A vision-based hybrid method for facial expression recognition

X. Huang
Department of Mechanical and Industrial Engineering
Northeastern University
Boston, MA 02115
(001) 617-373-4892
xianyih@coe.neu.edu

Y. Lin*
Department of Mechanical and Industrial Engineering
Northeastern University
Boston, MA 02115
(001) 617-373-8610
yilin@coe.neu.edu
*: The corresponding author

ABSTRACT
Facial expression is a very useful channel for intelligent human computer communication. In this paper we propose a hybrid method to recognize facial expression. Our main contributions in this study are: first, face region is detected by combing Adaboost, Skin color model and motion history image; second, feature points representing different facial expressions are separated using optical flow; third, a support vector machine classifier is used to classify these feature point's info (location, distance, angle); last, tests to explore the whole facial expression recognition process are conducted and the results are satisfactory.

Categories and Subject Descriptors
D.3.3 [**VISIONS**]: pattern recognition and computer vision

General Terms
Human factors

Keywords
Facial expression recognition, Adaboost, Skin color model, Motion history image, Optical flow, Support vector machine.

1. INTRODUCTION
There is a high need for the computer to be able to understand human emotional behavior. Ref. [Picard, 1997] suggested that knowing the user's emotions, the computer can become a more effective "tutor", "helper" or "companion". In our real life, there are a variety of ways a human express his emotions such as verbal language, and non-verbal means, e.g. body posture, facial expressions, and others. As cited in Ref. [Mehrabian, 1972], 93% of human communication is conveyed by nonverbal communicative behavior (e.g. body language, facial expression), only 7% of the impact accounts from spoken words. Considering the non-verbal parts, facial expression contributes 53% of the total effect. There is a good reason to believe that acts of non-verbal communication (facial expression) play a central role in human-computer interaction. In recent years, the use of sensors, particularly contact-less sensors, to track human operators and understand their behaviors has become an active topic. Web camera, with its obvious virtue of non-intrusiveness and low-cost high quality, has made visual analysis become one of the most promising technologies for facial expression recognition.

Ambi-sys 2008, February 11-14, 2008, Quebec, Canada.

Currently, most of existing studies are on the consideration of visual expression analysis. There are two objectives of the present paper. First, we review existing studies on facial expression recognition. Second, we present our approach which combines Adaboost, skin-color, and motion history image. We call our approach "hybrid approach" for short. At the end, we also present our experimental results to demonstrate the effectiveness of our hybrid approach and to suggest some future work.

2. PREVIOUS RESEARCH AND RELATED WORK
The pioneering work by Ekman [Ekman, 1975] formed the basis of visual face expression recognition. His studies indicated that there are six universally-recognized prototypes of face expression: happiness, anger, disgust, sadness, fear, and surprise. His work on action units (AU), described in Facial Action Coding System (FACS), and inspired researchers in this field. For example, an Automatic Face Analysis (AFA) system was developed to recognize fine-grained changes in facial expression into action units (AU) of the Facial Action Coding System; the system achieved an average recognition rate of 96.4 percent (95.4 percent if neutral expressions are excluded) for upper face AU and 96.7 percent (95.6 percent with neutral expressions excluded) for lower face AU [Tian, 2001]. Based on FACS, a description of body action units (BAU) was even presented [Gunes and Piccardi, 2005]. To enhance the robustness in identification, a hybrid approach was introduced with the use of facial feature point tracking, dense flow tracking with principal component analysis (PCA), and high gradient component detection combined by Hidden Markov Model (HMM) [Lien, 1998].

Facial expression recognition consists of two major parts: facial feature extraction and classification of the extracted facial feature. Different methods for facial expression recognition differ in the feature extraction and representation method, type of classification, and whether the recognition is done from still image or video. For the facial feature extraction and

representation method, there are three main types: template-based, feature-based, and appearance-based.

