Hand Posture Recognition from Disparity Cost Map

Hanjie Wang[1,2], Qi Wang[1], Xilin Chen[1]

[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. In this paper, we address the problem of hand posture recognition with a binocular camera. As bare hand has a few landmarks for matching, instead of using accurate matching between two views, we define a kind of mapping score–*Disparity Cost Map*. The disparity cost map serves as the final hand representation for recognition. As we use the disparity cost map, an explicit segmentation stage is not necessary. Local Binary Pattern (LBP) is used as feature for classification in this paper. In order to align the LBP feature, we further design an annular mask to deal with the problem of scaling, rotation, translation (RST) and search for an accurate bounding box of hand. The experimental results demonstrate the efficiency and robustness of our method. For 15 hand postures in varies cluttered background, the proposed method achieves an average recognition rate of 95% with a SVM classifier.

1 Introduction

Hand posture recognition attracts more and more researchers from both academy and industry as it has many applications, such as TV remote control, computer game navigation, sign language recognition [1], virtual navigation, etc. However, it is still a challenging problem due to the variability of hand shapes, orientations, etc.

Erol et al. provide a comprehensive review on hand posture recognition until 2005 [2]. Lockton et al. [3] propose a method providing a real-time performance by a combination of exemplar-based recognition and a deterministic boosting algorithm which can allow for fast online retraining. The correct rate is up to 99% for 46 postures. However, in Lockton's work, the backgrounds and shapes of these hand postures must be strictly the same. Similarly, the background in [4, 5] is just white wall. The hand can been perfectly segmented from the strongly constricted background and their feature can been easily extracted. Unfortunately, accurate hand segmentation is very challenging in real-world scenario. This limits the real application of many advanced technologies of hand posture recognition.

To deal with posture recognition in cluttered background, depth could be a useful cue for segmentation. Fujimura and Xia use Time of Flight (ToF) camera for gesture / sign language recognition [6, 7]. As TOF camera can only provide limited resolution, Van den Bergh and Van Gool [8] propose to combine RGB and

ToF cameras for 3D hand posture interaction with 6 hand postures. Recently, Kinect shows a significant advantage on capturing both color and depth images simultaneously with low cost. More and more researchers begin to explore action recognition including posture recognition with it. Oikonomidis et al. do a series of work on full DoF tracking of hand using Kinect [9–11]. However, such cameras can only be used in in-door environment and at a certain distance.

Compared with depth sensor, binocular camera simply captures a scene from different view angles. The disparity from two views provides an equivalent of depth. As the resolution can be very high, and the baseline of the binocular camera can be adjusted to fit the distance, the binocular camera can be used for more flexible scenarios. When a pair of rectified stereo images is prepared, disparity map, i.e., the depth cue, can be obtained by searching the correspondence of the pixels in the right image to the pixels in the left image on each epipolar lines. Among the state of the art, there is stereo matching algorithm that can provide enough accurate disparity map where hand segmentation is easy by simple threshold method. However, there is still a dilemma to be tradeoff between accuracy and computational cost. As one of the best results, Wang et al. [12] present a stereo matching algorithm based on inter-regional cooperative optimization. A processing with 4 iterations costs about 20s for one image pair. Felzenszwalb et al. [13] propose a method that substantially improves the running time of the belief propagation approach. However, it still takes approximately 1s to obtain the results. This is still far from the requirement of real-time processing.

The requirement of real-time processing makes some researchers incline to a low quality disparity map. Many of them think that, even with a low quality disparity map, accurate hand segmentation could still be possible by combining appearance cues [14, 15]. However, the perceived color of skin varies significantly depending on the user's skin color and the light condition of background. What's more, uncontrolled background may give rise to numerous false alarms. So, the number of hand posture class is limited. For example, Hadfield and Bowden [14] extract features in both the appearance and depth channel for 3 static postures recognition. Similarly, Grzeszczuk et al. [15] conduct their experiment on only 6 kinds of hand postures using the feature combining depth and skin color. In addition, introducing of skin color increases the chance of these methods being sensitive to lighting variance and confusing background.

To achieve real-time and more efficient hand posture recognition, we need other binocular visual cue rather than accurate disparity map or the combination of low quality disparity map and appearance cues. We found that the disparity cost map of hand, which is derived by matching the hand region in right view with that in the left view according to a series of disparities, can provide significant information about hand posture. Based on this observation, we propose to recognize hand posture directly from disparity cost map. The related work is Deselaers et al.'s [16], where they use disparity cost slice, similar to our disparity cost map, to find the correct depth for the hand. Different from their work, we utilize disparity cost map directly for hand posture recognition.

