Recognizing Activity for Computer Assistant Posture Learning[1]

Changbo HU, Tao FENG , Songde MA and Hanqing LU
National Lab of Pattern Recognition, P.O.Box 2728, Beijing,100080, China
cbhu,fengt,masd,hqlu@nlpr.ia.ac.cn

Abstract

This paper presents a system to retrieve human exercising video by means of activity recognition. By a Gaussian model of skin color in HSI color space, human face and hands regions are grouped and extracted firstly. Then human face and hands are tracked applying the constraints of motion and region continuity. And the normalized motion parameters of faces and hands are used to match with the parameter curves of models in PCA framework. Examples of Taiji postures are presented and discussed to illustrate our method.

1 Introduction

Human activity recognition is not only an attractive goal of computer vision theory, but also has many applications such as visual surveillance, virtual reality, human-machine interaction and aerobic analysis. When learning exercising postures from video teacher, learners can be provided much convenience with the help of computer by making an activity to let computer know which part of video he needs.

It is obvious that this human action recognition process is composed of two stages. The first and most important stage is human motion tracking and analysis. The second stage is the interpretation of this motion. In the first stage, two kinds of methods are often adopted. One approach applies human models, varying from 1 dimension to 3 dimensions [1]. These model-based methods all need to label out human body parts. The second approach does not apply models. In [2], by computing the optical flow fields between consecutive frames and dividing each flow frames into small windows, the motion magnitude in each window is summed to form a high dimensional feature vector for recognition. Similar idea is used by J.W.Davis[3] to interpret human motion by motion-energy images (MEI) and motion-history images (MHI).

However, both approaches have their shortcomings. In the model approach, motions of each part enable us to recognize very complex activities. But automatically tracking and labeling the human body parts are rather difficult especially in complex scene, and occlusion is another challenge. In addition, computation load need also be considered. In the non-model approach, although the computation of motion is easy to implement, yet the significance of whole image motion can not be easily determined. Without strict prior constraints the interpretation result may be completely wrong.

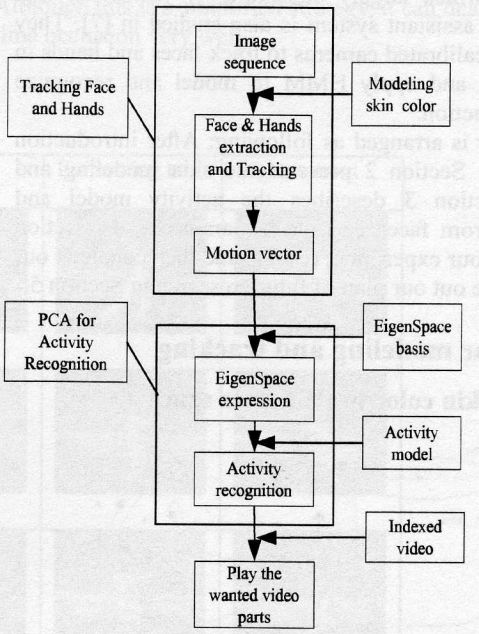

Fig.1 The pipeline of our system

According our experience, it is often enough to recognize many kinds of human actions by a few parts of human body for some applications. Faces and hands are important parts of human body and have a unique color property that facilitates tracking. In our system, we model skin color by a 2D Gaussian model using H and S in HSI color space because H and S are insensitive to illumination. Then the three areas are corresponded by motion prediction and minimization the region matching error between successive frames. After the face and hands are found and tacked, the motion vector is calculated from the position parameters and then we begin interpreting stage.

[1] This work is funded by research grants from the NSFC （No.69865005） and the 973 research project（G1998030500）.

There are also various approaches in this stage, such as hidden markov model (HMM) [4,9], primary component analysis (PCA) [5,10,11] and probabilistic framework [6,12]. Similar to [5], we apply PCA to model and recognize activity because PCA or an eigenspace representation can construct activities in a largely reduced dimensionality. In the system, normalization of the activity property is considered to assure right recognition by different human or under different imaging condition. The system is simple in design and computation, but robust and useful in recognition. The process of our system is demonstrated in figure 1.

