Hand Gesture Recognition Using Fast Multi-scale Analysis

Yikai Fang[†] Jian Cheng[†] Kongqiao Wang[‡] Hanqing Lu[†]

[†]National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
Beijing, 100080
{ykfang, jcheng, luhq}@nlpr.ia.ac.cn

[‡]Nokia Research Center
Nokai House 1, No.11 He Ping Li Dong Jie
Beijing, 100013
kongqiao.wang@nokia.com

Abstract

Hand gesture has been used as a natural and efficient way in human computer interaction. Due to independence of auxiliary input devices, vision-based hand interfaces is more favorable for users. However, the process of hand gesture recognition is very time consuming, which often brings much frustration to users. In this paper, we propose a fast feature detection and description approach which can significantly speed up hand gesture recognition. Firstly, integral image is used to approximate Gaussian derivatives to calculate image convolution in feature detection. Then multi-scale geometric descriptors at feature points are obtained to represent hand gestures. Finally gesture is recognized with its geometric configuration. Experiments show that the proposed method needs much less time consumption while obtains comparative performance with its counterpart in literatures.

1. Introduction

With the development of ubiquitous computing, computer is becoming more and more important in our daily life. Computer applications require more and more unrestricted interaction between human and computers, which is a great challenge to traditional input devices such as mouse, keyboard or pen etc. Hand gesture is frequently used in people's daily life. It's also an important component of body languages in linguistics. Compared with those devices mentioned above, hand gestures are more natural in interaction. The use of hand as a device for human-computer interaction (HCI) makes HCI easy.

The key problem in gesture interaction is how to make hand gestures understood by computers. Extra sensors and instruments, such as data gloves, may be easy to collect hand configuration and movement. However, these equipments are usually expensive and bring much cumbersome experience to users. Vision based gesture interaction have many appealing characteristics. The prominent one is that it realizes a natural interaction between human and computers independent of external dedicated devices. In general, the literatures of hand gesture recognition fall into two categories: learning based and model based method. Learning based method get classifier or detector with machine learning (e.g. boosting) from the training data which is constructed by multi-cue features with plenty of sample images. Ong and Bowden [1] distinguished hand shapes with boosted classifier tree and obtained fairly good results. However, their method is time consuming and unpractical for interactive interfaces. Kolsch [2] used fanned boosting detection for classification and got nearly real time results, while the detector's training is computational expensive. What's more, the detector makes constraints on the resolution and aspect ratio of gesture template. As for the model based method, there are some researchers who have achieved satisfactory accuracy in hand gesture recognition. Lars and Lindberg used scale-space color features to recognize hand gestures [3]. In their method, gesture recognition is based on a stable hand gesture representation generated by a hierarchy of feature detection. This hand gesture representation is effective in recognition. Although the authors show nearly real-time application on a dual-processor computer, the computation costs expensively as feature detector and descriptor involve a great deal of large-scale Gaussian convolution.

Viola uses integral images as Haar wavelet features in rapid object detection [4]. Integral images allow for the fast implementation of box type convolution filters, which makes very fast feature extraction. The pixel of an integral image $I_\Sigma(xx)$ at point $\mathbf{x} = (x, y)$ has the intensity of the sum of all pixels of a rectangular region formed by the point $\mathbf{x}$ and the origin in the input image I, $I_\Sigma(\mathbf{x}) = \Sigma_{i=0}^{i\leq x}\Sigma_{j=0}^{j\leq y} I(i,j)$. With I_Σ, it needs only four additions to compute the sum of the intensities over any upright, rectan-

0-7695-2929-1/07 $25.00 © 2007 IEEE
DOI 10.1109/ICIG.2007.52

gular area. The computation cost is independent of its size. Herbert etc. [5] utilize integral images in robust feature detection and obtain satisfactory results. In this paper, inspired by the work in [5], we propose a fast multi-scale feature detection and description method for hand gesture recognition. We firstly approximate complex Gaussian derivatives with simple integral image in feature detection. Then multi-scale geometric descriptors at feature points are obtained to represent hand gestures. Finally gesture is recognized with its geometric configuration.

The rest of the paper is organized as follows. Section 2 describes the hand gesture representation by hierarchy of feature detection. Section 3 introduces how to approximate image convolution and Gaussian derivatives with integral image. Section 4 provides experiments results for our method. Finally we will conclude and propose some future directions.