Template-based methods use the holistic info of facial recognition patterns and their advantage is the simplicity; however, they need to allocate large amounts of memory and cannot handle face with different poses well. The k-nearest-neighbor (*k*-NN) algorithm is a commonly used method in template-based facial expression recognition. Unlike template-based methods, feature-based methods can better solve such a pose problem, as they are based on the extraction of facial geometric features (i.e., position and shape of eyes, nose, and mouth, or eyebrow arches). They are often used in the face normalization, such as scaling or rotating correction. Generally speaking, the feature-based methods have a smaller memory requirement and a higher recognition speed than the template-based ones. However, perfect extraction of features is still an arduous task. Many techniques have been developed to detect facial features in a single image or sequence of images, including Optical flow, Gabor wavelets. Optical flow is used to detect the displacement and velocity of some key facial features determined by finding all the pixels in one frame whose values match the values of pixels belonging to that object in the previous frame. For example, the facial features from FERET were extracted by a geometric face model and edge finding and high threshold were used to find face boundary [Shih and Chuang, 2004]; an idea of using Gabor wavelets to perform feature extraction was described in Ref. [Lades, 1993]: convolving a gray-level input image with a series of Gabor wavelets, thereby performing a filter operation. The idea of appearance-based method is to project face images or subimages onto a low-dimensional feature space. Eigenface is such a representative method. It was first proposed in 1991 by Turk and Pentland [Turk and Pentland, 1991] and used to represent a face by projecting original images onto a face space defined by eigenvectors using principal component analysis (PCA). Any human face can be represented by a linear combination of eigenfaces. In other words, one person's face can be merged by one portion of an eigenface of one kind and other portion of an eigenface of another kind, and so on. Its procedure includes the calculation of average face templates, derivation of eigenvectors, and reconstruction of original images using calculated eigenvectors, and comparison of a new face to known faces by calculating the distance between their projections onto face space (**Figure 1**).

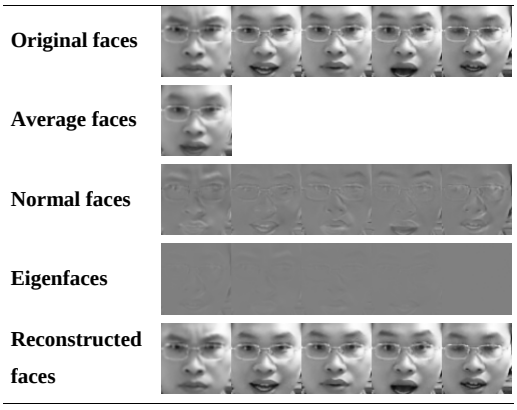


Figure 1 Eigenface sample procedure for face recognition

With regard to the method of classification, there are also two major types: image-based or sequence-based. Neural networks (NN), support vector machine (SVM), and Bayesian networks (BN) belong to the image-based classification, while Hidden Markov Model (HMM) and Dynamic Bayesian Networks (DBN) to the latter. A comprehensive review of these methods can be found in [Li, 2005] [Kaliouby, 2003].

So far, there is no technique that is sound enough to work perfectly in all cases. One of the most challenging issues is how to deal with variable illumination on human face. Most face images obtained from daily lives are more or less affected by the light. Therefore, a well-developed face recognition system has to handle this issue properly. In terms of this consideration, we choose a number of technologies and make a comparison in the following subsections.

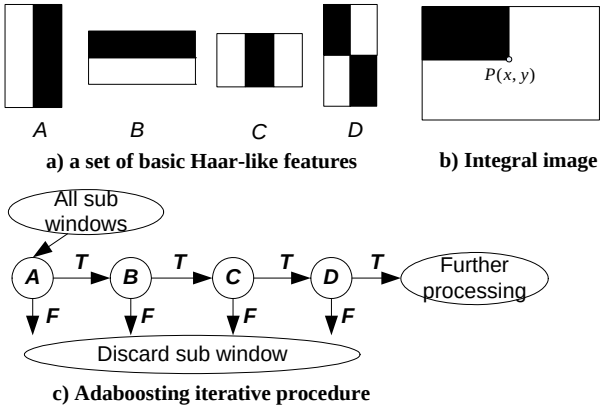
3. METHODOLOGY

3.1 Face detection

Before facial expression recognition is performed, the first thing we should do is to determine whether a particular face in a given image. In a sense, the success of face detection determines the performance of facial expression recognition. Once the face is detected, the face region should be extracted from the scene for later-on face recognition. Usually, the face detection and face region segmentation are performed simultaneously. In the following subsections, we analyze three face detecting algorithms, which will be applied in our facial expression recognition system.