We use Taiji video to testify our system. Taiji is a quintessence exercise of ancient China, which is learned all around the world. Its primary function is building body, although it is said to be a very trenchancy fighting skill. Taiji is performed largely depending on hands moving. Taiji training assistant system is also studied in [7]. They use two self-calibrated cameras to track faces and hands in 3 dimension, and apply HMM to model and recognize gestures and action.

This paper is arranged as following: After introduction in Section 1, Section 2 presents the skin modeling and tracking. Section 3 describes the activity model and recognition from faces and hands parameters. In section 4,we present our experiment results, and then conclude our work and give out our plan of future research in Section 5.

2 Skin color modeling and tracking

2.1 Model skin color in HSI color space

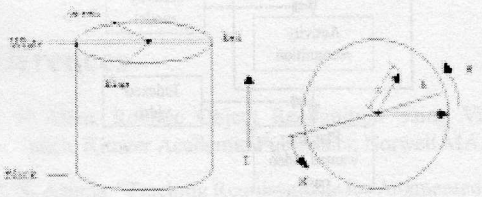

Fig.2 HSI color space.

In Rits Eye system [8], face skin color is modeled in RGB space with a 3D Gaussian model. Here we use HSI color space and model the distribution of skin color of faces by a 2D Gaussian model. The definition of HSI is demonstrated in Figure 2. Among the three components, we apply H and S, because they are insensitive to illumination. Skin color region can be separated from non-skin color region by Mahalanobis distance from the center of Gaussian distribution. Because H is the angle around the circle normalized to [0 1), the distance of 0 and 0.9 is not 0.9 but 0.1. As most test pixels distribute around the skin hue h (h is about 0.1 in our experiment), in order to avoid computation error caused by some pixels whose hue is around 1, we rotate the start point of H to the opposite angle of the hue of skin during our computation (see Figure 2), i.e.

$$H' = Mod(H + 0.5 - h, 1) \qquad (2.1)$$

where $Mod(x, y)$ is the remainder of x / y.

A general 2D Gaussian distribution is expressed by,

$$p(x) = \frac{1}{2\pi}\|\Sigma\|^{-1/2} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right] \qquad (2.2)$$

According a predefined skin pixels extracted manually, we learn the center and the covariance matrix by

$$\mu = \frac{1}{n}\sum_{k=1}^{n} x_k, \quad \Sigma = \frac{1}{n}\sum_{k=1}^{n}(x_k-\mu)(x_k-\mu)^T \qquad (2.3)$$

Where $x = [H', S]^T$.

If a color vector whose Mahalnobis distance to the center of the Gaussian distribution, expressed by

$$D = (x-\mu)^T \Sigma^{-1}(x-\mu) \qquad (2.4)$$

is less than a predetermined threshold, it is declared as a skin color; otherwise it is declared as non-skin color.

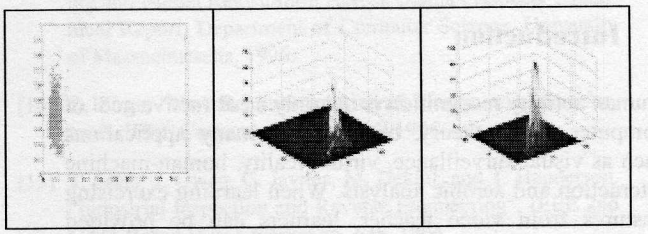

Fig3. Left: the distribution of skin color in HS plane; Middle: the histogram ; Right: the estimated Gaussian distribution in HS plane.

We used sample images of human faces to compute the distribution of skin color. Fig.3 shows the distribution and histogram of skin color and the estimated Gaussian distribution in HS plane.

