A REAL-TIME HAND GESTURE RECOGNITION METHOD

Yikai Fang[1]*, Kongqiao Wang*[2]*, Jian Cheng*[1] *and Hanqing Lu*[1]

[1] National Lab of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China
{ykfang, jcheng, luhq}@nlpr.ia.ac.cn
[2]Nokia Research Center, No.11 He Ping Li Dong Jie, Nokia House 1,Beijing 100013,China
kongqiao.wang@nokia.com

ABSTRACT

Compared with the traditional interaction approaches, such as keyboard, mouse, pen, etc, vision based hand interaction is more natural and efficient. In this paper, we proposed a robust real-time hand gesture recognition method. In our method, firstly, a specific gesture is required to trigger the hand detection followed by tracking; then hand is segmented using motion and color cues; finally, in order to break the limitation of aspect ratio encountered in most of learning based hand gesture methods, the scale-space feature detection is integrated into gesture recognition. Applying the proposed method to navigation of image browsing, experimental results show that our method achieves satisfactory performance.

1. INTRODUCTION

With the development of ubiquitous computing, current user interaction approaches with keyboard, mouse and pen are not natural enough for them. On PC platform, there are applications such as interactive entertainments and augmented reality requiring more natural and intuitive interface. For mobile or hand held devices, their relatively small size leads to limited input space and encumbered experience with tiny keyboard or touch screen. Hand gesture is frequently used in people's daily life. It's also an important component of body languages in linguistics. So a natural interaction between humans and computing devices can be achieved if hand gestures can be used for communication between human and computing devices.

Vision based hand gesture interface has been attracting more attentions due to no extra hardware requirement except camera, which is very suitable for ubiquitous computing and emerging applications. Methods for vision based hand gesture recognition fall into two categories: 3D model based methods and appearance model based methods. 3D model may exactly describe hand movement and its shape, but most of them are computational expensive to use. Recently there are some methods to obtain 3D model with 2D appearance model such as ISOSOM and PCA-ICA in [1]and [2].

In this paper, we focus on appearance model based method. There have been a number of research efforts on appearance based method in recent years. Freeman and Weissman recognized gestures for television control using normalized correlation [3]. This technique is efficient but may be sensitive to different users, deformations of the pose and changes in scale, and background. Cui and Weng proposed a hand tracking and sign recognition method using appearance based method [4]. Although its accuracy was satisfactory, the performance was far from real-time. Just et al introduce modified census transform into hand gesture classification [5]. For the purpose of classifying each gesture respectively, their method obtains fairly good results. While the performance in recognition experiments was not so satisfactory and the recognition result of different gesture great disparity. Elastic graphs were applied to represent hands in different hand gestures in Triesch's work with local jets of Gabor filters [6]. It locates hands without separate segmentation mechanism and the classifier is learned from a small set of image samples, so the generalization is very limited.

The performance of vision based gesture interaction is prone to be influenced by illumination changes, complicated backgrounds, camera movement and specific user variance. Many researchers have made effective efforts to deal with these problems. Lars and Lindberg used scale-space color features to recognize hand gestures [7]. In their method, gesture recognition method is based on feature detection and user-independent while the authors showed real-time application only under uniform backgrounds. Mathias and Turk developed a vision gesture interface with extended Adaboost for wearable computing, named HandVu [8]. It's insensitive to camera movement and user variance. The hand tracking acquired promising results, but the segmentation was not so reliable. Moreover all hand gesture images are required to have the similar aspect ratios, which restrict the scope of applications. In this paper, inspired by Lars and Lindberg's works and HandVu, we present a robust real-time gesture recognition method. The main idea is to segment hand with color and motion cues generated by detection and tracking. Then scale-

space feature detection method is used to recognize hand gestures. Our method is not confined by aspect ratio of hand image and can deal with cluttered background. Its also immune to camera movement in virtue of stable hand tracking. The whole process of gesture recognition is as follows:

(1) Firstly hand detection with Adaboost is used to trigger tracking and recognition.

(2) Then Adaptive hand segmentation is executed during detection and tracking with motion and color cues.

(3) Finally, scale-space features detection is applied to find palm-like and finger-like structures. Hand gesture type is determined by palm-finger configuration.