The rest of this paper is organized as follows: Section 2 presents the proposed method and the experimental results in section 3 demonstrate the efficiency of the proposed method. Section 4 concludes this paper and gives the future directions.

2 Proposed Method

We propose a method to recognize hand postures directly from disparity cost map without requiring accurate segmentation of hand. Besides, our method takes scaling, rotation and translation variance into consideration. Experimental results show that our method is robust and fast.

2.1 The Basic Idea

In our work, we denote the right view as I_r and the left view as I_l. The binocular camera is horizontally aligned. Therefore only horizontal displacement is relevant, and we can set $m(x, y, d)$ as a measurement of similarity between pixel (x, y) in the right view and corresponding pixel $(x + d, y)$ in the left view. We use a simpler Birchfield-Tomasi sub-pixel metric from [17] as the dissimilarity measurement.

$$m(x, y, d) = \min_{x-\frac{1}{2} \le x_l \le x+\frac{1}{2}} |I_r(x, y) - I_l(x_l + d, y)|, \tag{1}$$

$m(x, y, d)$ servers as the value of points in disparity cost map. Let R be a rectangle region and d be a specific disparity in the right view. We define

$$M(R, d) = \{m(x, y, d) | (x, y) \in R\}, \tag{2}$$

as a 2D patch in the disparity cost map. If both hand posture region R_h and its disparity d_h are obtained, the patch $M(R_h, d_h)$ is thus determined. Fig. 1 shows the construction of disparity cost map M and patch $M(R_h, d_h)$ of hand posture.

Then, let's explain the reason why $M(R_h, d_h)$ can contribute to hand posture recognition. Different from accurate segmentation, $M(R_h, d_h)$, which is only a patch of disparity map, can not indicate one pixel belongs to hand or background. What $M(R_h, d_h)$ is able to indicate is the potential of pixels satisfying the disparity d_h. Although so, it is enough for the hand posture to emerge from $M(R_h, d_h)$ since most of the hand pixels satisfy the disparity d_h while most of background pixels do not, as seen from Fig. 1.

There must be ambiguous pixels in $M(R_h, d_h)$, especially in the case of the hand in the background with similar appearance in both color and texture. However,statistical prior can help to clarify this ambiguous. In the sense of statistics, ambiguous pixels would become a kind of white noise due to the uncertainty compare with hand. What's more, due to the different angles of light, positions of shade and other imaging factors, the non-texture regions between fingers or even at the edge of fingers can not match well between two views. It is also

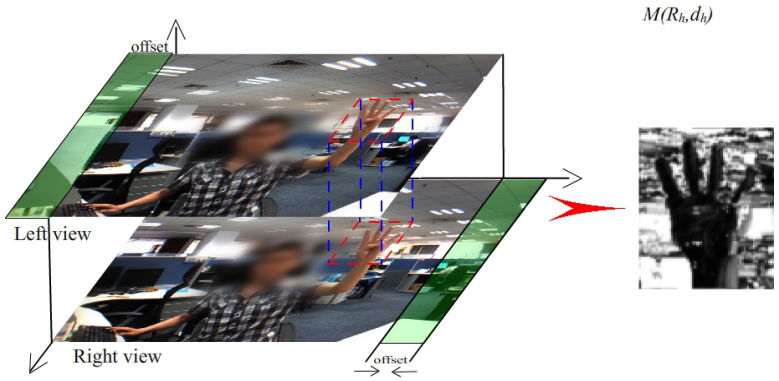


Fig. 1. The basic idea of hand posture from disparity cost map. Gray levels in $M(R_h, d_h)$ indicate the value of matching costs.

notable that the hand posture would emerge more clearly with cluttered background since the background behind the hand would have high matching cost in $M(R_h, d_h)$. It means that, different from other methods which require clean background, $M(R_h, d_h)$ would be able to contribute more clear information under cluttered background. In general case, cluttered background is one of the main obstacles to accurate hand segmentation.

In a word, $M(R_h, d_h)$ provides more information than non-segmented map but less than accurate segmented map. Based on the above reasons, we propose a framework for hand posture recognition. The main steps are shown in Fig. 2 and a brief explanation is as follows.