2.2 Faces and hands tracking

After the classification of skin and non-skin pixel, the skin pixels are grouped to skin region. The skin region is represented in white pixel, outside region is represented by black pixel. We then track the regions by area constraint and motion estimation. Notice that many regions may exist in one frame. What we need to do is finding out the right one. Under the assumption of small continuous motion, we can describe the motion by orthogonal transformation. And the transformation matrix is

$$A = \begin{bmatrix} \cos\theta & -\sin\theta & \Delta x \\ \sin\theta & \cos\theta & \Delta y \\ 0 & 0 & 1 \end{bmatrix} \qquad (2.5)$$

Clearly three parameters are independent. So we can

estimation them respectively from past frames. i.e., exists filter Γ, and parameter vector $\vec{T} = (\theta, \Delta x, \Delta y)^T$. Then we have

$$\begin{aligned} \vec{T}_{t+1} &\sim N(\Gamma(\vec{T}_0, \ldots, \vec{T}_t), \delta\vec{T}(t+1)) \\ &= N(\vec{T}^*_{t+1}, \delta\vec{T}(t+1)) \end{aligned} \quad (2.6)$$

By (3.1), skin region in t frame $R(t)$ is likely to be

$$\hat{R}(t+1) = A_{\vec{T}^*_{t+1}} R(t) \quad (2.7)$$

and among the field

$$R'_{t+1} = \bigcup_{d \in [-3\delta\vec{T}(t+1), 3\delta\vec{T}(t+1)]} A_{\vec{T}^*_{t+1}+d} R(t) \quad (2.8)$$

Then skin regions inside this aspect are matched using a similarity criteria

$$S(R_1, R_2) = e^{-(\alpha_1 \cdot p + \beta_1 \cdot n)} \cdot (1 - \alpha_2 \cdot p - \beta_2 \cdot n) \quad (2.9)$$

under orthogonal transformation $R_2 = \hat{A}R_1$.

where $\hat{A}$ is the 'best' orthogonal transform in the meaning of maximizing $S(R_1, R_2)$.

$$\begin{aligned} p &= Card_{\neq 0}(I_1 \cap I_2^c) / Card_{\neq 0}(I_1 \cup I_2) \\ n &= Card_{\neq 0}(I_2 \cap I_1^c) / Card_{\neq 0}(I_1 \cup I_2) \end{aligned} \quad (2.10)$$

measure the positive and negative error and $\alpha_1, \beta_1, \alpha_2, \beta_2$ are degree of emphasis on these errors. $Card_{\neq 0}$ denotes the number of pixels which are not black.

The final corresponding region is selected as its similarity S is maximum among all candidates.

Then we obtain a sequence of motion parameters of the region.

If there is no candidate region inside the prediction aspect or the similarity is too low, occlusion is assumed to happen and the region in frame $(t+1)$ is assumed as Equation (2.7). Although this is a simplified strategy, it can deal with short time occlusion very well.

