Collaborated Directional Chamfer Matching for 3D Hand Pose Estimation

Wei Liu, Wei Fan, Yuan He, Jun Sun
Fujitsu Research & Development Center Co.,Ltd.
{willie, fanwei, heyuan,sunjun}@cn.fujitsu.com

Akihiro Minagawa, Yoshinobu Hotta
Fujitsu Laboratories Ltd., Japan
{minagawa.a, y.hotta}@jp.fujitsu.com

Abstract

We describe a shape-based chamfer matching method for recovering 3D hand pose using a monocular camera. Although some variants of chamfer matching methods incorporated edge direction information, of which the direction term is set empirically, so they may fail in different application scenario. In this paper, a collaborated directional chamfer matching (CDCM) is proposed. The direction term can be directly converted to a distance term, so a normalization factor is no more needed, and the similarity measurement can be directly calculated in scale space without adjusting the normalization factor. At the same time, a two layered structure is constructed using both the Active shape models (ASM) and the proposed CDCM to get the precise pose. Experimental results show the effectiveness of the proposed method.

1. Introduction

3D hand pose estimation is an important technology in the hand gesture recognition, and it can be widely applied in the human computer interaction. Many researches based on multiple visual cues such as color, texture and shape are proposed to recognize the hand pose. As a widely used cue, shape features exhibit large invariance to different lighting conditions and it get significant attention in recent decades.

Extensive literatures on shape matching exist, for example, Belongie et.al proposed a shape context method which represents the relationship between edge points using a histogram in [1], this approach can well handle the deformations and complex backgrounds. Opelt et.al [2] proposed a boosting based boundary fragments selection method to construct a boundary model detector. Ferrari et.al used several scale invariant descriptors through the k-connected near contour fragments to represent the shape [3]. Felzenszwalb et.al proposed a hierarchical representation which captures shape properties at different resolution levels, and used a dynamic programming algorithm in an elastic matching framework in [4].

Although the above mentioned methods can get effective performance in many applications, most of them suffer from the high computational cost. While the chamfer matching [5] is more efficient as it calculates the shape similarity just using the edge map, so it remains the preferred way for shape matching. Chamfer matching is not always reliable in complex scenes as the difference measure just by matching the chamfer distance between two images, so several approaches which incorporate the edge direction into the matching cost have been proposed. In [6], the images are represented by their edge maps, with a local direction associated with each edge pixel. In [7] and [8], the edge direction difference serves as an additional cost for shape matching. Although the introducing of the direction term can decrease the error matching, a normalization coefficient is set manually to make the direction values comparable with the location values, and it is not convenient to use in different scenes.

To solve the above problems, we propose a collaborated directional chamfer matching for 3D hand pose estimation. Through the proposed method, the direction difference term can be directly converted to a distance term, so a normalization factor is no more needed, and the similarity measurement can be directly calculated in the scale space.

In addition, we utilized the appearance information and used an ASM [9] algorithm to estimate the rough hand pose. For precise pose estimation, the proposed collaborated directional chamfer matching gives satisfied results.

2. Chamfer Matching

Chamfer matching [10] is a popular way to get the best matching between two point sets. It calculates the mean distance of the edges in template to its nearest edge in the query image.

$$D_C(T,E)=\frac{1}{|T|}\sum_{x_t\in T}\min_{x_e\in E}|x_t-x_e| \tag{1}$$

where $E=\{x_e\}$ is the edge maps of the query image, and $T=\{x_t\}$ is the edge points in the template image, $|T|$ is the total number of points in the template. The edge map can be calculated through the Distance Transform (DT) [10]. After the transform, each pixel in the new transformed image represents its distance to the nearest edge. The chamfer matching may suffer from noises in many applications, so we use a Partial Chamfer Matching [10] method which considers the matching result of several best matching points along the contour as the final distance.

3. Collaborated Directional Chamfer Matching

Gradient Calculation: In each feature window, the gradient vector is calculated for each pixel, the edge image is extracted by considering whether the gradient is beyond a certain threshold.