3.1.1 Viola and Jones algorithm

In face detection, the recent representative work is done by Viola and Jones. In the paper [Viola and Jones, 2001], Viola and Jones presented an algorithm for face detection embodied with three techniques: Haar-like features (**Figure 2a)**), integral image (**Figure 2b)**), and Adaboost learning rule (**Figure 2c)**). A Harr-like feature is a sub-window composed of two or three or four connected "black" and "white" rectangles. The value of a Haar-like feature is the difference between the sums of the pixel gray values within the black-and-white rectangular region [Wilson, 2006]. The Haar-like features in a rectangle can be calculated rapidly using integral image. Integral image at location (x, y) includes the sum of the pixel values in the above and left of the x and y [Wilson, 2006]. Adaboost learning rule is an iterative algorithm, in order to improve the detection accuracy step by step based on a series of "weak" classifiers, and the final output is all the sub-windows are classified as face or non-face region.



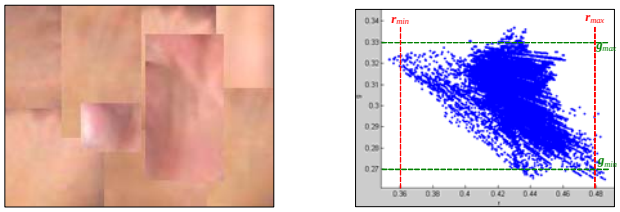
a) a set of basic Haar-like features **b) Integral image**

c) Adaboosting iterative procedure

Figure 2 Viola and Jones algorithm

3.1.2 *Skin color model*

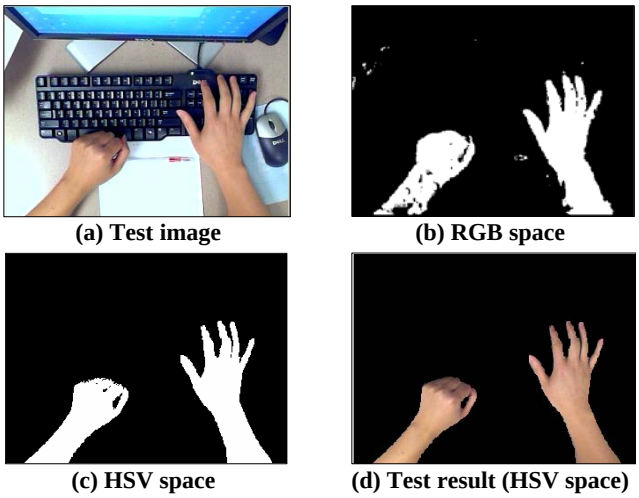
The distribution of skin color samples is found concentrated in a small area (**Figure 3**). Based on this finding, the skin regions can be separated from the rest of the image through a threshold process. Since color images are mostly stored in RGB mode, the RGB space is commonly chosen as the color space for skin-color model. In the common RGB color space, the triple components (r, g, b) of images represent not only color but also brightness. Brightness may vary due to the lighting changes in the surrounding environment in which the RGB model is not a reliable measure of skins separated from non-skin regions [Yang, 1997].



a) Sampled skins **b) Skin region**

Figure 3 Skin color clustering in 2D space

Another alternative people prefer to use for skin detection is hue saturation value (HSV) model. Same as the RGB color space, HSV color space contains three components: hue, saturation and value. Hue is defined as the color type, ranging between 0 and 360, each value corresponding to one color. Saturation is the intensity of the color, ranging from 0 to 100%, which is a shade of grey between black and white. Value represents the brightness of the color, which also ranges from 0 to 100%. Compared with RGB model on the lighting issue, HSV model can achieve better results (**Figure 4**). Therefore, we use the HSV color space to build up a skin-color model to segment face from non-face region.



