap. To simulate muscle forces and the dynamics of muscle contraction, mass-spring model is typically utilized [18,25,37]. Waters and Frisbie [41] proposed a two-dimensional mass-spring model of the mouth with the muscles represented as bands. The facial mass-spring model used is similar as in [8]. Each node in the model is regarded as a particle with mass. The connection between two nodes is modeled by a spring. The node in the model can move to the position until it arrives at equilibrium point.

3.2.2 Emotion dynamics

One common limitation of the existing works is that the recognition is performed by using static cues from still face images without considering the temporal behavior of facial expressions. The psychological experiments by Bassili [2] have suggested that facial expressions are more accurately recognized from a dynamic image than from a single static image. The temporal information often reveals informa-

(a) Tester 1 (b) Tester 2

Fig. 5 Facial feature extraction results



(a) Neutral (b) Starting (c) Apex (d) Ending (e) Neutral

Fig. 6 Smile expression motion starting from the neutral state passing into the emotional state

tion about the underlying emotional states. For this purpose, our work concentrates on modeling the temporal behavior of facial expressions from their dynamic appearances in an image sequence.

The facial expression occur in three distinct phases which can be interpreted as: the beginning of the expression, the apex and the ending period [42]. Different facial expressions have their unique spacial temporal patterns at these three phases. Figure 6 shows a smile expression motion starting from the neutral state passing into the emotional state and end with a neutral state. Figure 7 shows the temporal curve of one mouth point of smile expression.

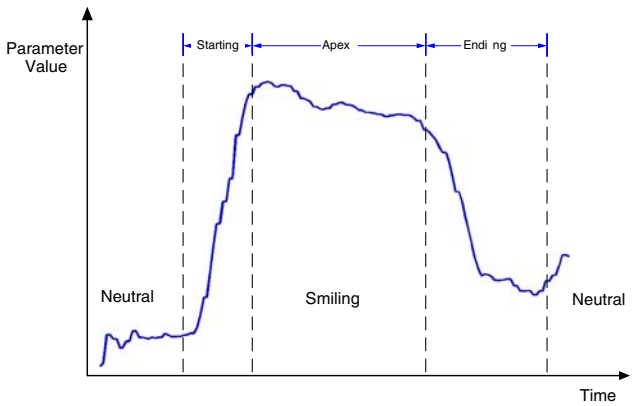


Fig. 7 The temporal curve of one mouth point in smile expression. Three distinct phases: starting, apex and ending

3.2.3 Potential energy

Expression potential energy is the energy that is stored as a result of deformation of a set of muscles [7]. It would be released if a facial expression in a facial potential field was allowed to go back from its current position to an equilibrium position (such as the neutral position of the feature points). The potential energy may be defined as the work that must be done in the facial expression, the muscles' force so as to achieve that configuration. Equivalently, it is the energy required to move the feature point from the equilibrium position to the given position. Considering the contractile properties of muscles, this definition is similar to the elastic potential energy. It is defined as the work done by the muscle's elastic force. For example, the mouth corner extended at the extreme position has greater facial potential energy than the same corner extended a bit. To move the mouth corner to the extreme position, work must be done, with energy supplied. Assuming perfect efficiency (no energy losses), the energy supplied to extend the mouth corner is exactly the same as the increase of its facial potential energy. The mouth corner's potential energy can be released by relaxing the facial muscle when the expression is to the end. As the facial expression fades out, its potential energy is converted to kinetic energy.

According to the feature movement information we obtained, we can define potential energy E_p at time t as:

$$E_p(p_i,t) = \frac{1}{2}\, k_i\ f_i(t)^2 = \frac{1}{2}\, k_i\ (D_{i\text{Neutral}} - D_i(t))^2 \tag{9}$$

- $f_i(t)$ is the distance between p_i and p_j at time t, expressed in m.
- $k_{i,j}$ is the the muscle's constant parameter (a measure of the stiffness of the muscle) linking p_i and p_j, expressed in N/m.

The nature of facial potential energy is that the equilibrium point can be set like the origin of a coordinate system. That is not to say that it is insignificant; once the zero of potential energy is set, then every value of potential energy is measured with respect to that zero. Another way of saying it is that it is the change in potential energy which has physical significance. Typically, the neutral position of a feature point is considered to be an equilibrium position. The potential energy is proportional to the distance from the neutral position. Since the force required to stretch a muscle changes with distance, the calculation of the work involves an integral. Equation (9) can be further written as follows with $E_p(p_i) = 0$ at the neutral position:

$$E_p(p_i,t) = -\int_{\mathbf{r}=0}^{\mathbf{r}} -k_i \mathbf{r}\ \mathrm{d}\mathbf{r} = -\left(\int_0^x -k_i x\ \mathrm{d}x + \int_0^y -k_i y\ \mathrm{d}y\right) \tag{10}$$

Potential energy is energy which depends on mutual positions of feature points. The energy is defined as a work against an elastic force of a muscle. When the face is at the neutral state, all the facial features are located at its neutral state, the potential energy is defined as zero. With the change of displacements of the feature points, the potential energy will change accordingly.

