HAND POSTURE RECOGNITION IN VIDEO USING MULTIPLE CUES

Liang Sha[1], *Guijin Wang*[1], *Anbang Yao*[1], *Xinggang Lin*[1], *Xiujuan Chai*[2]

[1]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[2]Nokia Research Center, Beijing, 100084, China

ABSTRACT

Hand posture conveys profound information for computer vision applications, but the articulated hand structure and restraint capture condition cast a tough obstacle on practical implementation, especially in real time video. This paper presents a framework to recognize hand postures in consecutive video frames. Mixture of Gaussian skin/non skin models is constructed for hand region detection, followed by particle filter to track hand. Then a soft-decision scheme based on extended Histogram of Orientated gradient is proposed to refine the best posture region and recognize it from pre-defined posture set. Experimental result shows promising performance under various capture conditions.

Index Terms—Hand posture recognition, object tracking, particle filter, histogram of oriented gradient, mixture of Gaussian model

1. INTRODUCTION

Hand posture has attracted much attention in computer vision community like in robot vision[2], sign language recognition[3], virtual reality[4] and human computer interaction(HCI)[5].

Intensive researches have been devoted to recognize hand postures [1][6]. Currently, there are two kinds of approaches in this field: 3D model based and 2D appearance based. The 3D model based approach estimates hand configurations to reconstruct 3D hand model [11][12], it is plagued by capture condition, computational burden and user-dependant calibration. The 2D appearance based approach aims to extract hand feature from images, and map it to a certain pre-defined posture configuration. Compared with the 3D model based approach, its computation complexity is relatively mild and the user independence could be delicately achieved. There are two core issues to be addressed for such an approach: 1) How to detect hand region from background; 2) How to extract discriminative and robust feature from a coarse detection/tracking result for further recognition. Considering the first issue, skin color [7][13] or feature points like finger tips[14] are usually employed as the a-prior cue. While finger tips meet self-occlusion problems, skin color shows robust efficiency on deformable hand appearance. As for the second problem, color or feature points suffer from illumination and orientation variance thus lack of robustness and discrimination. However, other features like silhouettes and edges of hand region [15] rely on the precise extraction from detection/tracking cue, and also they are weak in self-occlusion. Consequently, it is necessary to make necessary calibration on hand location and extract features inside the silhouettes, being independent of capture condition or user. From the above analysis, the hand posture recognition still faces significant challenges, such as posture collection, modeling, tracking in consecutive frames, feature selection and evaluation strategy.

This paper focuses on tracking hand location in monocular video and recognizing postures from pre-defined posture set. The posture set is 2D appearance based, and the video is captured by portable computer and computer attached video camera. To solve aforementioned difficulties, this paper proposes a solution combining multiple cues and a coarse-to-fine refined posture recognition strategy. Our contributions concentrate on three aspects. First, to effectively initialize hand region and improve tracking performance, we train and online update a mixture of Gaussian skin/non skin model to detect skin region from background. Second, to robustly and discriminatively characterize postures to be recognized, we extend histogram of oriented gradient (HOG) as posture feature descriptor. Third, to provide an accurate location for recognition and determine the best posture, we propose a "soft-decision" search procedure based on the tracking result. During such a procedure, descriptors from the posture set are employed to select among candidate regions; and the posture that compromises similarity and discrimination is finally settled. Experimental results show that our system could track the hand and recognize its posture in satisfying performance.

The whole paper is organized as follows: Section 2 described the system and its principle; Section 3 discusses the skin region detection, and posture recognition strategy in detail; experimental results are given in Section 4 and we summarize the whole paper in Section 5.

2. SYSTEM FRAMEWORK

Fig 1 gives the framework of our tracking system. From video containing hand posture as input, it recursively tracks

the hand in consecutive frames using particle filter (PF) tracker [8]. From PF's coarse location, we search posture location from grid candidate regions. Among those regions, HOG feature is extracted to provide similarity and discrimination information with potential posture set. Finally, the posture is recognized by a compromised way called "soft decision".