2. Multi-Scale Features

Since Lindberg made seminal work on scale-space framework for geometric features detection [6], scale-space feature detection has been widely applied in object recognition, image registering etc. Lars uses scale-space feature detection to detect blob and ridge structures of hand [3], i.e. palm and finger structures. Blobs are detected as local maxima or minima in scale-space of the square of the normalized Laplacian operator.

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

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 Ellipse parameters such as orientation and axis length are defined by a windowed second moment matrix in (3) as described in [6]. L_x and L_y are Gaussian mixture derivative operator and g is Gaussian kernel at a certain integration scale t_{int}.

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

Gaussian derivatives in blob and ridge detector involve a great deal of large-scale image convolution in implementation. The computation cost of the detectors is expensive for real-time gesture interaction.

3. Fast Feature Detection and Description by Integral Image

Multi-scale local feature detection method in [3] for gesture recognition is computationally expensive. Simplification and approximation are possible ways to speed it. Rectangle features, which can be computed very rapidly using integral images, have been widely used in face detection and recognition. They are also used in learning based approach for hand gesture recognition. Ong [1] and Kolsch [2] use combination of rectangle features for rapid hand detection. In this paper, we present a fast feature detection and description method for hand gesture recognition. The proposed method utilizes a simple integral image to replace the complex Gaussian derivatives in the multi-scale feature detection.

3.1. Rectangle Filter

As shown by Lindberg in [6], Gaussian is optimal for scale-space analysis. However, the Gaussian needs to be discrete and cropped in practice. So aliasing still occurs as long as the resulting images are sub-sampled [5][7]. Lindberg shows that the property that no new structures can appear while going to lower resolutions have been proven in the 1D case, but it is still unknown in the 2D case. In practice, the Gaussian may not be necessarily indispensable as described in [7]. David Lowe shows the LoG approximation is effective in SIFT [8]. Further approximation in [5] with box filters get comparable performance with discrete and cropped Gaussians. Moreover the box filters can be computed much faster with integral images.

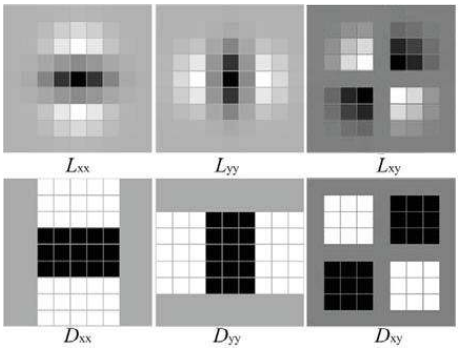

Figure 1. Discrete Gaussian derivative and their approximations.

Gaussian derivatives and the corresponding box filters used in our method are shown in Fig.1. The 9x9 filters in the first row are Gaussian derivates with $\sigma = 1.2$. The second row gives corresponding box filters. Derivates and

box filters at other scales are similar. Speed performance is improved just with these box filters. In [6], scale- space is usually imagined as image pyramids. A series of Gaussian filters at different scales repeatedly smooth the images, which are down-sampled in a higher level of the pyramid. It has disadvantages of iteratively image smoothing prevalent floating-point operations. In consideration of speed performance, we don't directly apply Gaussian derivatives to smoothed images, but instead apply rectangle filters at different sizes on original images and avoid floating-point operations. Thus the computation cost of scale space implementation is significantly reduced.

3.2. Fast Detectors

Laplacian detector in [3] locates geometric blob feature points both in spatial and scale space. $L_x x$ and $L_y y$ are the convolution of the second order Gaussian derivatives with the image I at point x. We denote $D_x x$ and $D_y y$ as approximated second order Gaussian derivatives and $D_x y$ as mixture partial Gaussian derivative. Then we extend rectangle filter to first order Gaussian derivatives denoted by D_x and D_y as shown in Fig. 2 and apply them in multi-scale geometric feature detection. With these approximated

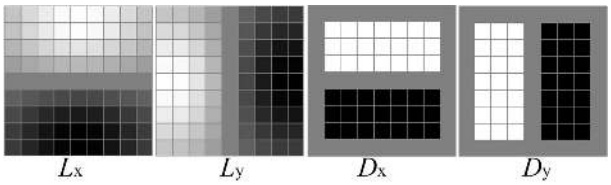

Figure 2. Extended box filters for first order Gaussian derivatives.