2. THE PROPOSED METHOD

2.1. Hand detection

Most of hand detection methods are sensitive to complicated background. Skin color based hand detection is unreliable for the difficulty to be distinguished from other skin-colored objects and sensitivity to lighting conditions. Approaches using shape models require sufficient contrast between object and background. There have been some effort to detect hand in grey image like Adaboost in [9][10]. It's similar to method in face detection and is adopted in our hand detection.

Hand detection in our method is an initial step of interaction. It's important for a gesture interface as it functions as a switch to turn on the interface. The detection uses extended Adaboost method which adopts a new type of featurefour box feature like image (d) and (e) in Fig.1.[9]. It's a feature type similar to the diagonal rectangle feature in an extension of Viola and Jones's work in [11]. It allows for almost arbitrary area comparisons since the rectangles' locations and sizes are less constrained; even overlapping areas are permitted.

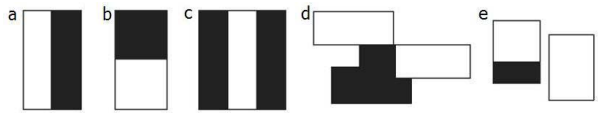


Fig. 1. (a-c) traditional feature type. (d-e) extended feature type.

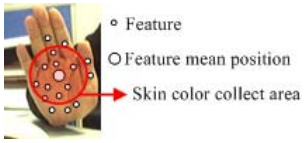
2.2. Hand tracking

It is a challenging task to track the articulated objects. Although shape based methods achieve better results for rigid objects, it is not suitable for the articulated objects, such as hand. Texture or appearance based methods have been improved to be more robust for the non-rigid objects. Some approaches resort to background modeling with limitation of stationery camera. Optical flow based method can produce good results for tracking when the object exhibit limited deformations.

In our method, we use a multi-modal technique which combines optical flow and color cue [12] to obtain stable hand tracking. Flock of features, which is first presented in [13] to describe behavior of birds or herds is adopted to make the traditional optical flow method feasible in the articulated object tracking. "Flock" means a loose global constraint on features' position, for example KLT features. It's a feature updating mechanism which replacing lost features with new ones. In every frame during tracking, there are two constraints: no two features must be closer to each other than a threshold distance and no feature must be further from the feature median than another threshold distance. Features violating the above restrictions will be removed [12].

2.3. Hand segmentation

Once the hand is detected, the color of hand is collected from the neighborhood of features mean position. Considering tradeoff between computational cost and accuracy of description, we use a single Gaussian model to describe hand color in HSV color space. In Fig.2(a), the bigger white dot denotes features' mean position. Only features within the circle are used to get instant hand color model. Compared with normalized RGB histogram in [12], our method can get better segmentation results. Histogram method is based on the assumption that no other exposed skin color part of user in the certain area around the hand. If there are wooden objects or part of user's face passing by such area, the histogram will deviate and segmentation results will be rapidly degraded. In that case our method can get better results like Fig.2(b).



(a) Skin color collect method

(b) Hand segmentation results

Fig. 2. Hand Segmentation method and results.

2.4. Gesture recognition

In general, recognizing various hand configurations is a difficult and largely unsolved problem. Ong and Bowden [10]

distinguished hand shapes with boosted classifier tree and obtained fairly good results. However, their method is time consuming and unpractical for interactive interfaces. Hanning Zhou and T. Huang made effort to recognize static hand gestures using local oriental histogram feature distribution model, but background in experiments are quite simple and sleeve color and texture are restricted [14]. Kolsch used fanned boosting detection for classification and got nearly real time results. In his method, all gestures' template should have identical resolutions and the hand areas must have identical aspect ratios [8]. Since Lindberg made seminal work on scale-space framework for image geometric structures detection [15], scale-space features detection have been widely applied in object recognition, image registering, etc. For planar hand shape, the scale-space feature detection can be used to detect blob and ridge structures, i.e. palm and finger structures. Blobs are detected as local maxima in scale space of the square of the normalized Laplacian operator [7].

$$\nabla^2_{norm} L = t(L_{xx} + L_{yy}) \tag{1}$$

L_{xx} and L_{yy} are Gaussian derivative operators at scale t along two dimensions of image. Elongated ridge structures, usually represented as ellipses are localized where the ridge detector

$$\mathcal{R}_{norm} L = t^{3/2}\left((L_{xx} - L_{yy})^2 + 4L^2_{xy}\right) \tag{2}$$

assumes a local maximum in scale-space [15]. Ellipse parameters such as orientation and axis length are defined by a windowed second moment matrix in (3) as described in [15]. L_x and L_y are Gaussian mixture derivative operator and g is Gaussian kernel at a certain integration scale t_{int}.

$$\Sigma = \int_{\eta \in R^2} \begin{pmatrix} L_x^2 & L_x L_y \\ L_x L_y & L_y^2 \end{pmatrix} g(\eta; t_{int}) d\eta \tag{3}$$

Fig. 3. Blob and ridge detection of hand gestures.