(a) Test image **(b) RGB space**

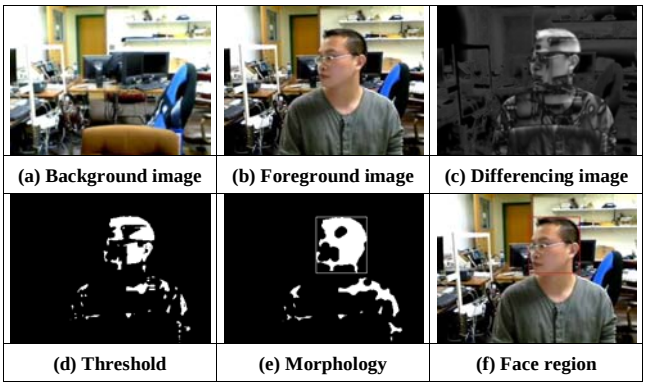
(c) HSV space **(d) Test result (HSV space)**

Figure 4 An example of skin detection

3.1.3 *Background subtraction technique*

The background subtraction technique is used to detect moving objects in videos from stationary cameras. Assuming that the camera is fixed and only the recorded subject's head is moving, the facial part can be easily identified by differentiating the foreground part containing head movement from the fixed background.

The most effective way to do localization is Motion History Image (MHI). MHI is defined as the cumulative motion over a specified duration and used for identifying object motion [Bobick, 2001]. Sometimes it is subdivided into two types: point-wise and region-wise. Point-wise MHI is defined as the amount of pixel-wise spatial-temporal variation [Essa, 1997], while region-wise MHI is the amount of region-wise variation [Park, 2004]. Motion history images, regardless of the type, can be achieved by the absolute difference between adjacent frames to get the frame motion silhouettes, and then the profiles of these silhouettes is used as a threshold in binary images to get the motion history over the duration. To improve the robustness, image noises are removed through morphological operations (erosion and dilation) (**Figure 5**).



(a) Background image **(b) Foreground image** **(c) Differencing image**

(d) Threshold **(e) Morphology** **(f) Face region**

Figure 5 An example of face region localization by Motion history image.

In short, although Viola and Jones's algorithm has been tested by many researchers that it shows very robust performance against illumination conditions and scale invariant, it shows poor performance against in-plane rotation ($>15^0$) and out-of-plane rotation. Therefore, this algorithm is well suited for the upright frontal face detection. On the contrary, skin-color model is rotationally invariant, but it is sensitive to light changes. Finally, motion history image is suitable for tracking face motion. Based on the above analysis (**Table 1**), a hybrid approach, combining the virtues of each technique, will achieve a better detection performance than using only one approach. For example, Yoon and Kim [Yoon and Kim] adopted an integrated approach for human detection, which uses skin color and motion information to first find the candidate foreground objects, and then uses a more sophisticated technique to classify the objects.

As a result, for frontal to frontal face detection, we choose Viola and Jones's algorithm, while for profile to frontal face detection, we use skin color model combined by motion history image in order to overcome the pose problem.

Table 1. Comparison with adaboost, skin-color model, and motion history image (MHI)

	Pros	Cons
Adaboost	Scale invariant, efficient image representation, very fast on non-faces	Templates of known object is required, poor performance for rotated object
Skin color model	Processing very fast, orientation invariant under certain lighting conditions, stable object representation	Device dependant, user dependent, environment dependent (lighting problem)
Motion history image	Fit for head pose, good for foreground segmentation	Lighting sensitive

3.2 Feature extraction

Feature extraction refers to the problem of matching features extracted from frame to frame in long sequences of images or video. The choice of feature depends on a variety of factors, like the kind of objects being tracked, the overall conditions of illumination, and the image brightness contrast. Good features can be found using eigenvalue analysis: two large eigenvalues mean a reliable pattern.

Optical flow is a commonly used method for tracking the movement of feature points by finding all the pixels in an image. Given point (μ_x, μ_y) in image I_1, find the point $(u_x + \delta_x, u_y + \delta_y)$ in image I_2 that minimizes ε, assuming brightness constancy within a small region surrounding the feature point

$$\varepsilon(\delta_x, \delta_y) = \sum_{x=u_x-\omega_x}^{u_x+\omega_x} \sum_{y=u_y-\omega_y}^{u_y+\omega_y} \left(I_1(x,y) - I_2(x+\delta_x, y+\delta_y)\right)$$

The process of computing optical flow includes three steps: the first to detect facial feature points in the first image according to the specific face model (Shi and Tomasi algorithm); secondly, to match the corresponding points in another image according by pyramidal L. K.'s optical flow; third, to calculate the displacement vectors of all these feature points (e.g. Euclidean distance, angle).