Step 1: Disparity map is calculated with fast block matching.

Step 2: Similar to the method proposed by Song et al. [18], we locate the hand by human face detection combined with a generative model of the human upper body. Different from [18], to raise the accuracy, the detection step also have an assumption that hand is usually the foremost part in HCI scene. Finally, we obtain the region R_h and the average disparity value d_h of hand.

Step 3: $M(R_h, d_h)$ is then determined through Eq. (2).

Step 4: To extract LBP feature, a refinement of R_h and d_h must be conducted. Then, the label of the hand posture is obtained by feature extraction and classification

This paper emphasize constructing $M(R_h, d_h)$ and extracting LBP feature from $M(R_h, d_h)$ for the purpose of hand posture recognition, i.e., step 3 (section 2.1) and step 4 (section 2.2).

2.2 Recognize Hand Posture Using LBP Feature

After obtaining $M(R_h, d_h)$ at step 3 in the proposed method, a more accurate bounding box of hand is derived with the following searching method in this section. The scaling, rotation and translation of hand are taken into account at

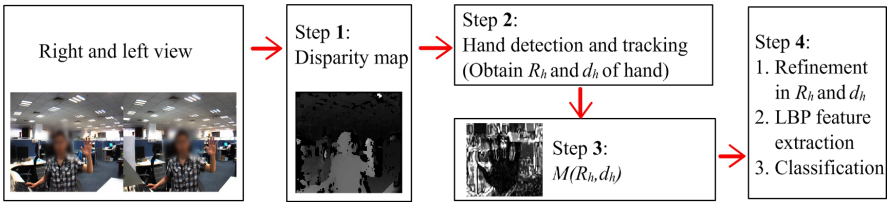


Fig. 2. The procedure of proposed method for hand posture recognition.

the same time. Following that, a LBP feature is extracted from the bounding box.

2.2.1 Bounding Box for LBP Feature Extraction

We discover that $M(R_h, d_h)$ is not an ideal bounding box for extracting LBP feature since the size of R_h is too large for hand and a single d_h is not robust. So, a refinement of d_h and R_h is necessary.

Refinement of d_h. By considering that it is not common for all pixels belong to hand posture to satisfy disparity d_h well, we replace the single $M(R_h, d_h)$ with the triple of $(M(R_h, d_{h-1}),\ M(R_h, d_h),\ M(R_h, d_{h+1}))$. The utilization of this triple would increase the robustness and efficiency for extracting feature. In fact, we simply combine the triple through Eq. (3) to produce a new $\bar{M}(R_h, d_h)$ for successive feature extraction.

$$\bar{M}(R_h, d_h) = \min\{M(R_h, d_{h-1}), M(R_h, d_h), M(R_h, d_{h+1})\} \tag{3}$$

Refinement of R_h. LBP feature is widely used in human face recognition and requires faces to be aligned. However, hands are not aligned in $\bar{M}(R_h, d_h)$ and R_h is not the perfect bounding box of hand. In general case, in problems of face recognition, the eyes, nose and mouth always appear in the same configuration, with similar proportions and contrasts. However, robust hand posture recognition is a more difficult problem due to the high number of degrees of freedom. In this paper, we propose to extract a special annular region from hand. The annular region is to hand what the eye or nose is to human face. To be specific, we design a rectangle with a annular mask in it to search for a more accurate bounding box of hand in $\bar{M}(R_h, d_h)$, considering the scaling, rotation and translation at the same time.

In order to achieve the best discriminative ability, the annular region changes its radius r_{min} and r_{max} automatically and so does the corresponding bounding box (see Fig. 3). We represent the pixel in annular region as $p(r\cos\theta, r\sin\theta)$ in polar coordinates. At this step, we only consider the pixels on the annulus between r_{min} and r_{max}. We define $H_{mean}(r)$ as the mean of pixels' values along the radius and $V_{mean}(\theta)$ as the mean of pixels' values along the circle. H_{var} and V_{var} are their corresponding variances.