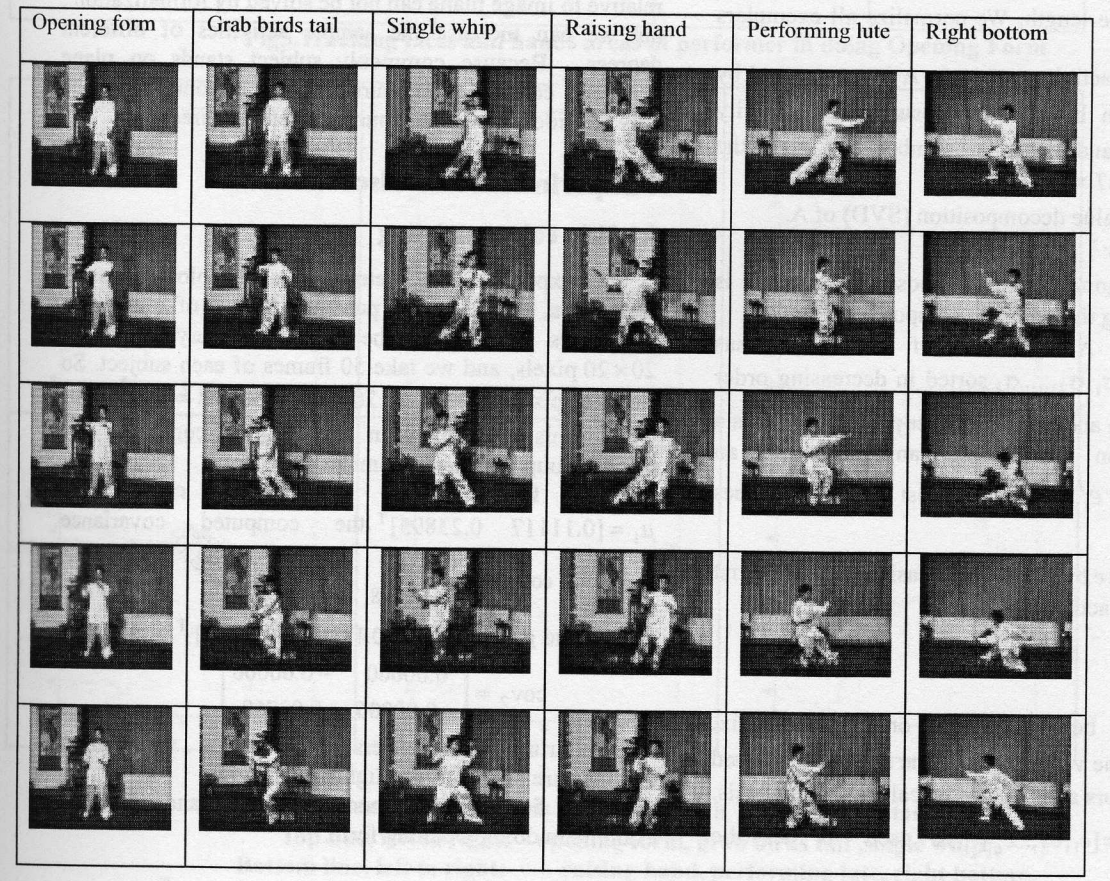

Fig.4. Six Taiji forms by an expert.

3 Activity model and recognition

3.1 Model activity using PCA

Figure4 is six Taiji forms of Yang style we intent to recognize: opening form, grab birds tail, single whip, raising hands, performing lute and right bottom from left to right. We note them as O,G,S,Rh,P,Rb.

We use the position parameters of face and hands to describe the changing with motion, which are face centroid (fx, fy), hands centroid (lx, ly) and (rx, ry).

Since the absolute coordinates depend on the position that the person locates in the image, we apply the difference of them as the parameters. i.e.,

$$\begin{aligned} flx &= fx - lx \qquad frx = fx - rx \\ fly &= fy - ly \qquad fry = fy - ry \end{aligned} \tag{3.1}$$

Then we construct the model of activity of face motion by exemplars using PCA. We present the main steps of model activity as following:

Step 1.Construct an activity e_i^j. e_i^j is a nT column vector of i th exemplar of activity j. n is the number of parameters. T is the length. We normalize all exemplars to a unit length.

Step 2.Construct activity matrix A. A is constructed by arranging e_i^j column by column. Assume total activity class number is J and exemplar number is I, then the dimension of A is $nT \times IJ$ $(IJ < nT)$.

Step3. Singular value decomposition (SVD) of A.

$$A = U \Lambda V^T \tag{3.2}$$

U is an orthogonal matrix whose dimension is $nT \times nT$ representing the principle component directions in the training set. Λ is a matrix of $nT \times IJ$ who has IJ singular values $\sigma_1, \sigma_2 ..... \sigma_k$ sorted in decreasing order in diagonal direction and zeroes in other positions; V^T is a matrix of dimension $IJ \times IJ$. We can approximate an instance of activity e_i^j using the largest q singular values $\sigma_1, \sigma_2 ..... \sigma_q$.