Chamfer Direction Map for the Whole Image: The edge image of the test image is shown in Figure 1a and the Distance Transform image of the test image in shown in Figure 1b. The red arrows represent the chamfer directions of the test image (*P1* is the closet edge point of *P2*, and the direction of vector $\overrightarrow{P1P2}$ is defined as the chamfer direction of *P2*).

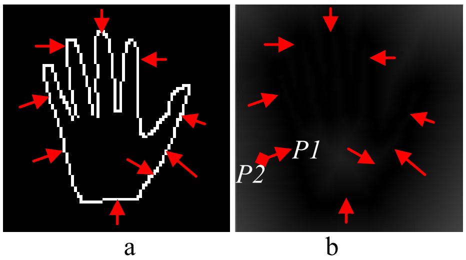


Figure1. Edge image of the test image and its DT image.

To get the chamfer direction map for the whole image, a Nearest Edge Point Searching algorithm was proposed to calculate the directions of all the pixels in the image. The algorithm is illustrated as follows:

Nearest Edge Point Searching Algorithm:

1, An input edge image (Figure 2a) is transformed to the distance map image using Distance Transform [10].

2, For each pixel in the distance map image, through searching for its 8 near neighbor pixels, a nearest pixel can be obtained as a local minimum to represent for the propagation trend to the edge (note that the values of the edge pixels are 0, and the value of each pixel represents its distance to the closest edge, so the direction along which the pixel value decreases represent the direction to the edge).

3, Based on the propagation trend, an iterative searching is used to obtain the global minimum of each pixel in the distance map, and finally, a corresponding closest edge point can be obtained.

4, Then the direction of each pixel can be calculated using the skew angle of the line connecting the current pixel and its closest edge point.

Figure 2b shows the directions of an example points on the edge image (Figure 2a). The white points denote the edge points, and black ones denote the example points, white lines denote their directions.

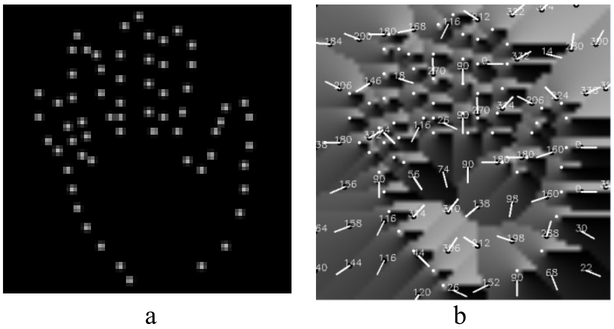


Figure 2. Chamfer direction for the whole image.

Collaborated Directional Chamfer Matching: Suppose the red dash curve in Figure 3 denotes the edge points in the template, and the green solid one denotes the edge points in the DT image of the test image. The collaborated directional chamfer distance can be calculated through the following steps:

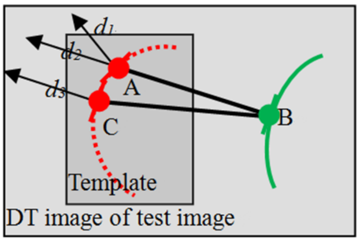


Figure 3. Collaborated directional chamfer matching.

(1) Point *A* has its direction d_1 in the template, and its direction in test image is d_2. Suppose Point *B* is the nearest point of *A* in the DT image of the test image, so the value of point *A* is |*AB*|.

(2) After d_2 was obtained, we can search in a near neighbor of *A* in the template to find the point *C* which has the most similar direction as d_2: d_3. Then |*BC*| is considered as the real distance between *A* and *B*. It is also defined as the Collaborated Directional Chamfer Distance.

(3) By accumulating all the points in template, we can get the total collaborated directional chamfer distance between the template and the test image.

$$D_{CDC}(T,E)=\frac{1}{|T|}\sum_{x_t\in T}|\mathop{\arg\min}_{x_e\in E}|x_t-x_e|-f(\mathop{\arg\min}_{x_e\in E}|x_t-x_e|)|$$

$$f(x)=\mathop{\arg\min}_{x_t\in T}|z(x_t)-z(x)| \tag{2}$$

where *z(x)* returns the direction of pixel *x* in test image, and *f(x)* returns the pixel with most similar direction as the pixel *x* in the template. Equation (2) calculates the |*BC*| in Figure 3.