The potential energy can be viewed as description of the muscle's tension state. The facial potential energy is defined with an up-bound. It means that there is a maximum value when the feature point reach their extreme position. It is natural to understand because there is an extreme for the facial muscles's tension. When the muscle's tension reach the apex, the potential energy of the point associated with the muscle will reach its up-bound. For each person, the facial muscle's extreme tension is different. The potential motion energy varies accordingly.

Figure 8 shows the potential energy of two points: the left mouth corner and the lower mouth. The black contour

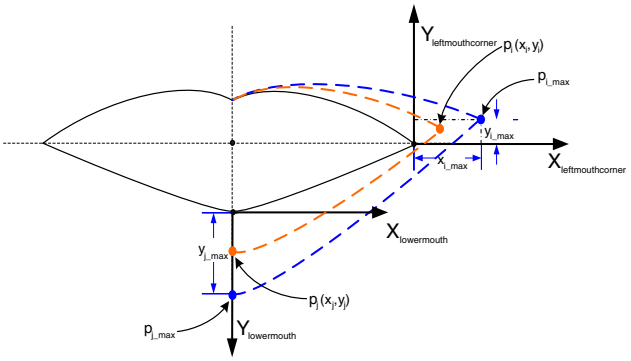


Fig. 8 The potential energy of mouth points

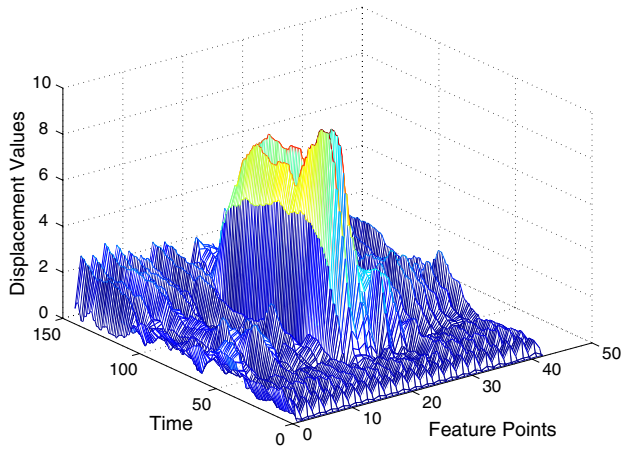


Fig. 9 The 3D spatio-temporal potential motion energy mesh of the smile expression

represents the mouth at its neutral position, the blue dash line represents mouth's extreme contour while the orange dash line is mouth contour at some expression. For the left mouth corner, we define a local coordinate that could be used for the computation of potential energy. The extreme point of the muscle tension is represented by $E_{pi\_\max}$. At this position, this feature point E_{pi} has the largest potential energy computed along the X-axis and Y-axis. When this feature point located between the neutral position and the extreme position, as illustrated of E_{pi}, its corresponding potential energy can be computed following Eq. (10). The same rule can also applied to the lower mouth point. According to the nature of human month structure, the movement of this feature point is mostly limited along the Y-axis.

Figure 9 shows the 3D spatio-temporal potential motion energy mesh of the smile expression. At the neutral state, all the facial features are located at their equilibrium positions. Therefore, the potential energy is equal to zero. When one facial expression reaches its apex state, its potential energy reaches the largest value. When the expression is at the ending state, the potential energy will decrease accordingly.

For each facial expression pattern, there are great varieties in the feature points' movements. Therefore, the potential energy value varies spatially and temporally. When an expression reaches its apex state, the potential value will also reach its maximum. Therefore, the pattern can be classified accordingly.

3.2.4 Kinetic energy

Kinetic energy is defined as a work of the force accelerating a facial feature points. It is the energy that a feature point possesses as a result of facial motion. It is a description energy.

Our system not only considers the displacement of the feature points in one direction, but also takes the velocity into account as movements pattern for analysis. The velocity of each feature points is computed frame by frame. It is natural that the feature points remain nearly static in the initial and apex state. During the change of the facial expressions, the related feature points' movements are fast. By analyzing the moving features' velocity, we can find the cue of a certain emotion.

According to the velocity obtained using Lucas and Kanade (L-K) optical flow method [21], we can define kinetic energy E_k as:

$$E_k(p_i, t) = \frac{1}{2} w_i \|v_i\|^2 \tag{11}$$

where w_i denote the ith feature point's weight, and v_i is the velocity for point i.