Step 4.Descript e by activity basis. Select first q columns of U as activity basis.

$$e \approx \sum_{l=1}^{q} c_l U_l \tag{3.3}$$

where $c_l = e \bullet U_l^T$ because U is a orthogonal matrix. That is, projecting the vector $\bar{e}$ onto the subspace spanned by the q basis vectors and U_l can be called activity basis. Row vector $c^T = [c_1, c_2, ... c_q]$ is used to describe activity e.

3.2 Recognizing activity

Recognizing activity means matching newly performed activity with exemplars. To a newly performed activity α, Equation (3.3) is solved by minimization the form

$$E(\bar{c}) = \sum_{j=1}^{nT} (\alpha_j - \sum_{l=1}^{q} c_l U_{l,j}) \tag{3.4}$$

Upon recovery of the coefficient vector $\bar{c}$, we classify α to the exemplar activity e whose coefficient vector $\bar{m}$ is nearest to $\bar{c}$.

The important issue in recognition stage is normalization. In order to recognize the human activity, we must consider the imaging change not due to human motion. Commonly we consider three influencing factors, i.e., translation, scale and rotation. Since relative coordinate is applied, translation is naturally avoided. Then in order to eliminate the influence of imaging scale and the anthropometric difference of subjects, we compute the Euclidean distance of activities by normalized coefficients.

$$d = \sum_{i=1}^{q} \left| c_i / \|\bar{c}\| - m_i / \|\bar{m}\| \right| \tag{3.5}$$

Without depth information, the rotation of subject relative to image plane can not be solved by formalization , but we can increase the sample activities of different degrees. Because commonly subject stands on plane floor and can not rotate surrounding horizontal axis, we only consider rotating surrounding vertical direction.

4 Experiments and discussion

4.1 Skin color modeling

In the experiment of determine skin color Gaussian parameters, we select the patches of faces of 8 different performers manually. The size of every patch is 20×20 pixels, and we take 50 frames of each subject. So the total pixel number is $20 \times 20 \times 50 \times 8 = 160,000$. We do this experiment in two lighting condition (the second group images are dimmer).

The first group computed center is $\mu_1 = [0.11117 \quad 0.23895]^T$,the computed covariance matrix is $\text{cov}_1 = \begin{bmatrix} 0.00070 & -0.00008 \\ -0.00008 & 0.00802 \end{bmatrix}$;

The second group , $\mu_2 = [0.11312 \quad 0.24191]^T$

$$\text{cov}_2 = \begin{bmatrix} 0.00060 & -0.00006 \\ -0.00006 & 0.00768 \end{bmatrix}$$

Comparing the parameters of two groups , we can find that they are insensitive to lighting changing.

Figure 5 is a performer faces and hands tracking example in doing Opening form.

4.2 Taiji forms recognition

In the experiment of activity model and recognition, we increase the activity examples by asking subjects to

perform activity in every 10° through -30° to 30° in horizontal direction. We do not care the rotation in vertical direction because our camera is fixed to a degree and subject can not rotate in vertical direction. So we have 7 exemplars of every activity performed by two experts, and total exemplar number is $7\times2\times6=84$. We determine the start time by the curves changing from a constant state and end time as the curves entering static state. We normalize the time of an activity to 100 frames. The matrix A is of dimension 400×84.

Figure 6 is the graphs of six forms performed with subject orientation o°. Figure 7 is examples of performing Opening form and Grab birds tail with orientation 10° and -30°. In fact , *fly* and *fry* do not change. Only *flx* and *frx* change with orientation. Here the orientation does not mean the face orientation, but the rotation relative to the exemplar form.