$$H_{mean}(r) = \left[\sum_{\theta=0}^{2\pi} p(r\cos\theta, r\sin\theta)\right] \Big/ \left[\tfrac{\theta_{\max}-\theta_{\min}}{\Delta\theta} + 1\right] \; r_{\min} \le r \le r_{\max}\,, \quad (4)$$

$$V_{mean}(\theta) = \left[\sum_{r=r_{\min}}^{r_{\max}} p(r\cos\theta, r\sin\theta)\right] \Big/ \left[\tfrac{r_{\max}-r_{\min}}{\Delta r} + 1\right] \; 0 \le \theta \le 2\pi\,, \quad (5)$$

$$H_{\text{var}} = \sum_{r=r_{\min}}^{r_{\max}} \sum_{\theta=0}^{2\pi} [p(r\cos\theta, r\sin\theta) - H_{mean}(r)]^2, \quad (6)$$

$$V_{\text{var}} = \sum_{\theta=0}^{2\pi} \sum_{r=r_{\min}}^{r_{\max}} [p(r\cos\theta, r\sin\theta) - V_{mean}(\theta)]^2, \quad (7)$$

In experiment, we set $r_{\max} = r_{\min} \times 1.5$, radius step length $\Delta r = 1$ and radian step length $\Delta\theta = 0.01$. Intuitively, on one hand, H_{var} has its local maximum in one range of r_{min} (in Fig. 3, the range is [5, 15]), where the annulus covers the edge of palm. Keep expanding the annular region, H_{var} decreases to a local minimum, which indicates that the annular region no longer covers the palm. On the other hand, V_{var} should have kept increasing when the annulus expands. However, it decreases after $r_{min} = 15$ in Fig. 3. That is because the number of ambiguous pixels increases when the annulus expands. Experimentally, we design a D as the criteria to search the best r_{min} as follows:

$$D = \max_{r_{\min}} \frac{\lambda V_{\text{var}}}{H_{\text{var}}}, \quad (8)$$

where $\lambda = 0.9$ in our experiment. The solid (red) curve in Fig. 3 (a) shows the values of D with its corresponding 21 r_{min} candidates. Experimental results also demonstrate the robustness of this method.

Dealing with the problem of RST. The changing radius of annulus decides the size of the bounding box and further solves the problem of scaling. When refer to variable translation, disparity map has already provided a coarse location of hand. However, our annular region needs a center of palm and we have to solve the problem as follows. In the training step, the accurate center of palm is labeled manually. In the testing step, the center of the annular region moves in a small range and chooses the best one. The best one is chosen as follows. We just simply move a window in the bounding box to search the region with maximum density. The center of the window is determined as the center of annulus. In order to take rotation of hand into consideration, we obtain the principal direction using the ultimate optimal annular region, which theoretically and experimentally covers the wrist or parts of lower arm nearby. The principal direction is determined by the position of wrist. A simple clustering in the annular region locates the wrist (see Fig. 3 (d)). Following that, the annulus is kept still and the bounding box is rotated according to the principal direction. Attention, the rotation angle in our work is no larger than 45 degree.

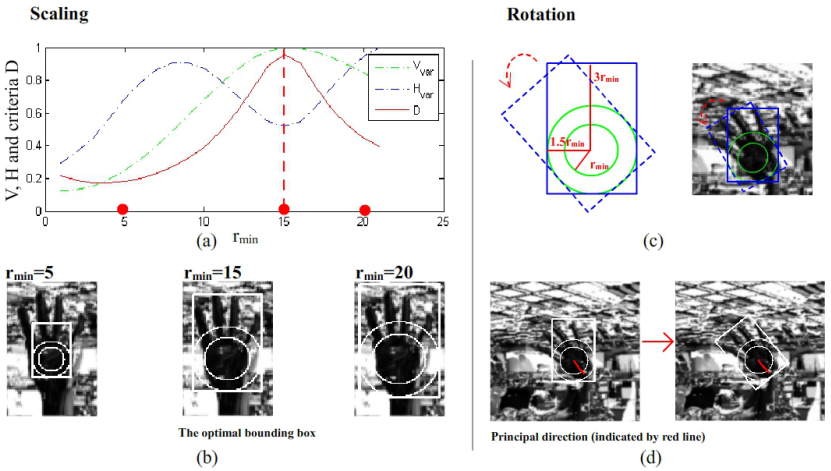


Fig. 3. (a) Red curve in diagram shows the values of D with its corresponding 21 r_{min} candidates. (b) Samples with $r_{min} = 5, 15, 20$ around the palm to search for a bounding box of hand in $\bar{M}(R_h, d_h)$. (c) Rotation method. (d) Principal direction for rotation.