For each facial expression pattern, it will occur from the starting, translation and vanishing. At the neutral state, since the face is static, the kinetic energy is nearly zero. When the facial expression is at the starting state, the feature points are moving fast, the kinetic energy will vary temporally–increase first and decrease later. During this state, the muscle's biological energy is converted to feature points' kinetic energy. The kinetic energy is converted to feature points' potential energy. When an expression reaches its apex state, the kinetic energy will decrease to a stable state. If the facial muscle is still then, the kinetic energy will decrease to zero. At this time, the potential energy will reach to its apex. When the expression is at the ending state, feature points will move back to the neutral positions. Therefore, the kinetic energy will increase first and decrease later again. By analyzing and setting a set of rules, associated with the potential energy value, the pattern can be classified accordingly.

At the same time, the feature points' movement may temporally differ a lot when an expression occur, e.g. when someone is angry, he may frown first and then extend his mouth. Therefore, the kinetic energy for each feature points may not reach the apex concurrently.

We use a normalized dot product similarity metric to compare the differences between facial expressions. A simple form of similarity metric is the dot product between two vectors. We employ a normalized dot product as a similarity metric. Let X_i be the ith feature of the facial expression vector for expression X. Let the normalized feature vector, be defined as

$$\bar{X}_i = \frac{X_i}{\sqrt{\sum_j^m X_j^2}} \tag{12}$$

where m is the number of elements in each expression vector. The similarity between two facial expression vectors, X and Y, for the normalized dot product is defined to be $\bar{X} \cdot \bar{Y}$, the dot product on the normalized feature vectors.

4 Experiments and results

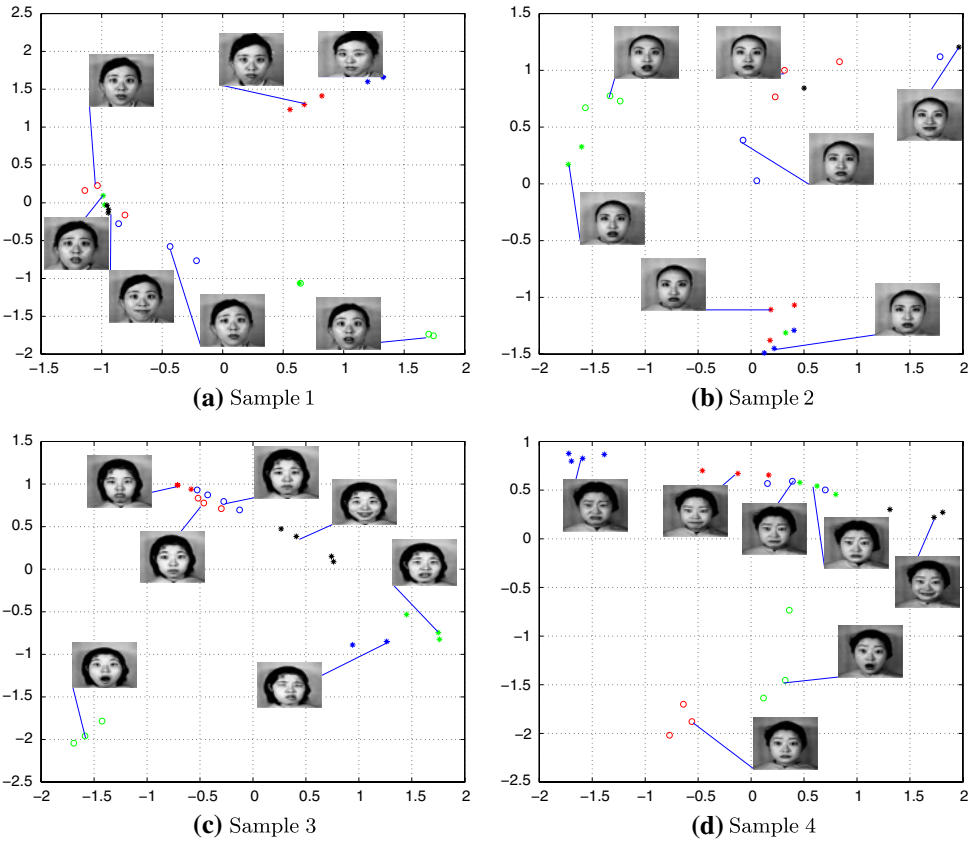
In this section, we present the results of simulation using the proposed static person dependent and dynamic person independent facial expression recognition methods. In our system, resolution of the acquired images is 320 × 240 pixels. Any captured images that are in other formats are converted first before further processing. Our system is developed under Microsoft Visual Studio.NET 2003 using VC++. The Intel's Open Source Computer Vision Library (OpenCV) is employed in our system [16]. The OpenCV Library is developed mainly aimed at real-time computer vision. It provides a wide variety of tools for image interpretation. The system is executed on a PC with Pentium IV 2.8G CPU and 512M RAM running Microsoft XP. Our experiments are carried out under the following assumptions:

- There is only one face contained in one image. The face takes up a significant area in the image.
- The image resolution should be sufficient large to facilitate feature extraction and tracking.
- The user's face should be kept stationary during the time when the initialization or re-initialization takes place.
- While tracking, the user should avoid much fast global movement. Sudden, jerky face movements should also be avoided.