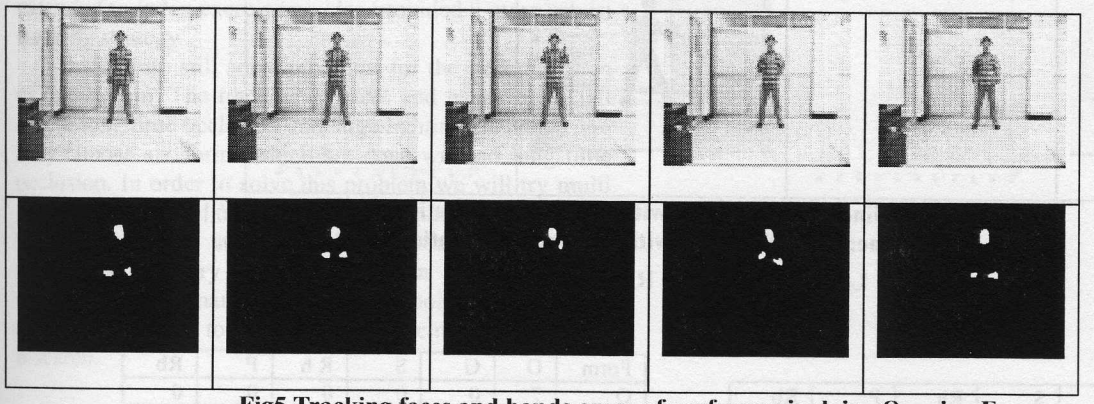

Fig5.Tracking faces and hands areas of performer in doing Opening Form

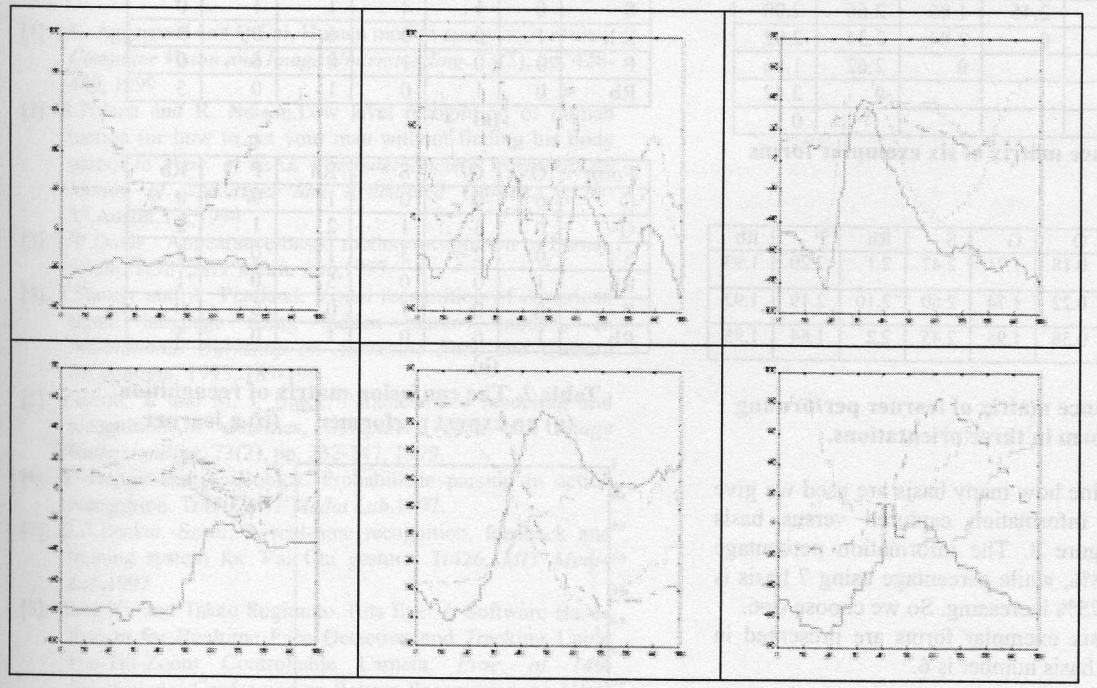

Fig.6. Graphs of six forms performed with subject orientation o°.
Top line, left to right: opening form, grab birds tail ,single whip;
Bottom line, left to right: raising hand, performing lute, right bottom.
flx -.- *frx* ..... *frx* _ _ *fry*