Similarity Measurement in Scale Space: As the units of the chamfer distance and the edge direction are not comparable, so many methods fuse these two cues by setting the normalization factor between them manually. When the scale changes, the factor parameter need to be refined, this would induce great inconvenience to the application. However, based on the real distance (CDCM), manual setting of normalization factor is no more needed. Besides, as the linear characteristic of Euclid distance in the scale space, the similarity in different scale can be calculated directly as follows:

$$D^{S}_{CDC}(T,E) \;=\; \frac{1}{S}D_{CDC}(T,E) \tag{3}$$

We can use this similarity measurement to judge how well the images match.

4. 3D hand pose estimation

The flowchart of the 3D hand pose estimation algorithm is shown in Figure 5. In the offline training stage, with the samples of depth images and the corresponding color images obtained from a depth sensor, we can calculate the poses of the detected hands through manually marking several points on the hand plane. So the ground truth which contains the color image and its corresponding pose parameters are obtained.

Then a cluster algorithm such as k-means is utilized to cluster the hand samples according to the Euclid distance among their pose parameters. Until now, several categories of hand samples with similar poses are obtained.

At the same time, with the marked points on the color image, several hand models for each pose can be trained with the samples in the same category using ASM [9].

In the online stage, all the ASM models are matched with the test image to get the best matching model with the minimum difference. As each ASM model corresponds to a specific pose (a specific category of similar hand samples), the best matching ASM model can provide several candidate samples in its corresponding category. Then the proposed Collaborated Directional Chamfer Matching method can be used to find the best matching candidate and get the precise hand pose.

4.1 Rough hand pose category selection

In our online hand pose estimation system, the initial detection of the hand is performed through a hand classifier trained using the Viola's [11] object detection method.

Then all the trained ASM models are matched with the cropped hand image to select the best matching model. The candidates in the corresponding category of the best matching model are later used for precise matching. The flowchart is shown in Figure 4.

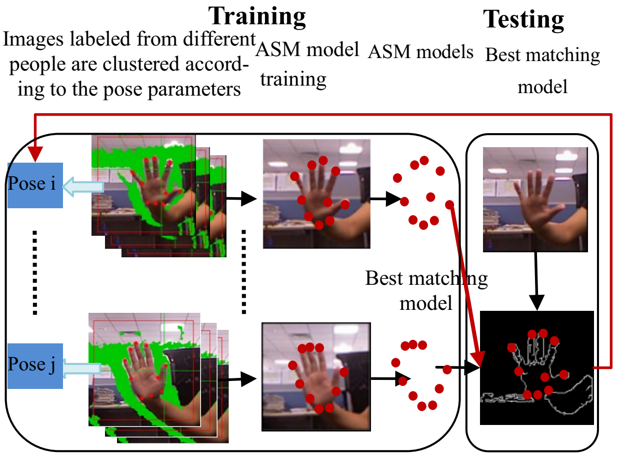


Figure 4. Low layer rough pose category estimation.

4.2 Precise hand pose estimation using CDCM

As long as the candidate template library is prepared in section 4.1, the precise estimation for the 3D hand pose can be performed. The flowchart of the precise hand pose estimation using CDCM is shown in Figure 5.

The candidate templates library which is a subset of the whole template library contains images with unique label of pose parameter as the ground truth. All templates are sliding on the test image to find the best location which has the maximum similarity score. Then all the scores are compared to find the best one and the template which corresponds to this score is considered as the best matching one.

The pose parameters of the best matching template are considered as the 3D poses of the hand. So the hand poses can be obtained.