The face tracking method does not require that the hand gesture must be centered in the image. It is able to detect frontal views of human faces under a range of lighting conditions. It can also handle limited changes in scale, yaw, roll and tilt.

4.1 Person dependent recognition

In this section, we make use of the similarity of facial expressions appearance in low-dimensional embedding to classify different emotions. This method is based on the observation



(a) Sample 1 **(b)** Sample 2 **(c)** Sample 3 **(d)** Sample 4

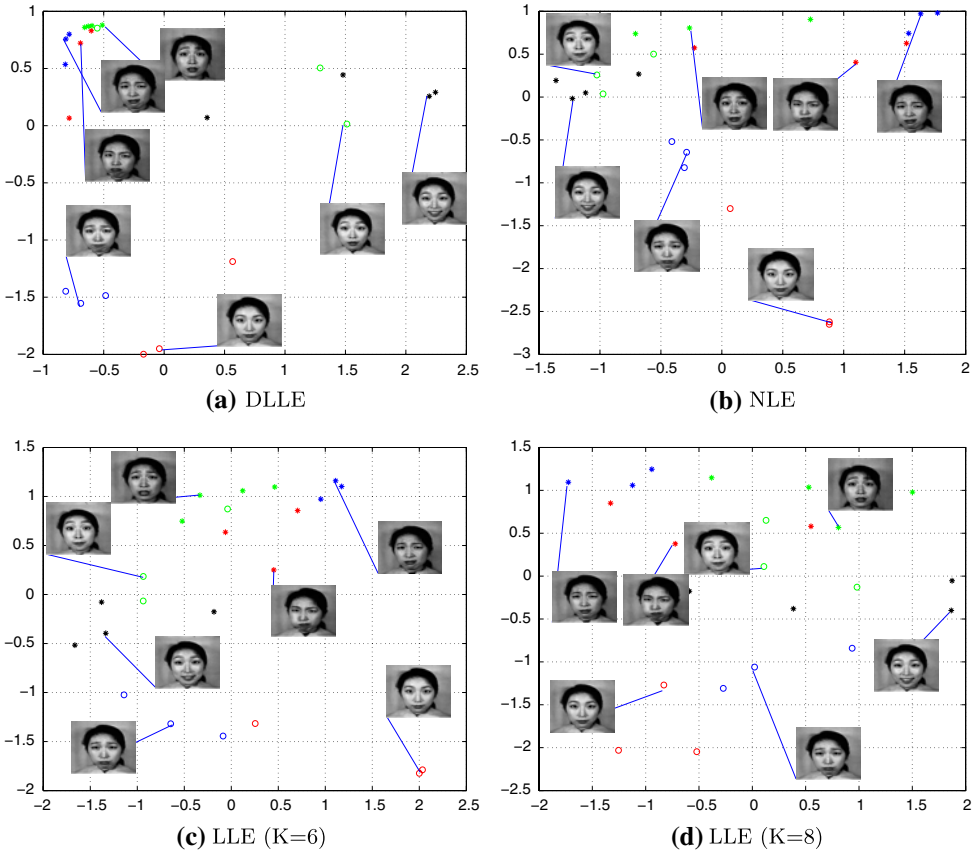
Fig. 10 The first two coordinates of DLLE of some samples of the JAFFE database [22]

(arguments) that facial expression images define a manifold in the high-dimensional image space, which can be further used for facial expression analysis. On the manifold of expression, similar expressions are points in the local neighborhood, while different expressions separate apart. The similarity of expressions depends greatly on the appearance of the input images. Since different people have great varieties in their appearances, the difference of facial appearance will overcome the discrimination caused by different expressions. It is a formidable task to group the same expression among different people by several static input images. However, for a certain person, the difference caused by different expressions can be used as the cues for classification.

As illustrated in Fig. 10, according to the DLLE algorithm, neighborhood relationship and global distribution can be preserved in the low dimension data set. The distances between the projected data points in low dimension space depend on the similarity of the input images. Therefore, images of the same expression are comparatively closer than images of different expressions in low dimension space. At this time, the training samples of the same expressions are "half clustered" and only a few of them may be apart from their corresponding cluster. This makes it easier for the classifier to categorize different emotions. Seven different expressions are represented by: anger, red star; disgust, blue star; fear, green star; happiness, black star; neutral, red circle; sadness, blue circle; surprise, green circle.