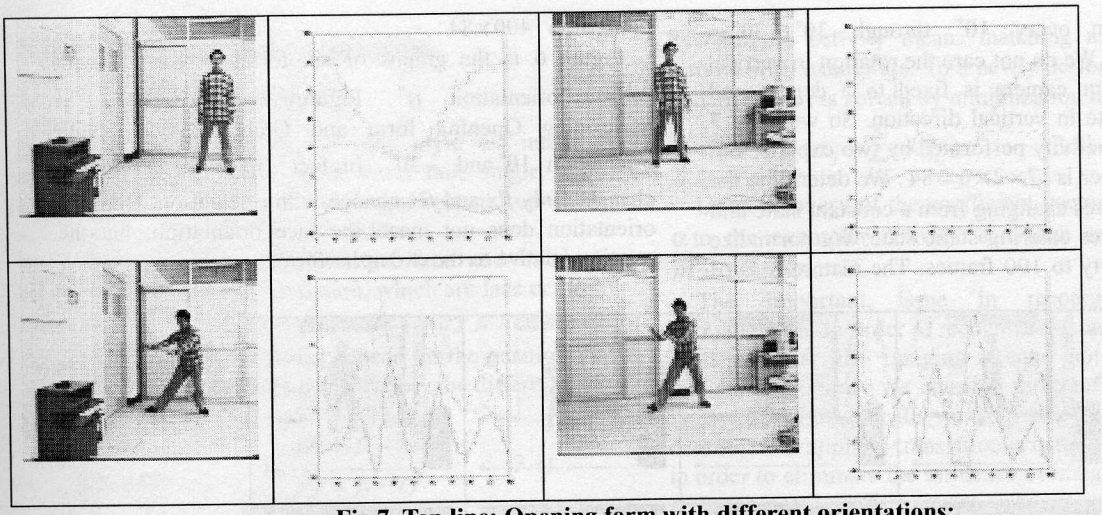

Fig.7. Top line: Opening form with different orientations; Bottom line: Grab Birds tail with different orientations

Left: 10° **Right:** -30°

Form	O	G	S	Rh	P	Rb
O	0	1.63	2.55	2.17	2.37	2.01
G		0	2.45	1.86	2.66	2.00
S			0	2.04	2.34	2.02
Rh				0	2.62	1.86
P					0	2.42
Rb						0

Table 1 (a) Distance matrix of six exemplar forms

Forms	O	G	S	Rh	P	Rb
Opening form 0°	0.18	1.71	2.47	2.1	2.20	1.93
Opening form 10°	0.22	1.74	2.50	2.10	2.19	1.92
Opening form -30°	0.58	1.95	2.45	2.2	1.84	1.95

Table 1(b) Distance matrix of learner performing opening form in three orientations.

In order to determine how many basis are used we give out the cumulative information captured versus basis number curve in Figure 8. The information percentage using 6 basis is 79.48%, while percentage using 7 basis is 80.23%, with only 0.75% increasing. So we choose q=6.

The distances of six exemplar forms are presented in Table 1(a), where the basis number is 6.

The recognition result is described by confusion matrix in Table 2. Every form is performed 7 times with orientation between -30° and 30°. The confusion matrix (a) in table 2 is done by an expert and (b) is done by a learner.