2.2.2 LBP Feature and Classification

For feature extraction, LBP texture feature is selected. It is originally proposed for gray-scale and rotation invariant texture classification [19]. In our work, it is performed in the refined bounding box of hand. We divide the bounding box to cells. For each pixel in a cell, compare the pixel to each of its 8 neighbors along a circle, i.e., clockwise. The label of each pixel is thus determined by its neighbors as follows:

$$P_c^{LBP} = \sum_{i=0}^{7} \begin{cases} 2^i & P_i > P_c \\ 0 & P_i \leq P_c \end{cases}, \tag{9}$$

where P_i are the neighbors of P_c. For each hand posture image, we divide it into 49 cells. In each cell, a 10 bin normalized histogram is developed. The final feature vector concatenates normalized histograms of LBP in all cells and its total dimension is 490.

In our work, we use SVM for classification. The standard SVM targets to maximum the boundary of two classes. In this paper, given a set of training examples, a multi-class SVM model is trained. The model is the classifier we need to predict new examples to a category based on which side of the gap they fall on. Experimental result demonstrates that the correct recognition rate increases when the number of training data becomes larger.

3 Experiments

This section reports results of our proposed method in static hand posture recognition.

Dataset. The test vocabulary includes 15 classes of hand postures. We collect hand posture samples by sampling all the 15 postures against 10 different backgrounds under 2 different lighting conditions. Each posture is sampled 5 times at each situation. So, we obtain altogether $1500 = 15 \times 10 \times 2 \times 5$ hand posture samples, in other words, each posture has 100 samples. In order to evaluate our method against variable illumination and background, two datasets were developed for each task. In dataset I, 50 samples from almost same lighting condition are set as training data and 50 samples in different lighting condition are set as testing data for each hand posture. In dataset II, 50 samples from 5 kinds of background are set as training data and the rest 50 samples from other 5 kinds of background are set as testing data. The size of input image is 512×384 for each view. Fig. 4 shows all the 15 postures and also the typical $\bar{M}(R_h, d_h)$ related to each posture.

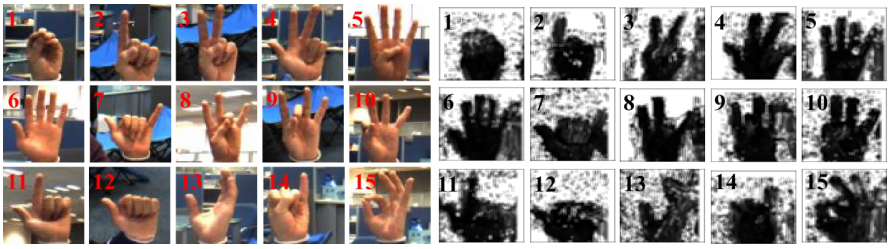
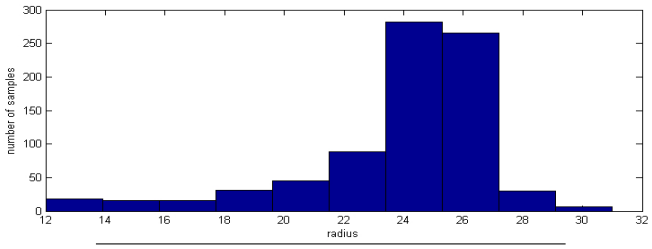


Fig. 4. 15 hand postures in cluttered lab environment and $\bar{M}(R_h, d_h)$ of each corresponding hand posture in our experiment. They have already been aligned and scaled according to the annular masks.

Optimal radius of annular region. By expanding the annular region from the center of palm, we obtain an optimal radius and further obtain the optimal bounding box of hand. For the purpose of comparison, an additive dataset of ground truth is labeled manually. They are scaled to equal size. The optimal radiuses are nearly 25, which are also estimated manually. The purpose of this experiment is to find the optimal radius of hand in the additive dataset automatically for comparing with the ground truth. The result demonstrates the robustness and efficiency of our method proposed in section 2.2. The histogram of the optimal radiuses from 750 samples is shown in Fig. 5.

The histogram shows that most samples obtain their optimal radiuses at nearly 25, which is accord with the ground truth. Some "errors" are mainly occurred in hand postures like No. 1 and No. 12, which have no long fingers spread and thus obtain the optimal radius at nearly 14. However, such small



Process	Dataset I	Dataset II
Before refinement (%)	87.73	81.35
After refinement (%)	91.07	86.00

Fig. 5. The histogram of the optimal radius from 750 samples and the results comparison.

bounding box is big enough to include all the pixels of such hand postures like No. 1 or No. 12 and still performs correct recognition.