To adapt to current capture condition, we train offline and periodically online update the mixture Gaussian skin/non skin model to improve the performance of traditional PF tracker.

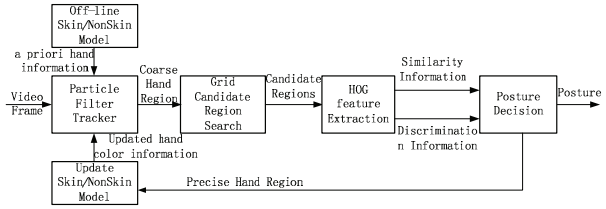

Fig.1 System Framework

3. POSTURE RECOGNITION METHODS

3.1 Skin Detection

Skin reflects strong centralization in color distribution, thus an appropriate skin color model gives important cue for posture background subtraction. In practice, skin model is used as a classifier to divide the original image into two parts, skin region and non-skin region. At present, skin model (and corresponding non-skin model) is constructed based on statistics of manually labeled skin color pixels from picture database, the learned model often results in two kinds of formats: histogram-based model and density function model. Compared with the former one [7], density based model could be trained from relatively small amounts of data and is agile to update or reform in online environment. In our system, a mixture Gaussian probability based skin model and a corresponding non-skin model are constructed in the form below,

$$p(\mathbf{x}) = \sum_{i=1}^{N} \frac{\omega_i}{(2\pi)^{3/2}\left|\Sigma_i\right|^{1/2}} \exp\{-\frac{(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i)}{2}\}, \quad (1)$$

where x is the pixel's color vector and the contribution of the *i*th Gaussian kernel is determined by a scalar weight ω_i, mean vector μ_i and diagonal covariance matrix Σ_i.

For a pixel, it is labeled as a skin pixel if its color vector (in RGB space in our context) satisfies

$$\frac{P(\mathbf{x}|skin)}{P(\mathbf{x}|nonskin)} \geq \lambda, \quad (2)$$

where $\lambda \in (0,1)$ is a threshold which can be adjusted to trade-off between miss-detections and false positives. In our system, considering the detection precision and computation burden, the skin detection procedure is implemented in off-line and on-line steps.

During the off-line step, to get the initial model for detection, we capture videos under different illumination, posture and cluttered background. Then, we manually label images as skin region and non-skin region containing over 150 frames and 40000 pixels to generate candidate skin Gaussian set $\{\omega_i, \mu_i, \Sigma_i\}_{i=1}^{16}$ and non skin Gaussian set $\{\omega_j, \mu_j, \Sigma_j\}_{j=1}^{16}$. Fig.2 shows the contour plot of our skin and non-skin Gaussian model. The color in RGB space is marginalized among gray axis and chrominance axes and then calculates probability in either model. It could be observed that the skin model distribution is a high bias on red half plane, while non-skin model distribution is relatively homogenous along the gray axis. This separation is the basis for the good performance of our skin detector, which is further demonstrated in Section 4.

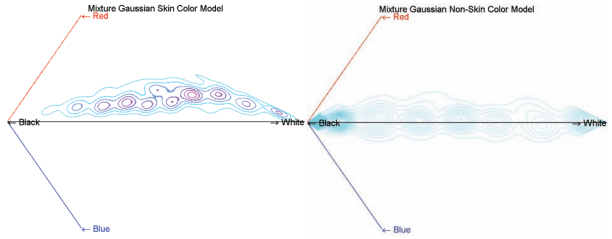

Fig.2 Contour plot for skin model

During the on-line step, to accelerate the computation, use 5 kernels from skin set and 5 kernels from non-skin set for skin detection as in (1) and periodically update 1 of each set from other candidate kernels. To be exactly, suppose $\{\omega_k, \mu_k, \Sigma_k\}_{t-1}$ is the kernel in contemporary skin set which is least like the skin region among all 5 kernels, we choose $\{\omega_k, \mu_k, \Sigma_k\}_t$ from candidate set if

$$\begin{cases} \max\{p(\mathbf{x}|\{\omega_k, \mu_k, \Sigma_k\}_t) \\ p(\mathbf{x}|\{\omega_k, \mu_k, \Sigma_k\}_t) > p(\mathbf{x}|\{\omega_k, \mu_k, \Sigma_k\}_{t-1}) \end{cases} \quad (3)$$