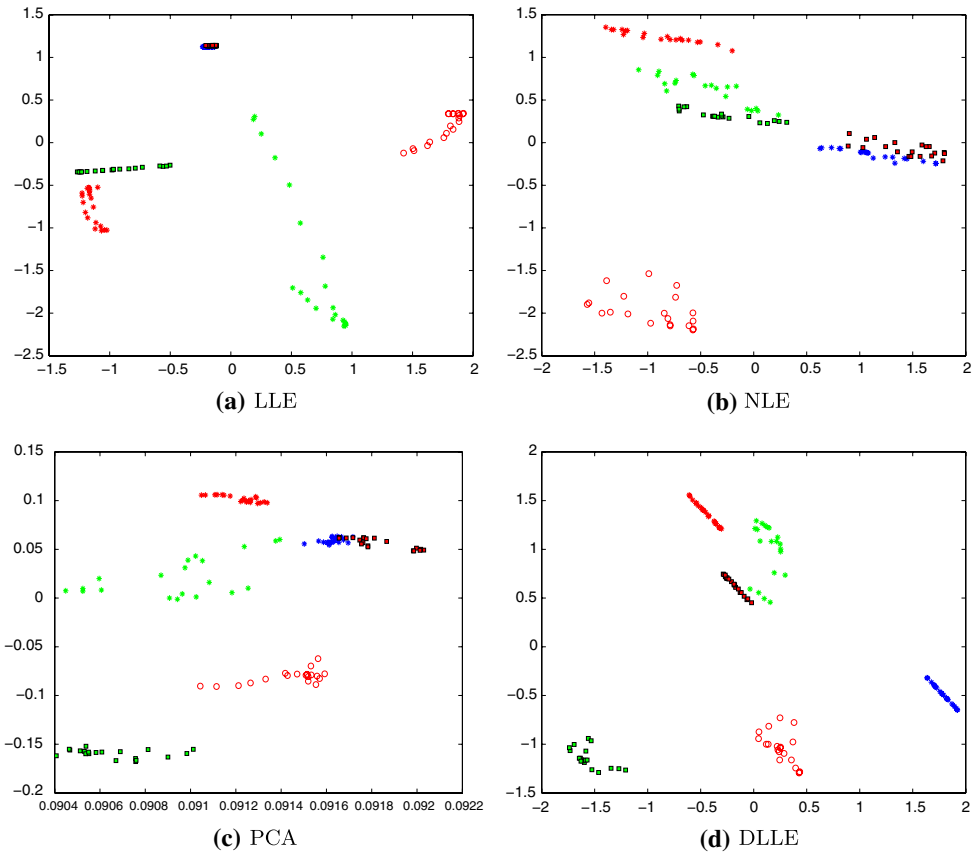
Static images taken at the expressions can also be employed. Figure 11 shows the result of projecting our training data (set of facial shapes) in a two dimensional space using DLLE, NLE and LLE embedding. The facial expressions are roughly clustered and the classifier works on a low-dimensional facial expression space. For the purpose of visualization, we can map the manifold onto its first two dimensional space. Figure 11 compares the two dimensional embeddings obtained by DLLE, NLE and LLE for 23 samples of one person from seven expressions respectively. We can see from Fig. 11a that for $d = 2$, the embedding of DLLE separates the seven expressions well. Samples of the same gesture clustered together while only a few different gesture samples are overlapped. Figure 11b shows that the embedding of NLE can achieve similar result as DLLE. The LLE is very sensitive to the selection of number of nearest neighbors. The images of different expressions become mixed up easily when we increase the number of nearest neighbors as shown in Fig. 11c, d.

In Fig. 12, we compare the properties of the LLE, NLE, PCA and DLLE after the sample images are mapped to 2D dimension using the feedtum database [40]. Six different expressions are represented by: anger, blue star; disgust, red