Gaussian derivatives, we construct a fast Laplacian detector for blob structures. The detector for blobs is

$$\nabla^2_{norm} D = t(D_{xx} + D_{yy}) \tag{4}$$

For the elongated ridge structures represented as ellipses, the fast ridge detectors turn into

$$\mathcal{R}_{norm} D = t^{3/2}\left((D_{xx} - D_{yy})^2 + 4D^2_{xy}\right) \tag{5}$$

The matrix that defines orientation and axis length of ellipses corresponding to ridges becomes

$$\Sigma = \int_{\eta \in R^2} \begin{pmatrix} D_x^2 & D_x D_y \\ D_x D_y & D_y^2 \end{pmatrix} g(\eta; t_{int}) d\eta \tag{6}$$

4. Experiments

In order to validate the proposed method, we implement the experiments on hand gesture data collected by ourselves and a public hand gesture image dataset.

4.1. Our Dataset

We collect some hand gesture images by a 1.3 mega pixels camera. Images in Fig.3 are some samples. To compare with Lars's method in [3], we give results by our method and Lars's method respectively. The detection results are superposed on sample images in Fig3. Ridge feature points are located at fingers. The ellipse on each finger is the union of elliptical descriptors of all ridge features. Blobs are marked as circles with blob features at the center of palm.

Our detectors get comparable results with Lars's method. The ridge and blob structures corresponding to palm and fingers are all detected with both methods. As depicted in Tab.1, It's obvious that our detectors obtain less feature points than Lars's method, which result from the approximation mechanism. Lars's method exactly calculates image convolution and finds local extrema in entire image area. While in our method, the filters have much bigger size than Gaussian kernels in Lars's method at the same scale. Thus the adjacent pixels are prone to have the similar derivatives, which results in less extrema. In practice, the features obtained by our method are enough to describe geometric structures of hand. On the other hand, since our method is not so sensitive as Lars's method, it's more robust to noise and artifacts than Lars's method. In the right two columns, no features are found at non-hand regions with our method, while Lars's method has false features.

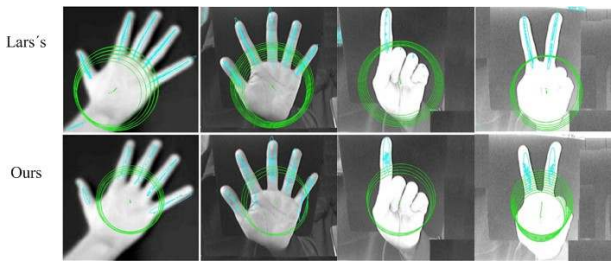

Figure 3. Blob and ridge features on sample images.

Tab.1 gives quantitative comparison on the image in the first column of Fig.3 between the two methods for blob and ridge structures, respectively. Because some little extrema are noise or non-hand features, we retain the same percentage of features for both methods. The percentages are empirically set as 10% for blob, 50% for ridge.

4.2. Standard Image Dataset

The standard database used in this paper is the Jochen Triesch database [9] which is a benchmark database in the field of hand gesture recognition. It consists of 10 hand

Table 1. Comparison between Lars's method and ours.

Detectors	Number of features (percentage)	Time(ms)
Lars's(blob)	8(10%)	590
Ours(blob)	6(10%)	9
Lars's(ridge)	590(50%)	860
Ours(ridge)	109(50%)	108

gestures in ASL (American Sign Language) performed by 24 different people against different backgrounds. To validate our method, we select the sequences of 4 hand gestures which denote 4 letters: a, d, g and v, respectively in Fig.4 because these 4 gestures are separable with geometric features representation. The backgrounds in selected images are of two types: uniform light and uniform dark. So we have totally 192 images in our experiments.

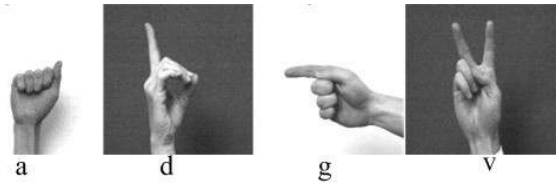

Figure 4. Gesture images for letters a, d, g and v.