,where $\{\omega_k, \mu_k, \Sigma_k\}_t \in$ candidate kernel set

At an extreme situation, none of the candidate kernels satisfies the current observation, we replace $\{\omega_k, \mu_k, \Sigma_k\}_{t-1}$ with a new kernel $\{\omega_k, \mu_k, \Sigma_k\}_t$ with

$$\begin{aligned} \mu_{k,t} &= (1-\rho)\mu_{k,t-1} + \rho\mathbf{x}_{center} \\ \Sigma_{k,t} &= (1-\rho)\Sigma_{k,t-1} + \rho(\mathbf{x}_{center} - \mu_{k,t})^T(\mathbf{x}_{center} - \mu_{k,t}) \end{aligned} \quad (4)$$

,where $\rho = \alpha P(\mathbf{x}_{center}|\mu_{k,t}, \Sigma_{k,t})$,

where $\mathbf{x}_{center}$ is the color vector of the newly observed hand region's center pixel, and $\alpha \in [0,1)$ is the learning rate. Therefore, the weight of each kernel is updated using (5):

$$\begin{aligned} \omega_{k,t} &= (1-\alpha)\omega_{k,t-1} + \alpha \\ \omega_{i,t} &= (1-\alpha)\omega_{i,t-1}, i \neq k. \end{aligned} \tag{5}$$

Taking advantage of skin/non skin detection, the non skin pixels are labeled as low confidence, thus the output image provides a-prior cue for PF tracker.

3.2 Posture Recognition using HOG and Soft Decision

As is analyzed above, it is critical to choose a discriminative and robust feature to express posture appearances. Histogram of oriented gradient (HOG) [9] was originally proposed as the feature to express the shape of the object, and was used to detect pedestrian. It shows robust performance against illumination, target scale variance and certain tolerance on rotation. We extended such a feature to express hand posture, and discovered that it is effective for hand posture estimation both in discriminative power and robustness.

During the HOG extraction procedure, the input region is first divided into 4 cells, each of which is converged to gradient in X and Y component by basic gradient operator. Then the gradient orientation and magnitude are calculated using cardinal-polar coordinate transform. The local 1-D histogram of orientation gradient for each cell is accumulated. Finally, we join the local histograms into HOG in a contrast-normalized way [10]. From training, we generate hand posture template set. Fig.3 shows the discriminative capacity of HOG feature. The similarity is calculated by Bhattacharyya coefficient between posture template HOG features and scaled to gray value [0 255]. The discrimination between postures (Inter-Posture) is significant, and that between samples of the same posture (Intra-Posture) is relatively mild. Thus, it is persuasive that HOG feature could be used as a cue to recognize hand postures.

In our scenario, HOG descriptor is not only employed as recognition cue but also utilized for calibrating coarse tracking result. Due to the character of random sampling, restriction of particle number and observer measure of the PF tracker, the location estimate often results in quite a discrepancy from the accurate place. However, its accuracy is vital in recognizing the posture. To this end, we introduce a "soft-decision" scheme for posture recognition. During the recognition process, we first choose the tracking resultant region and its 4 neighborhood regions as in Fig.4. For each candidate region, we then calculate its similarity to each posture in our posture set. Considering the best region would be similar enough to a certain posture template and discriminative enough with other postures, we introduce a principle taking into account both similarity and discrimination. As in (6), the best region would either possess both the largest similarity and the largest discrimination among candidate postures, or compromise similarity and discrimination in the largest degree. Then the posture of current frame is the posture which the best region is most similar to.