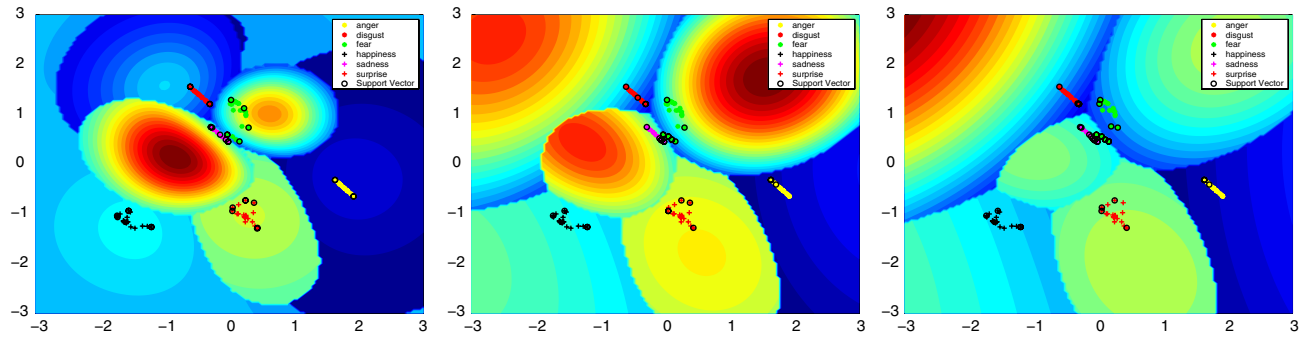
(a) DLLE **(b)** NLE **(c)** LLE (K=6) **(d)** LLE (K=8)

Fig. 11 2D projection using different Nonlinear Reduction methods using samples from JAFFE database [22]



(a) LLE **(b)** NLE **(c)** PCA **(d)** DLLE

Fig. 12 2D projection using different nonlinear reduction methods



(a) SVM (first order) **(b)** SVM (third order) **(c)** SVM (fifth order)

Fig. 13 The SVM classification results according to the 2D embedding

star; fear, green star; happiness, green square; sadness, black square; surprise, red circle. The projected low dimension data should keep the separating features of the original images. Images of the same expression should cluster together while different should be apart. There are 120 samples of one person from six expressions respectively (20 samples per expression). We can see from Fig. 12d that for $d = 2$, different expressions' embedding of LLE are separated. However, the red and blue points are overlapped and not separatable in 2D dimension. Figure 12b shows the embedding of NLE. It can be seen that in general they are separated, but the boundary between different groups are not clear. PCA achieves similar result as NLE which is shown in Fig. 12c. The samples of the same expression are not so centralized and the red and blue star samples are mixed up. As illustrated in Fig. 12d, we can see that DLLE can separate the six expressions well. Samples of the same expression cluster together while different expression samples are clearly separated.

The reason is that LLE is an unsupervised learning algorithm. It selects the nearest neighbors to reconstruct the manifold in the low dimensional space. There are two types of variations in the data set: the different kinds of facial expressions and the varying intensity for every kind of facial expression. Generally, LLE can catch the second type of variation-an image sequence is mapped in a "line", and LLE can keep the sequences with different expressions distinctive when there is only one sequence for each expression. When the data set contains many image sequences for the same kind of expression, it is very hard to catch the first kind of variation using a small number of nearest neighbors. But with the increased number of nearest neighbors, the images of different expressions are more prone to be mixed up.

Figure 13 demonstrates the SVM classification results on the 2D embedding of the original data Fig. 12d. The kernel was chosen to be the polynomial. The polynomial mapping is a popular method for non-linear modeling. The penalty parameter is set 1,000 ($C = 1,000$). Figures 13a–c illustrate the SVC solution obtained using a degree 1, degree 3 and degree 5 polynomial for the classification. The circled points are the support vectors for each classes. It is clear that SVM can correctly classify the embedding of sample data sets.

Tables 1 and 2 show the recognition results using DLLE and SVM(one against one algorithm) for the training and testing data. The database contains 480 images of 6 different type of expressions for training. These samples are used for training the SVM. Apart from the training samples, there are another 120 samples of six expressions are employed to be tested.

4.2 Person independent recognition

Although person dependent method can reach satisfactory results, it is required a set of pre-captured expression samples. If the robot has stored the someone' expression images and its computation speed is fast enough, it could recognize his/her expressions at run time using the person dependent recognition method. Most of the existing methods are not conducted in real-time [30,43]. A general method is needed which can recognize facial expressions of different individuals without the training sample images. By analysis of facial movements pattern captured by optical flow tracker, a recognition system based on facial expression motion energy is setup to recognize expressions in real time.

Initially, a front view image of the tester's neutral face is captured. This image is processed to detect the tester's face region, extract the eyebrows, eyes, nose and mouth features according to the methods described in Sect. 3.1. In fact, this process is done in a flash. Our system is able to complete the process by just clicking a button on the interface. The features locations are then mapped to the real-time video according to the video's resolution. Once the initialization is completed, the tester can express his emotion freely. The feature points can be predicted and tracked frame by frame using Lucas-Kanade optical flow method. The displacement and velocity of each feature points are recorded at each frame. By

Table 1 Recognition results using DLLE and SVM(1V1) for training data

Emotion	Happiness	Sadness	Fear	Disgust	Surprise	Anger	Rate (%)
Happiness	80	0	0	0	0	0	100
Sadness	0	80	0	0	0	0	100
Fear	0	0	80	0	0	0	100
Disgust	0	0	0	80	0	0	100
Surprise	0	0	6	0	73	1	91.25
Anger	0	0	0	0	1	79	98.75