Both our method and Lars's method in [3] are used in our experiments. The features and descriptors are shown in Fig.5. The blob corresponding to palm is marked as a green circle with its blob feature at the center, ridges corresponding to fingers represented as cyan ellipses with ridge features on the fingers. Similar to the results on our dataset, features obtained by our method are less than those by Lars's method, but are enough to capture the configuration of gestures. Blob feature in gesture image (d) by Lars's method is deviated from center of palm because of the dark annular area in palm area and many false ridge features found at palm. While our method gets blob and ridge features at proper position and scale. It's mentioned in section4.1 that our method is robust to noises or artifacts due to the mechanism of approximation. Experimental results here also testify its robustness. The recognition results are shown in Tab. 2. Our method achieves close accuracy for gestures (a), (g) and (v) to Lars's method. For gesture (d), our method exceeds the Lars's method with more than 20 percentages. The average accuracy is also comparable with the recent results obtained by Just A. using MCT features [10]. Our method has better speed performance compared with Lars's method. Time consumption of the two methods is also given in Tab.2. We average time consumption of two methods for each gesture sequence respectively. The improvement in speed is obvious. Notice that all the results of time consumption in this paper are measured with Matlab7.01 on a 2.8GHz PC with 512Mb RAM.

5. Conclusions

In this paper, we propose a fast multi-scale feature analysis method which can be used for hand gesture recognition. It is based on integral image approximation for Gaussian derivative in image convolution. With the approximation, fast multi-scale feature detectors are constructed to speed up computation. Experiments show that our method significantly decrease computation time while obtains comparable results both on our dataset and standard gesture image dataset.

6. 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 Nokia Research Center for the funding support.

References

[1] 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.

[2] 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.

[3] 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.

[4] P. Viola and M Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of Computer Vision and Pattern Recognition*. Hawaii, U.S., 2001, pp. 511–518.

[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *Proceedings of European Conference on Computer Vision*. Graz, Austria, Sept. 2006, pp. 404–417.

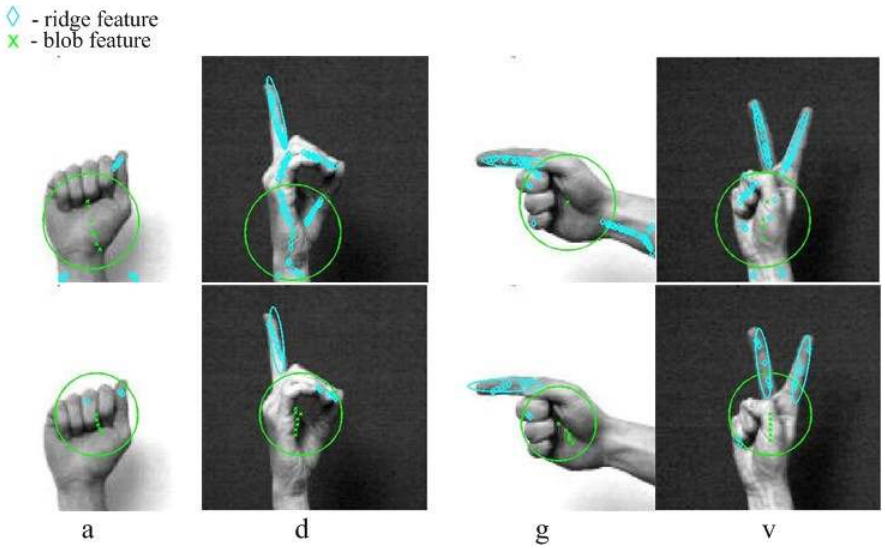

Figure 5. Comparison between Lars's and our methods.upper: Lars's method; bottom: our method.

Table 2. Results on Jochen Triesch database. (Each image is 128x128 pixels.)

Gesture	Method	Total	Correct	Accuracy	Average time (ms)
a	Ours	48	40	83.33%	52
	Larss	48	39	81.25%	232
d	Ours	48	41	85.41%	58
	Larss	48	30	62.50%	262
g	Ours	48	37	77.08%	57
	Larss	48	38	79.17%	251
v	Ours	48	43	89.58%	56
	Larss	48	46	95.83%	266
Total	Ours	192	161	83.85%	–
	Larss	192	153	79.69%	–

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

[7] T. Lindeberg, "Scale-space: A framework for handling image structures at multiple scales," *Technical Report CVAP-TN15,Royal Institute of Technology*, Sept. 1996.

[8] David G. Lowe, "Distinctive image features from scale-invariant keypointss," *International Journal of Computer Vision*, Feb. 2004.

[9] 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.

[10] 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, United Kindom, Apr. 2006, pp. 351–356.