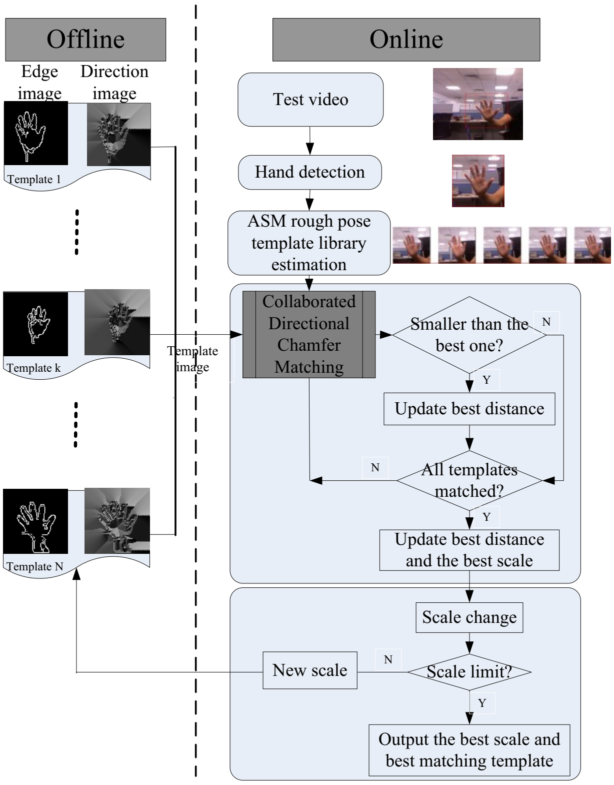

Offline
Online
Edge image
Direction image
Template 1
Template k
Template N
Template image
Test video
Hand detection
ASM rough pose template library estimation
Collaborated Directional Chamfer Matching
Smaller than the best one?
N
Y
Update best distance
All templates matched?
N
Y
Update best distance and the best scale
Scale change
Scale limit?
N
Y
New scale
Output the best scale and best matching template

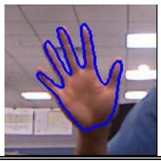
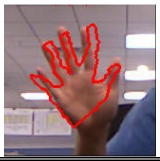
Figure 5. High layer precise pose estimation using Collaborated Directional Chamfer Matching.

5. Experiments

We collected videos from 10 persons who were asked to perform arbitrary hand movements. Through manually mark several salient edge points on the frames, hand samples including the color images and their corresponding 3D poses obtained as the ground truth with a depth sensor. There are totally 1157 images captured as the templates library, and 613 images are used as the test samples. The cropped template images have a resolution of 160*160.

Then an interpolation algorithm is used to generate dense edge points. In our experiments, the number of edge points is set as 191. With the edge points, the ASM model can be trained. Table 1 is the result of best matching model selection. Model 1 is trained with samples similar as the test image (shown in Table 1), while the model 2 is trained with samples which have large difference with the test image. The best model is selected with the minimum error, so model 1 is selected.

Table 1. Best ASM model selection.

Test	**Model 1**	Model 2
Matching error	8.57	26.03
Matching result		

When the best model is selected, the samples which are used to train the model are considered as the candidates for precise matching. The results of the precise matching are shown as follows.

With the sample dataset mentioned above, the proposed collaborated directional chamfer matching (CDCM) method is compared with the conventional partial chamfer matching (PCM) [10] and the directional chamfer matching (DCM) [8] methods. The estimation rates of the three methods are shown in Table 2. The matching process is executed from the base scale and multiplied with a factor of 1.2 in each loop, and the total number of scales searched is 5.

The computation complexity of the proposed method is $O(n^2)$, here n is the number of edge points in the template image, so speeding up the algorithm is still necessary. The average running time of the three methods on the bench of Intel(R) Core(TM) i5-2520M CPU @ 2.5GHz are also shown in Table 2.

Table 2. Pose estimation results of different methods.

Algorithm	PCM	DCM	CDCM
Estimation rate	92.9%	91.9%	94.0%
Running time (/ms)	97.8	358.5	494.2

From the Table 2, it can be seen that when the scale changes, a manually set normalization factor in the DCM will decrease the estimation rate.

Some pose estimation results of the propose method are shown in Figure 7, with different light conditions, sizes and persons. It shows that the proposed method is invariant to such situations.