Table 2 Recognition results using DLLE and SVM(1V1) for testing data

Emotion	Happiness	Sadness	Fear	Disgust	Surprise	Anger	Rate (%)
Happiness	18	2	0	0	0	0	90
Sadness	0	20	0	0	0	0	100
Fear	0	0	19	0	1	0	95
Disgust	0	0	0	20	0	0	100
Surprise	0	0	0	0	20	0	100
Anger	0	0	0	0	1	19	95

Fig. 14 Real-time video tracking results in different environment

analyzing the dynamic movement pattern of feature points, the expression potential energy and kinetic energy are computed out in real-time. Once an expression occur, the detection system will make a judgement using the method described in Sect. 3.2. The recognition result will be displayed at up-right corner of the video window. When one expression is over, the tester can express his following emotions or re-initialize the system if any tracker is lost.

Figure 14 shows the expression recognition results under different environments. It can be seen from these figures that the system can robustly recognize the human's expression regardless the background.

The results of real-time person independent expression recognition are given in Fig. 15. Our system can reach 30 FPS (frame per second). The pictures are captured while the expression occurs. The recognition results are displayed in

Fig. 15 Real-time video tracking results for other testers

(a) Happiness **(b)** Surprise

(c) Happiness **(d)** Happiness

real-time in red at the up-left corner of the window. From these pictures, we can see that our proposed system can effectively detect the facial expressions.

5 Conclusions and future directions

This paper investigated the emotion detection and recognition aspect of visual sensing that forms a crucial part of allowing empathetic interaction between intelligent service robots and humans. Both person-dependent and person-independent recognition approaches have been examined, and the proposed methods can successfully recognize the static, off-line captured facial expression images, track and identify dynamic on-line facial expressions of real-time video from camera. An unsupervised learning algorithm, DLLE, has been introduced to discover the intrinsic structure of the high dimensional data, and the discovered properties were used to compute their corresponding low-dimensional embedding. Associated with SVM, a high recognition accuracy algorithm has been developed for static facial expression recognition. Facial expression motion energy has also been introduced to describe the facial muscle's tension during the expressions for person-independent tracking. Extensive simulations verify the effectiveness of the proposed approach.

One limitation of the current system is that it can detects only one front view face looking at the robot. Multiple face detection and feature extraction could be further improved. Since the current system can deal with some degree of lighting and orientation variation, the resolution of the image would be the main problem to concur for multi-person expression analysis. One direction to advance our current work is to combine the human speech and make intelligent human robot interface, and explore robotic human companion for learning and information seeking.

Acknowledgments The authors thank the following persons: Yaozhang Pan, Yang Chen, Chengchen Li, Lan Zhang and Zhen Su for their time and effort in verifying the methods proposed in the paper.

References

1. ISO/IEC IS 14496-2 (1999) Visual A compression codec for visual data
2. Bassili J (1979) Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. J Personality Soc Psychol 37:2049–2059
3. Breazeal C (2003) Toward sociable robots. Robot Auton Systems 42:167–175
4. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor J (2001) Emotion recognition in human–computer interaction. IEEE Signal Process Magaz 18(1):32–80
5. Donato G, Bartlett MS, Hager JC, Ekman P, Sejnowski TJ (1999) Classifying facial Actions. IEEE Trans Pattern Anal Mach Intell 21(10):974–989