To be exactly, let $P = \{p_i\}_{i=1}^{M}$ be the candidate posture set, $L = \{l_j\}_{j=1}^{N}$ be the candidate search region set. For each candidate region l_j, we calculate its similarity with each candidate posture p_i, i.e. $P_j^{(i)} = P(l_j | p_i)$. Then we reorder $P_j^{(i)}$ and extract largest similarities $P_j^{(1)}, P_j^{(2)}$ the inter class discrimination $D_j^{(1)} = P_j^{(1)} - P_j^{(2)}$ and their corresponding postures $p_j^{(1)}, p_j^{(2)}$, therefore, we extend L to incorporate similarity information $L' = \{l_j, p_j^{(1)}, p_j^{(2)}, P_j^{(1)}, D_j^{(1)}\}_{j=1}^{N}$. Finally, the best location l^* and posture p^* are extracted from L'.

$$l^* s.t. \begin{cases} \max(P_l^{(1)}) \\ \max(\mathrm{D}_l^{(1)}) \end{cases} \text{or.} \begin{cases} P_{l^*} < \max(P_l^{(1)}) \\ \mathrm{D}_{l^*}\text{-}\max(\mathrm{D}_l^{(1)}) > thredhold \end{cases} \tag{6}$$

(a)Inter-Posture (b)Intra-Posture

Fig.3 Discrimination of Posture HOG, Intra Class and Inter Class

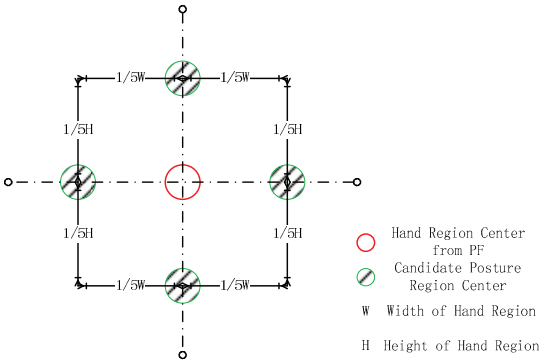

Fig. 4 Posture Soft Decision Search Region

4. EXPERIMENTAL RESULTS

To evaluate the proposed framework, we construct a test system using a Pentium 4 1.7GHz portable computer and a USB video camera with 1.3M pixel CCD, and the system could process real time video of resolution 320x240 pixels and frame rate 15fps. We update the Gaussian kernel every 10 frames and the parameters are M=10, N=5 as in Sec.3. In this section, we test our algorithm with real time video under different illumination, background texture, hand motion speed and posture scale/rotation change.

Table 1 gives numerical results, it demonstrates that our system improves tracking performance of PF tracker and achieves promising recognition performance, which is further shown in Fig.5. Each image series of Fig. 5 has three images, which are skin detection image, tracking location from PF (in red circle) and recognition output (corrected posture location in green circle and recognized posture sign on the up left corner). It could be observed that most of the non skin region is subtracted to aid the tracker. Under the soft-decision scheme, the coarse location is corrected to provide the most beneficial feature for posture discrimination. Notice that when background cluttering or ill white balance happens, the skin image or gradient information is disturbed, as in (b) and (e), but our system could still recognize the right posture. That is the benefits due to we combined color and HOG feature together during the whole tracking/recognition process. Currently, our system could not deal with over-exposure and drastic hand motion. The reason is that under the above two conditions, color feature based tracker could not well judges target from background, especially when ghost shadow or drift occurs.

5. CONCLUSION

In this paper, a posture recognition system containing multiple cues and soft decision tracking calibration is introduced. Our future work would focus on more robust feature and measurement fusion for posture recognition.

REFERENCES

[1] S. Thieffry, "Hand gestures", in *The Hand*, pp. 488-492, Philadelphia, PA: Sanders, 1981.

[2] T. I. Cerlinca, S. G. Pentiuc, R. D. Vatavu, and M. C. Cerlinca, "Hand posture recognition for human-robot interaction," *Proc. of Workshop on Multimodal interfaces in Semantic interaction*, 2007.

[3] J. S. Kim, W Jang, Z Bien, "A dynamic gesture recognition system for the Korean sign language (KSL) ," *IEEE Transactions on Systems, Man and Cybernetics*, Part B, 1996.