Result with variable illumination and background. To evaluate our method against variable illumination, the classifier I is trained from training data in dataset I by SVM. The testing data of dataset I is then predicted by classifier I. The result confusion matrix is shown in Table 1. The average recognition rate is 91.07%.

To evaluate our method against variable background, we did the same experiment in dataset II. The result confusion matrix is shown in Table 1 and shows that the average recognition rate is 86.00%, which is comparable to the performance of classifier I. See from Table 1, hand posture No.11 and No.12 have the lowest correct recognition rate. Because, No. 11 is very similar to No. 2, No. 12 is similar to No.7. Confusion pixels, which may be mis-recognized as fingers, have large probability to lead to error recognition in those situations. We will address this issue in our future work.

Table 1. The confusion matrix of result from dataset I (Left) and dataset II (Right). Numbers in **bold** are hand posture index.

No.	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
01	0.9	0.08													0.02
02		0.88			0.04		0.06				0.02				
03		0.02	0.88	0.06	0.02		0.02								
04			0.04	0.94							0.02				
05			0.08		0.62	0.04		0.08	0.08	0.06	0.02				0.02
06						0.92		0.04							0.04
07	0.02	0.02					0.84					0.04			0.08
08		0.02	0.04					0.94							
09			0.04						0.94						0.02
10										1					
11											1				
12						0.06						0.94			
13			0.02	0.02			0.02						0.88	0.02	0.04
14														1	
15														0.02	0.98

No.	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
01	0.86	0.04					0.06							0.04	
02		0.96	0.02					0.02							
03			0.94	0.04									0.02		
04				0.98		0.02									
05			0.02		0.88	0.04		0.06							
06				0.02	0.02	0.96									
07		0.02			0.04	0.02	0.86	0.02				0.04			
08			0.08		0.14			0.78							
09					0.06				0.94						
10					0.02					0.98					
11		0.24			0.04		0.08		0.04		0.6				
12							0.44					0.56			
13													0.96		0.04
14					0.02									0.94	0.04
15		0.02		0.02		0.14				0.1				0.02	0.7

Training data in dataset. Performance is also measured using SVM in other datasets with random split of training data and testing data. It is notable that the correct recognition rate increases when the number of training data becomes larger (see Fig. 6). This is because in the sense of statistics, confusing pixels would become white noise due to the uncertainty of their positions. The correct recognition rate of 96.33% is comparable to some results using ToF camera or Kinect [8, 6, 20]. What's more, our method requires much less constraints.

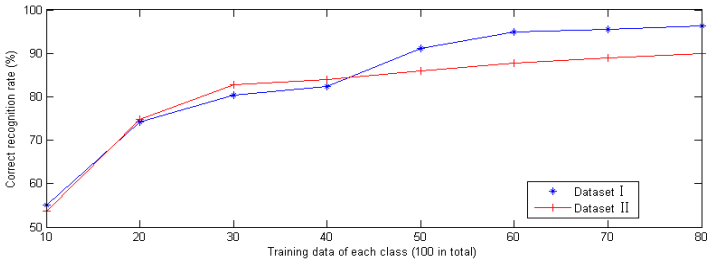


Fig. 6. The correct recognition rate with the number of training data

Feature comparison. To the best of our knowledge, there is no feature directly extracted from thus 3D matching cost map. In order to choose the best feature fit for the disparity map, we mainly compare the performance of LBP with Gray Level, Gradient and HOG. For each hand posture image, we divide it into 49 cells. Features of LBP, Gray Level, Gradient and HOG are individually concatenated to normalized histograms in all cells and their final feature vector are all 490. It is notable that Gaussian filter is also used to resist the confusion pixels in $\bar{M}(R_h, d_h)$. We do this experiment both in dataset I and dataset II by SVM and the results are shown in table 2.