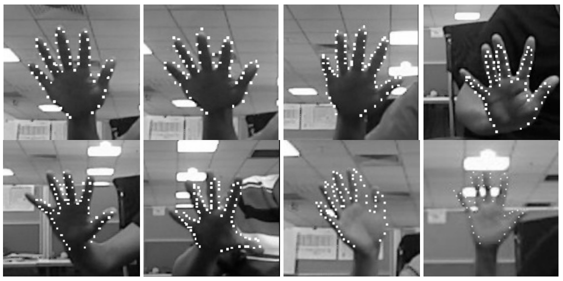


Figure 7. Some pose estimation results.

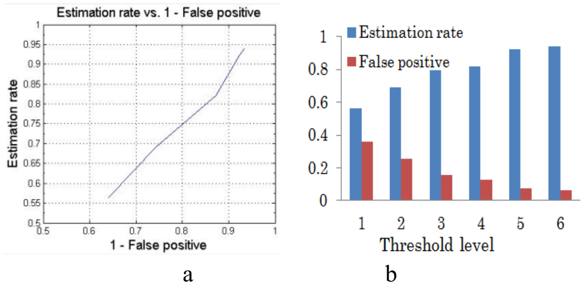


Figure 8. Relationship between the estimation rate and the false positive.

The Euclid distance between the best matching pose parameter and that of the ground truth is compared with a threshold to evaluate the efficiency of the match. We define the false positive as the matches with error unmatched 3D parameters but with good match in the 2D images. The estimation rate vs. false positive curve and their change with the threshold level are shown in Figure 8a and Figure 8b respectively.

6. Conclusions

A collaborated directional chamfer matching is proposed in this paper. The direction difference term can be directly converted to a distance term, so a normalization factor is no more needed, and the similarity measurement can be directly calculated in scale space. Besides, a two layered structure is constructed using both Active shape models and the proposed CDCM to get the precise pose. As there many complex gestures exist in real applications, more effective models which can represent complex gestures should be constructed in the next step.

References

[1] S. Belongie, J. Malik, and J. Puzicha: “Shape Matching and Object Recognition Using Shape Contexts”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.24, no.4, pp.509-522, 2002.

[2] A. Opelt, A. Pinz, and A. Zisserman: “A Boundary-Fragment-Model for Object Detection”, *In Proceedings of European Conference Computer Vision*, vol.2, pp. 575–588, 2006.

[3] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid: “Groups of adjacent contour segments for object detection”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no.1, pp.36-51, 2008.

[4] P. F. Felzenszwalb, and J. D. Schwartz: “Hierarchical Matching of Deformable Shapes”, *in IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2007.

[5] H. Barrow, J. Tenenbaum, R. Bolles, and H. Wolf: “Parametric correspondence and chamfer matching: Two new techniques for image matching”, *In International Joint Conference on Artificial Intelligence,* pp. 659–663, 1977.

[6] C. F. Olson and D. P. Huttenlocher: “Automatic Target Recognition by Matching Oriented Edge Pixels”, *IEEE Transactions on Image Processing*, vol.6, no.1, pp.103-113, 1997.

[7] M. Y. Liu, C. O. Tuzel, A. N. Veeraraghavan, and R. Chellappa: “Fast Directional Chamfer Matching”, *in IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1696-1703, 2010.

[8] J. Shotton, A. Blake, and R. Cipolla: “Multiscale categorical object recognition using contour fragments”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.30, no.7, pp. 1270-1281, 2008.

[9] T. F. Cootes, C. J. Taylor, D.H. Cooper and J. Graham: "Active shape models - their training and application", *Computer Vision and Image Understanding* vol.61, pp. 38–59, 1995.

[10] G. Borgefors: “Hierarchical chamfer matching: A parametric edge matching algorithm”, Pattern Analysis and Machine Intelligence, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.10, no.6, pp. 849-865, 1988.

[11] P. A. Viola and M. J. Jones: “Robust Real-Time Face Detection”, *International Journal of Computer Vision,* vol.57, no.2, pp.137-154, 2004.