Lars Bretzner in [7] using multi-scale color features to detect planar hand posture structures, i.e. fingers and palms. However in practice they also put many restrictions on background according to their video demo. In our method, we reduce computation expense by detect multi-scale feature across binary image and make hand gesture interface more practicable by combine this feature detection with hand tracking and segmentation.

3. EXPERIMENTS

In order to validate the proposed method, we design a hand-driven navigation of image browsing interface. We defined six

Fig. 4. Gesture definition.

gestures as shown in Fig. 4 for navigation interface which is indispensable in interactive interface: These gestures execute operations like shifting focus. LEFT, RIGHT, UP and DOWN are for shifting focus in four directions; OPEN and CLOSE are used to open and close preview of selected image. CLOSE is also interpreted as STOP when shifting focus.

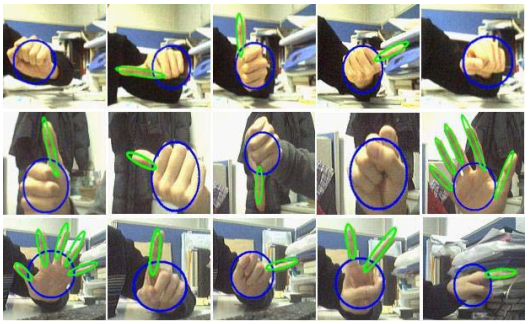


Fig. 5. Sample results for gesture recognition.

$$s = \exp\left[\frac{1}{k}\sum\nolimits_{(u,v)} \log\left(F(u,v) - P(u,v)\right)\right] \tag{4}$$

We probe the separability of the six defined gestures with spectrum analysis like (4)[9]. and get their separability values s as shown in Tab.1.$F(u,v)$ and $P(u,v)$ are Fourier transform of sample hand image and neutral grey image of the same size k. From the Tab.1, the OPEN has the highest separability value, which suggests it is easy to detect and distinguish from other gestures. Moreover, OPEN as the beginning is also coincident with user experience. So we define the OPEN as switch to open interface and train a detector for it using Adaboost with extended features in Section 2 .

Table 1. Separability value of six gestures.

Gesture	LEFT	RIGHT	UP	DOWN	OPEN	CLOSE
value	0.1744	0.1676	0.1414	0.1481	0.2108	0.1608

To assess the performance of this combined method with average accuracy, a total of 2596 frames were recorded in experiments. There are both simple and cluttered backgrounds and gestures were performed by different users. A few frames with different gestures are shown in Fig. 5. Tab. 2 shows results under simple (but not uniform) background and Tab. 3 contains results under cluttered background. Among the total of 2596 frames recorded, 2436 frames were recognized correctly. The average accuracy of recognition in experiment is 0.938. It can be noted that recognition accuracy for RIGHT are a bit lower than others. From the captured frames, it

Table 2. Recognition results under simple background.(Total: the total frames recorded. Correct: the correct recognized frames.)

Gesture	LEFT	RIGHT	UP	DOWN	OPEN	CLOSE
Total	222	185	205	228	212	282
Correct	218	176	203	225	211	278
Accuracy	0.982	0.951	0.990	0.987	0.995	0.986

Table 3. Recognition results under cluttered background. (Total: the total frames recorded. Correct: the correct recognized frames)

Gesture	LEFT	RIGHT	UP	DOWN	OPEN	CLOSE
Total	216	202	195	210	221	218
Correct	192	171	175	188	202	197
Accuracy	0.889	0.846	0.897	0.895	0.914	0.904

seems that gesture RIGHT is not as convenient as others to be put in proper view of camera scene for some right handedness. So there are some frames that contain non-standard gestures. But later we found that any finger besides thumb can be extended to the right freely to represent RIGHT without influence on performance. So if thumb is not so convenient for some people,they may try index finger, middle, or even pinkie. This makes the gestures much more usable. Gesture OPEN achieves relatively high accuracy with both backgrounds because it contains richer features in image and thus easier to be recognized.