3.3 Support vector machine for facial expression recognition

3.3.1 SVM method

Support vector machine (SVM), presented by Vapnik in 1990s [Cortes and Vapnik, 1995], has been widely used in pattern classification. Based on statistical learning theory, SVM is similar to the root of multilayer neural networks. Given some sets of data classified, when a new data added, SVM can predict which set it should belong to. Thus, we can consider SVM as a "black box", and just push data into SVM and use the output.

With a series of training data $M:\{(x_1, y_1), \{x_2, y_2\}, \cdots, (x_m, y_m)\}$ $(i = 1,2,\cdots,m)$, the associated labels are $y_i = 1$ for Class 1 and $y_i = -1$ for Class 2. If the training data is linearly separable, there exists a series of separating planes called hyperplanes, represented by a plane equation $w^T x + b = c$ $(-1 < c < 1)$, where w is an m-dimensional vector of weights, and b is known as bias. When $c = 0$, the separating hyperplane is in the middle of two hyperplanes with $c = 1$ and $c = -1$. Those points along the hyperplanes ($c = 1$ or -1) are called support vectors. If the margin, defined as the distance between the hyperplanes with support vectors, is maximized, the middle hyperplane within such margin is called the optimal hyperplane (**Figure 6**). Then it is very clear that SVM's generalization ability depends on the location of the optimal hyperplane [Shigeo, 2005], which can be obtained by minimizing $\phi(w) = \frac{1}{2}\|w\|^2$, and, for all training examples (x_i, y_i), by subjecting to the constraints $y_i(w^T x_i + b) \geq 1$.

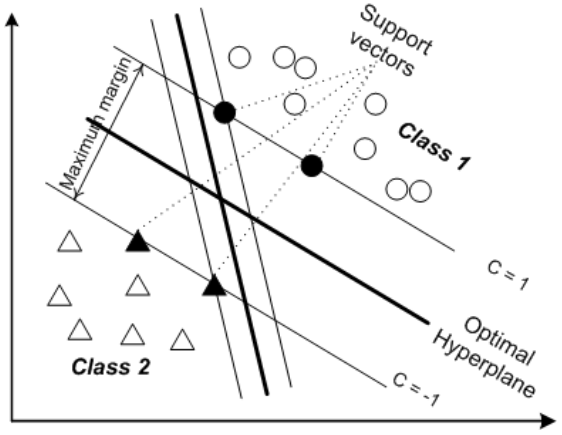
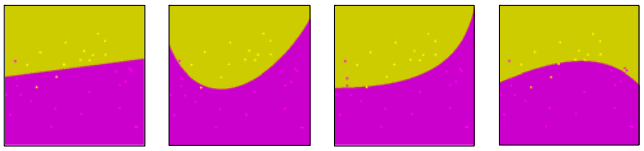


Figure 6 Optimal hyperplane in 2D space

In training a support vector machine, it is necessary to select an appropriate kernel function (**Figure 7**) and its parameters if a classification problem is not linearly separable in the input space. Kernel function performs as mapping the original input space into a higher dimensional feature space. Typically kernel function used for SVM is Radial Basis Function (RBF), which is well suitable for dealing with the nonlinear relationship between attributes and class labels (**Figure 7(b)**).



(a) Linear **(b) RBF** **(c) Polynomial** **(d) Sigmoid**

Figure 7 Sketch of different kernel function's training result on *C* = -100

3.3.2 SVM application in facial expression classification

In general, facial expression recognition is performed directly on raw image data. In considering the high storage requirement, Principal Component Analysis (PCA) is always used to reduce the dimensionality of the training data and capture the important components of datasets. Based on PCA analysis of a set of images with different facial expressions, the eigenvalues and eigenvectors are calculated, and based on these values and vectors, the decomposition coefficients are calculated for each image. It is assumed that the decomposition coefficient vectors represent the location of each image in the eigenspace, and vectors for similar images should have similar values. Based on this assumption, some representative vectors are chosen together with the corresponding expression labels as the inputs for the training stage. The trained SVM model is subsequently used to classify later-on in real time. The steps of SVM applied in image are described below, and the virtue of PCA is displayed in step 6 and 7.