Form	O	G	S	R h	P	Rb
O	7	0	0	0	0	0
G	0	4	0	1	0	2
S	0	1	4	1	1	0
R h	1	0	0	6	0	0
P	0	0	1	0	6	0
Rb	0	1	0	1	0	5

(a)

Form	O	G	S	Rh	P	Rb
O	6	0	0	1	0	0
G	0	3	1	2	1	2
S	0	2	3	1	0	1
Rh	1	0	0	5	0	1
P	0	0	1	0	6	0
Rb	1	0	0	1	0	5

(b)

Table 2. The confusion matrix of recognition (a) an expert performer (b) a learner

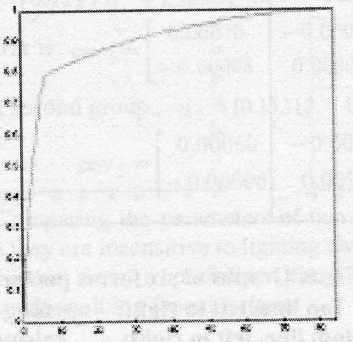

Fig.8 Cumulative information captured by 84 basis

5 Conclusion

In this paper we present a system retrieving exercising video by activity recognition. We track human face and hands by skin color with motion and area continuity constraints, and obtain activity vectors from hands and face position parameters. Activity vectors are modeled by PCA method with normalization. Experiments show that our system has potential to be utilized in practical environment. We believe, in fact, our system can be extended to index exercising video provided a more robust tracking strategy.

Yet there are still some problems for the generalization of our system. The tracking of faces and hands often fail due to long time occlusion and large lighting changing. We only choose six forms which are done with no long time occlusion. In order to solve this problem we will try multi cameras and extend our system to recognize all forty forms in Taiji of Yang style. Besides, normalization in our method is not very accurate. We plan to compute the orientation of human by other anthropologic feature, or apply 3D vision to solve this problem in our future research.

References

[1]. J.K. Aggarrwal and Q.Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*. 73(3), pp. 428-440, 1999

[2]. R.Polana and R. Nelson.Low level recognition of human motion (or how to get your man without finding his body parts). In *Proc. of IEEE Computer Society Workshop on Motion of Non-Rigid and Articulated Objects*, pp.83-88,Austin,TX,1994

[3]. J.W.Davis . Appearance-based motion recognition of human action. Tr387,*MIT Media Lab*,1997.

[4]. T.Starner and A. Pentland. Visual recognition of American signal language using hidden markov models. In *International Workshop on Automatic Face and Gesture Recognition*, 1995, 189-194.

[5]. Yacoob, Y. and M. J. Black, Parameterized modeling and recognition of activities, *Computer Vision and Image Understanding*. 73(2), pp. 232-247, 1999.

[6]. Y. Ivanov and A. Bobick. Probabilistic parsing in action recognition. Tr450, *MIT Media Lab*,1997.

[7]. D.A.Becker. Snsei: A real-time recognition, feedback and training system for T'ai Chi gesture. Tr426, *MIT Media Lab*,1997

[8]. Gang Xu and Takeo Sugimoto. Rits Eye: A Software-Based System for Realtime Face Detection and Tracking Using Pan-Tilt-Zoom Controllable Camera. *Proc. of 14th International Conference on Pattern Recognition*, pp.1194-1197, 1998.

[9]. J. Schlenzig, E. Hunter, and R.Jain. Recursive identification of gesture inputs using hidden markov models. *Proc. Second Annual Conference on Applications of Computer Vision*, pp.187-194, December 1994.

[10]. A. Bobick and J. Davis. An appearance-based representation of action. *Proc. of 13th International Conference on Pattern Recognition*, pp307-312, 1996.

[11]. S.X.Ju,M.J.Black and Y.Yacoob.Cardboard people: A parameterized model of articulated image motion. *Proc. International Conference on Face and Gesture*, Vermout,pp.561-567,1996.

[12]. C. Bregler, S.M. Omohundro et.al. Probabilistic models of verbal and body gesture, in *Computer Vision for Human--Machine Interaction*, R. Cipolla and A.Pentland, Eds. Cambridge University Press, 1998.