In our experiments, with 320x240 pixels sized camera video as input, processing time for each frame is between 90–110 ms on a 2.8Ghz desktop computer with 256Mb RAM. The speed of our method satisfy real-time requirements in human computer interface. The image features are detected in scale from 5 to 55 with a scale interval of 2. Few special optimization steps are used in implementation.

4. CONCLUSIONS

Altogether our method combines fast hand tracking, hand segmentation and multi-scale feature extraction to develop an accurate and robust hand gesture recognition method. It takes advantage of color and motion cues acquired during tracking to implement adaptive hand segmentation. On the basis of segmentation, multi-scale feature extraction is executed and gestures are recognized with palm-finger decomposition. Extensive experiments show this method has promising performance with various hand gesture posture aspect ratios and under complicated backgrounds.

5. ACKNOWLEDGEMENT

The research was supported by National Natural Science Foundation of China (Grant No. 60121302) and NSF of Beijing (Grant No. 4072025). The authors would also like to thank Mr. Roope Takala and the funding support from Nokia Research Center.

6. REFERENCES

[1] Haiying Guan, Rogerio S. Feris, and Matthew Turk, "The isometric self-organizing map for 3d hand pose estimation," in *Proceedings of Int. Conf. on Automatic Face and Gesture Recognition*. Southampton, UK, Apr. 2006, pp. 263–268.

[2] Makoto Kato, Yen-Wei Chen, and Gang Xu, "Articulated hand tracking by pca-ica approach," in *Proceedings of Int. Conf. on Automatic Face and Gesture Recognition*. Southampton, UK, Apr. 2006, pp. 329 – 334.

[3] W. T. Freeman and C. Weissman, "Television control by hand gestures," in *Proceedings of International Workshop on Automatic Face and Gesture Recognition*. Zurich, Switzerland, June 1995, pp. 197–183.

[4] Y. Cui and J. Weng, "View-based hand segmentation and hand sequence recognition with complex backgrounds," in *Proceedings of 13th ICPR*. Vienna, Austria, Aug. 1996, vol. 3, pp. 617–621.

[5] Just A., Rodriguez Y., and Marcel S., "Hand posture classification and recognition using the modified census transform," in *Proceedings of Int. Conf. on Automatic Face and Gesture Recognition*. Southampton, UK, Apr. 2006, pp. 351–356.

[6] J. Triesch and C. von der Malsburg, "Robust classification of hand posture against complex background," in *Proceedings of Int. Conf. on Face and Gesture Recognition*. Killington, Vermont, Apr. 1996, pp. 170–175.

[7] Lars Bretzner, Ivan Laptev, and Tony Lindeberg, "Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering," in *Proceedings of Int. Conf. on Automatic Face and Gesture Recognition*. Washington D.C., May 2002, pp. 423–428.

[8] Mathias Kolsch, "Vision based hand gesture interfaces for wearable computing and virtual environments," *PHD Dissertation,UCSB*, 2005.

[9] Mathias Kolsch and Matthew Turk, "Robust hand detection," in *Proceedings of Int. Conf. on Automatic Face and Gesture Recognition*. Seoul, Korea, May 2004, pp. 614 – 619.

[10] Eng-Jon Ong and Richard Bowden, "A boosted classifier tree for hand shape detection," in *Proceedings of Int. Conf. on Automatic Face and Gesture Recognition*. Seoul, Korea, May 2004, pp. 889 – 894.

[11] M. Jones and P. Viola, "Fast multi-view face detection," *Technical Report TR2003-96,MERL*, July 2003.

[12] Mathias Kolsch and Matthew Turk, "Fast 2d hand tracking with flocks of features and multi-cue integration," in *Proceedings of CVPR Workshop on Real-Time Vision for HCI*, 2004.

[13] C. W. Reynolds, "A distributed behavioral model," *Computer Graphics*, vol. 21, pp. 446–458, Dec. 1987.

[14] Hanning Zhou, D.J. Lin, and T.S. Huang, "Static hand gesture recognition based on local orientation histogram feature distribution model," in *Proceedings of CVPR*, 2004, pp. 161–169.

[15] T. Lindeberg, "Feature detection with automatic scale selection," *IJCV*, vol. 30, pp. 77–116, June 2004.