Step1: representing images as vector;
Step2: setting up a template database;
Step3: computing the mean;
Step4: subtracting the mean from training images;
Step5: calculating the covariance matrix *COV*;
Step6: calculating the Eigenvectors and Eigenvalues of the covariance matrix *COV*;
Step7: choosing Eigenvectors to use and form the transformation matrix *T*;
Step8: deriving the new data set;
Step 9: selecting appropriate kernel function and parameters to train the datasets;
Step 10: once the training model is built, using it to predict the new dataset.

4. EXPERIMENT RESULT

In this section, we describe an architecture that automatically finds faces in the video stream and codes each frame with respect to four categories (neutral, surprise, happy, and angry) in real time. Our facial expression recognition system includes three parts: i) locating human face; ii) extracting facial features, iii) classifying facial expressions. The system is developed in the Visual studio 2005 platform using Intel's OpenCV library and LIBSVM [Chang and Lin, 2001].

We adopt the above introduced hybrid method combined by Adaboost, skin color model (HSV) and motion history image (MHI) to do face detection (**Figure 8**). Input image or video sequence is captured by a single camera mounted in the workplace.

For MHI-based detection, first frame is used as a still image, and then the current frame is subtracted with the still frame to find the moving object (face), then connected objects are separated and noises are removed by morphological operation (erosions and dilations), and finally, we mask the subtracted frame containing objects(face).

For skin-based detection, we first build a skin pixel model based on a sample skin image and then converting this training image into HSV format, then analyzing each pixel among this image, and, finally computing the mean/variance for all three channels for a skin pixel. Once the model is built, for each image in the sequence, each pixel is compared to the skin pixel model. If the means of the pixels for all three channels are within a specified distance from the standard deviation of the means of the pixel model, these pixels are classified as skin. Next, we compute the connected components for the potential candidates, the largest of which is considered to be the face. Also, some morphological operations are used to filter out really small candidates. To be more robust, we apply each technique respectively, and the intersection is part of every one. It has been evaluated by 6 participants and it shows great performance in locating rotated face (**Figure 9**).

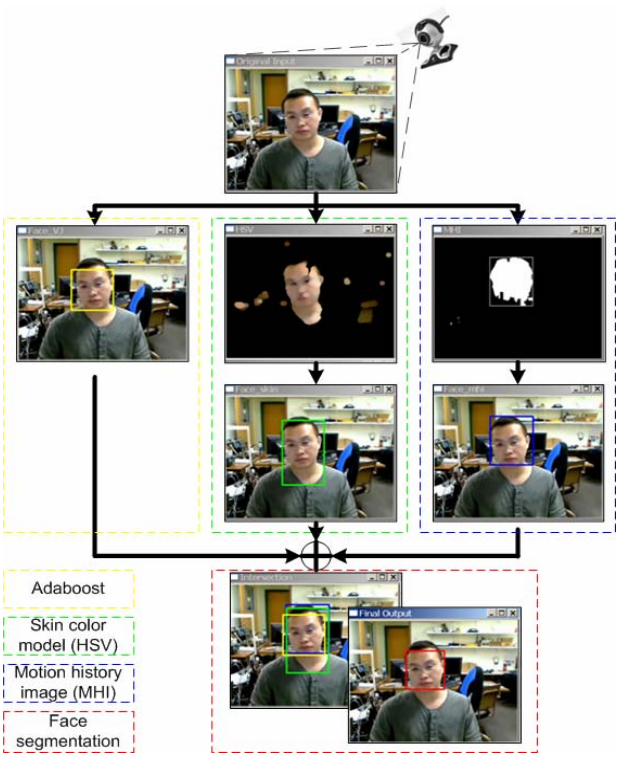


Figure 8 Flow chart of face detection in our recognition system

Figure 9 Sampled face detecting results

The purpose of doing feature extraction is to reduce the dimensionality of input data. To facilitate this process, we choose a neutral expression as a reference, and then compare them with other expression (e.g. happiness, anger, and surprise) using optical flow method, in order to find some key features which can be used to distinguish the facial expression change (**Figure 10**).