6. Essa I, Pentland A (1997) Coding, analysis, interpretation, and recognition of facial expressions. IEEE Trans Pattern Anal Mach Intell 19(7)
7. Essa IA, Pentland A (1995) Facial expression recognition using a dynamic model and motion energy. In: International conference on computer vision (ICCV), pp 360–367
8. Feng GC, Yuen PC, Lai J (2000) Virtual view face image synthesis using 3d spring-based face model from a single image. In: Proceedings of fourth IEEE international conference on automatic face and gesture recognition, pp 530–535
9. Ge SS (2007) Social Robotics: Integrating advances in engineering and computer science. In: Proceedings of electrical engineering/electronics, computer, telecommunications and information technology international conference. Chiang Rai, Thailand, pp xvii–xxvi
10. Ge SS, Guan F, Loh A, Fua C (2006) Feature representation based on intrinsic structure discovery in high dimensional space. IEEE International Conference on Robotics and Automation, pp 3399–3404
11. Ge SS, Loh AP, Guan F (2003) Sound localization based on mask diffraction. In: Proceedings of IEEE international conference on robotics and automation, pp 1972–1977
12. Ge SS, Loh AP, Guan F (2005) Robust sound localization using lower number of microphones. Int J Info Acqui 2(1):1–22
13. Ge SS, Yang Y, Lee TH (2006) Hand gesture recognition and tracking based on distributed locally linear embedding. In: Proceedings of 2nd IEEE international conference on robotics, automation and mechatronics. Bangkok, Thailand, pp 567–572
14. Guan F, Li LY, Ge SS, Loh AP (2007) Robust huamn detection and identification by using stereo and thermal images in human robot interaction. Int J Inf Acquis 4(2):161–183
15. Harris C, Stephens M (1988) A combined edge and corner detector. In: Proceedings of the 4th Alvey vision conference, pp 147–151
16. Intel Corporation: OpenCV Reference Manual (2001) http://www.intel.com/technology/computing/opencv/index.htm
17. Izard CE (1990) Facial expressions and the regulation of emotions. J Personality Soc Psychol 58(3):487–498
18. Kahler K, Haber J, Seidel H (2001) Geometry-based muscle modeling for facial animation. In: Proceedings of graphics interface
19. Kotropoulos C, Pitas I (1997) Rule-based face detection in frontal views. In: Proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP 97), vol IV, pp 2537–2540
20. Loh AP, Guan F, Ge SS (2004) Motion estimation using audio and video fusion. In: Proceedings of the 8th international conference on control, automation, robotics and vision (ICARCV), pp 1569–1574
21. Lucas B, Kanade T (1981) An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th international joint conference on artificial intelligence (IJCAI '81), pp 674–679
22. Lyons MJ, Akamatsu S, Kamachi M, Gyoba J (1998) Coding facial expressions with gabor wavelets. In: Proceedings of the third IEEE international conference on automatic face and gesture recognition, pp 200–205. http://kasrl.org/jaffe.html
23. Matsuno K, Tsuji S (1994) Recognizing human facial expressions in a potential field. In: International conference on pattern recognition, pp. 44–49
24. McKenna S, Gong S, Raja Y (1998) Modelling facial colour and identity with gaussian mixtures. Pattern Recogn 31(12):1883–1892
25. Nedel LP, Thalmann D (1998) Real time muscle deformations using mass-spring systems. In: Proceedings of the computer graphics international, pp 156–165
26. Otsuka T, Ohya J (1996) Recognition of facial expressions using HMM with continuous output probabilities. In: Proceedings 5th IEEE international workshop on robot and human communication (RO-MAN), pp 323–328
27. Padgett C, Cottrell G (1997) Representing face images for classifying emotions, vol 9. MIT Press, Cambridge
28. Padgett C, Cottrell G, Adolps B (1996) Categorical perception in facial emotion classification. In: Proceedings of cognitive science conference, vol 18, pp 249–253
29. Paiva A (ed) (2000) Affective interactions: towards a new generation of computer interfaces. Springer, New York
30. Pantic M, Rothkrantz LJM (2000) Automatic analysis of facial expressions: the state of the art. IEEE Trans Pattern Anal Mach Intell 22(12):1424–1445
31. Roivainen P, Li H, Forcheimer R (1993) 3-D motion estimation in model-based facial image coding. IEEE Trans Pattern Anal Mach Intell 15:545–555
32. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326
33. Tapus A, Mataric MJ (2006) Towards socially assistive robotics. Int J Robot Soc Jpn (JRSJ) 24(5):576–578
34. Tapus A, Mataric MJ (2007) Methodology and themes of human-robot interaction: A growing research field. Int J Adv Robot Systems 4(1):103–108
35. Tapus A, Mataric MJ, Scassellati B (2007) The grand challenges in socially assistive robotics. IEEE Robot Autom Magaz (RAM), Spec Issue Grand Chall Robot 14(1):35–42
36. Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323
37. Terzopoulos D, Waters K (1993) Analysis and synthesis of facial image sequences using physical and anatomical models. IEE Trans Pattern Anal Mach Intell 15(6):569–579
38. Tsapatsoulis N, Raouzaiou A, Kollias S, Cowie R, Douglas-Cowie E (2002) MPEG-4 facial animation, chap. Emotion recognition and synthesis based on MPEG-4 FAPs. Wiley, New York
39. Vapnik V (1995) The nature of statistical learning theory. Springer, New York
40. Wallhoff F: Facial expressions and emotion database. http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html
41. Waters K, Frisbie J (1995) A coordinated muscle model for speech animation. In: Proceedings of graphics interface, pp 163–170
42. Yacoob Y, Davis LS (1996) Recognizing human facial expressions from long image sequences using optical flow. IEEE Trans Pattern Anal Mach Intell 18:636–642
43. Zhang Y, Ji Q (2005) Active and dynamic information fusion for facial expression understanding from image sequences. IEEE Trans Pattern Anal Mach Intell 27:699–714