[4] A. G. Hauptmann and P. McAvinney, "Gesture with speech for graphics manipulation," *International Journal of Man-Machine Studies*, vol. 28, pp. 231-249, 1993.

[5] T. S. Huang, V. I. Pavlovi'c, "Gesture modeling, analysis, and synthesis," *Proc. of IEEE International Workshop on Automatic Face and Gesture Recognition*, 1995.

[6] Y. Wu, T. Huang, "Hand modeling, analysis and recognition," *IEEE Signal Processing Magazine*, vol. 18, pp. 51-60, 2001.

[7] M. J. Jones, J. M. Rehg, "Statistical color models with application to skin detection", *International Journal of Computer Vision*, vol.46, pp.81-96, 2002.

[8] M. Isard, A. Blake, "CONDENSATION - conditional density propagation for visual tracking", *International Journal of Computer Vision*, vol.29, pp.5-28, 1998.

[9] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition*, 2005.

[10] K. Onishiy T. Takiguchiz Y. Arikiz, "3D Human Posture Estimation Using the HOG Features from Monocular Image", *Proc. of IEEE International Conference on Pattern Recognition*, 2008.

[11] A. J. Heap and D. C. Hogg. "Towards 3-D hand tracking using a deformable model", In *2nd International Face and Gesture Recognition Conference*, 140.145, Killington,VT, October 1996.

[12] C. Tomasi, S. Petrov, and A. K. Sastry. "3D tracking = classification + interpolation". *Proc. 9th Int. Conf. on Computer Vision*, vol.2, pp.1441.1448, Nice, France, 2003.

[13] R. Lockton and A. W. Fitzgibbon, "Real-time gesture recognition using deterministic boosting," Proc. British Machine Vision Conference, vol.2, pp.817-826,Cardiff, UK, 2002.

[14] B. Stenger, "Template-Based Hand Pose Recognition Using Multiple Cues", Proc. Asian Conf. on Computer Vision, 2006.

[15] Y. Wu and T. S. Huang. "View-independent recognition of hand postures," *Proc. of Computer Vision and Pattern Recognition*, vol. 2, pp.88-94, Hilton Head, SC, June 2000.

TABLE I

NUMERICAL RESULT

Video Number	Frame Number	Labeled Postures	Posture Kinds
14	5892	4802	10
Tracked Rate(PF)	Tracked Rate(Soft Decision)	Recognition Rate	
91.5%	93.1%	81.8%	

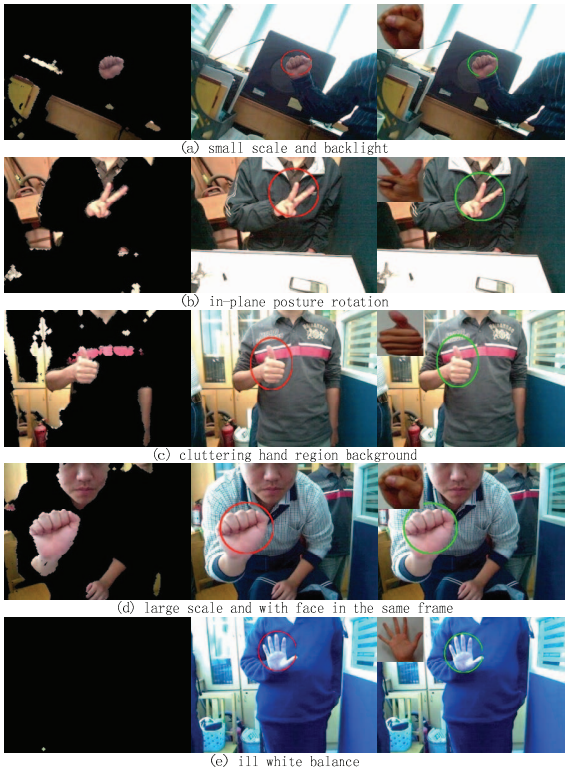

Fig.5 Posture Recognition Examples