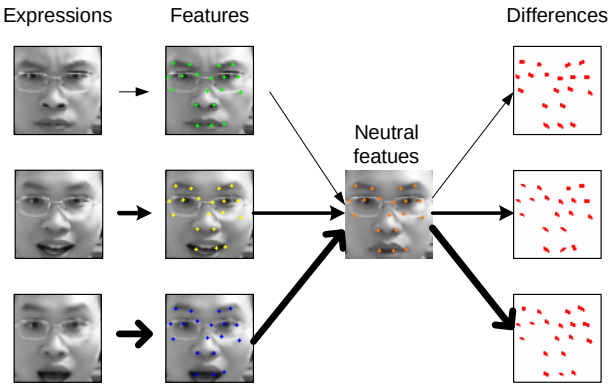


Figure 10 Differences between neutral expression with other expression (angry, happy, and surprise)

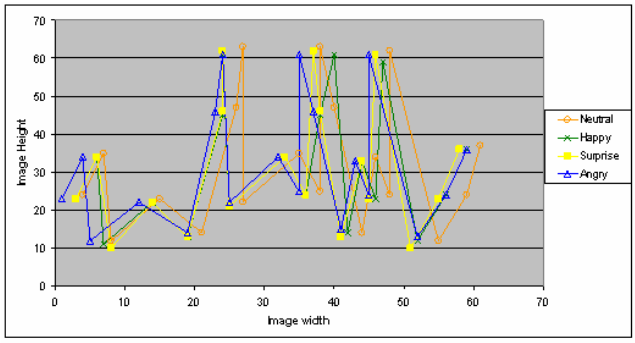


Figure 11 Locations of features point in different expression image

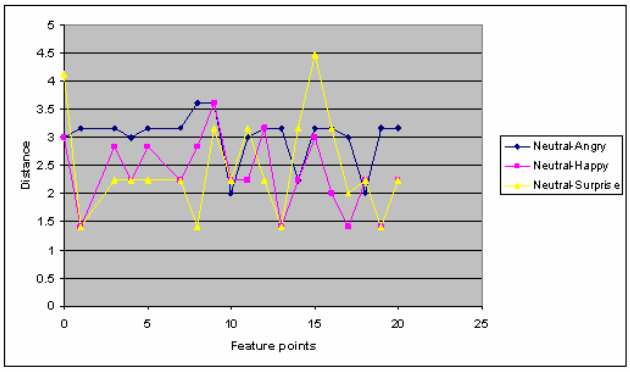


Figure 12 Feature points' movement of other expressions referenced to feature positions of neutral

From **Figure 11** and **Figure 12**, we do think that each expression has a corresponding characteristic curve. And the key points' info such as location in the image, Euclidian distance, or moving direction (angel) could be helpful for expression classification.

To test this, we selectively choose 19 feature points, and each point has four attributes: location (x, y), distance and angle. We use Support vector machine to train these data and predict the test images (**Figure 13**). On average, the recognizing accuracy is about 81.5%.

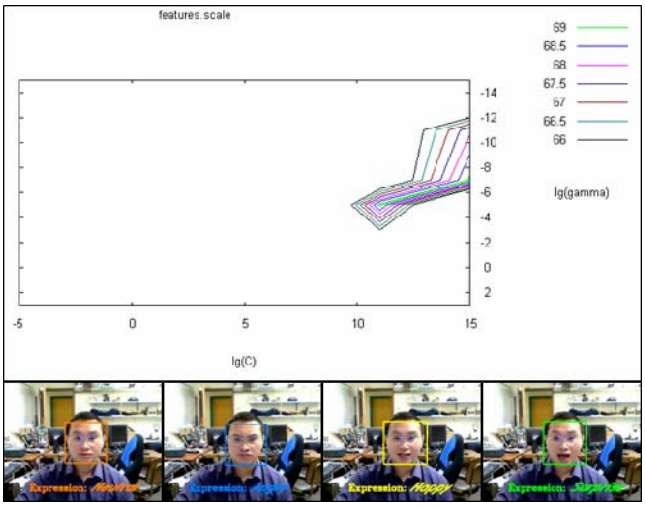


Figure 13 Demonstration of SVM's train and predict

5. CONCLUSIONS

This paper surveys the existing techniques for facial expression recognition and analyzes the strengths and limitations of template-based approach, feature-based based approach and appearance-based approach. To solve the facial pose problem, we present a hybrid algorithm combining Viola and Jones's algorithm, skin color model and motion history image, which achieves satisfactory results in real-time face localization, especially in the detection of rotated face. We also do a tentative test that some geometric feature info (location, distance and angle) could be used for facial expression recognition. Nevertheless, accurately positioning feature is the decision point for the recognition level.

6. REFERENCES

[1] Bobick, A.F. and Davis, J.W. (2001) The Recognition of Human Movement Using Temporal Templates. IEEE Transactions on PAMI, Vol. 23, 257–267.

[2] Chang, C. C. and Lin, C. J. (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[3] Cortes, C. and Vapnik, V. (1995) Support-vector network. Machine Learning 20, 273-297.

[4] Chen, Q., Georganas, N. D. and Petriu, E. M. (2007) Real-time vision-based hand gesture recognition using Haar-like features. IMTC 2007. Warsaw, Poland, May 1-3, 2007.

[5] Ekman, P. and Friesen, W. V. (1975) Unmasking the face: a guide to recognizing emotions from facial clues. Imprint Englewood Cliffs, N.J.: Prentice-Hall, 1975.

[6] Essa, I. and Pentland, A. (1997) Coding, analysis, interpretation, and recognition of facial expressions. IEEE Trans. on Pattern Analysis and Machine Intelligence 19(7), pp. 757–763.

[7] Gunes, H. and Piccardi, M. (2005) Fusing Face and Body Gesture for Machine Recognition of Emotions. IEEE International Workshop on Robots and Human Interactive Communication.

[8] Intel OpenCV Documents.

[9] Lades, M. et al. (1993) Distortion invariant object recognition in the dynamic link architecture. IEEE Trans. on Computers, vol. 42, pp. 300 – 311.

[10] Lien, J. J., Kanade, T., Cohn, J. F., and Li, C. C. (1998) Automated facial expression recognition based on FACS Action Units. Proceedings of FG's 98, April 14-16, Nara, Japan.

[11] Li, S. Z. and Jain, A. K. (2005) Handbook of Face recognition. Springer New York.

[12] Mehrabian, A. (1972) Nonverbal Communication. Aldine-Atherton, Chicago, Illinois.

[13] Michel, P. and Kaliouby, R. E. (2003) Real time facial expression recognition in video using support vector machines. Proceedings of the 5th international conference on Multimodal interfaces, Vancouver, Canada, pages 258 – 264.

[14] Park, H, and Park, J. I. (2004) Analysis and recognition of facial expression based on point-wise motion energy. Springer Berlin/Heidelberg, vol. 3212.

[15] Picard, R.W. (1997) Affective Computing. MIT Press, Cambridge, MA.

[16] Shigeo Abe. (2005) Support vector machines for pattern classification. Springer-Verlag, London, 39-40.

[17] Shih, F. Y., and Chuang, C. (2004) Automatic extraction of head and face boundaries and facial features. Information Sciences, vol. 158, pp. 117-130.

[18] Tian, Y., Kanade, T. and Cohn, J. (2001) Recognizing Action Units for Facial Expression Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23 (2): 97-116.

[19] Turk, M. A. and Pentland, A. P. (1991) Face Recognition Using Eigenfaces. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3-6 June 1991, Maui, Hawaii, USA, pp. 586-591.

[20] Viola, P. and Jones, M. (2001) Robust Real-time Object Detection. Second International Workshop on Statistical and Computational Theories of Vision Modeling, Learning, Computing, and Sampling, Vancouver, Canada, July 13.

[21] Wilson, P. I. and Fernandez, J. (2006) Facial feature detection using Haar classifiers. JCSC 21, 4**.**

[22] Yoon, S. M. and Kim, H. (2001) Real-time multiple people detection using skin color, motion and appearance information. International Workshop on Robot and Human Interactive Communication, pp. 331–334.