Table 2. The correct rate of different features and time cost for feature extraction

Features	**Gray**	**HOG**	**LBP**	**Gradient**
Dimension	490	490	**490**	490
Variable illumination (%)	70.19	83.87	**91.07**	73.11
Variable background (%)	84.40	88.67	**86.00**	66.67
Time cost (s/per frame)	0.2052	0.2120	**0.2078**	0.2090

From table 2, we may see that LBP feature is the best feature in dataset I. This is because LBP label is calculated depending on its neighbors, which makes LBP feature robust to the variable illumination. However, HOG feature has a slightly better performance than LBP feature in dataset II. But HOG

feature is sensitive to confusion pixels in the $\bar{M}(R_h, d_h)$, which is preprocessed by Gaussian filter. In experiment, we discovered that a small change of the filter vastly decreases HOG feature's performance while LBP feature does not. A trivial example is simply using the raw $\bar{M}(R_h, d_h)$ without preprocessing. The correct recognition rate of HOG feature is 67.07% while the rate of LBP feature is 81.33%.

Method comparison. As there is no dataset can be used for comparison in this field, the only way for comparison is to re-implement theirs' methods and test on our datasets.

In [16], in order to segment foreground from background, learning approach is employed. The background is manually segmented for training. They even train backgrounds under different lighting conditions and camera positions. They also uses fusing feature, including appearance, shape and depth. Same hand postures with different depths are regarded as different classes in the work of [16]. However, depth is not considered in the task of hand posture recognition in this paper. Besides, see from Fig. 4, it is not wise to extract shape information of hand postures. To this end, appearance is the only cue we can use to re-implement their methods on our datasets.

It is more convenient to compare our method with [14] on our datasets since both of us use binocular cameras in office environment. The difference is that we have 15 classes of hand postures while they have only 3 classes. See Table 3 for details.

Table 3. The correct recognition rates of 3 methods

Class Number	**15 classes**	**8 classes**	**3 classes**
Our method	0.93	0.93	0.99
Hadfield et al.[14]	0.84	0.88	0.96
Deselaers et al.[16]	0.88	0.91	0.96

4 Conclusion and Future Work

This paper presents a method to recognize hand posture with binocular camera from disparity cost map. The proposed method avoids segmenting hand accurately, and it works well even in cluttered background. Our experimental results demonstrate that the proposed method is efficiency and robustness. Besides, annular region is applied to deal with hand changing in scaling, rotation, and translation. This helps to align the LBP feature and improve the performance.

Acknowledgement. This work is partially supported by National Basic Research Program of China (973 Program) under contract 2009CB320902; and

Natural Science Foundation of China (NSFC) under contract Nos. 61025010, 60832004, 61001193 and 60973067; and the FiDiPro program of Tekes.

References

1. Cooper, H., Holt, B., Bowden, R.: Sign Language Recognition. In: Visual Analysis of Humans 4. Springer (2011) 539–562
2. Erol, A., Bebis, G., Nicolescu, M., Boyle, R.D., Twombly, X.: Vision-based hand pose estimation: A review. CVIU **108** (2007) 52–73
3. Lockton, R., Fitzgibbon, A.: Real-time gesture recognition using deterministic boosting. In: BMVC. 2. (2002) 817–826
4. Ho, M.F., Tseng, C.Y., Lien, C.C., Huang, C.L.: A multi-view vision-based hand motion capturing system. PR **44** (2011) 443–453
5. Tomasi, C., Petrov, S., Sastry, A.: 3d tracking = classification plus interpolation. In: ICCV. (2003) 1441–1448
6. Fujimura, K., Xia, L.: Sign recognition using depth image streams. In: FG. (2006) 381–386
7. Xia, L., Fujimura, K.: Hand gesture recognition using depth data. In: FG. (2004) 529–534
8. Van den Bergh, M., Van Gool, L.: Combining rgb and tof cameras for real-time 3d hand gesture interaction. In: WACV. (2011) 66–72
9. Oikonomidis, I., Kyriazis, N., Argyros, A.: Efficient model-based 3d tracking of hand articulations using kinect. BMVC. 2. (2011)
10. Oikonomidis, I., Kyriazis, N., Argyros, A.: Tracking the articulated motion of two strongly interacting hands. In: CVPR. (2012) 1862–1869
11. Oikonomidis, I., Kyriazis, N., Argyros, A.: Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In: ICCV. (2011) 2088–2095
12. Wang, Z., Zheng, Z.: A region based stereo matching algorithm using cooperative optimization. In: CVPR. (2008)
13. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient belief propagation for early vision. IJCV **70** (2006) 41–54
14. Hadfield, S., Bowden, R.: Generalised pose estimation using depth. In: ECCVW on Sign Gesture & Activity. (2010)
15. Grzeszcuk, R., Bradski, G., Chu, M., Bouguet, J.: Stereo based gesture recognition invariant to 3d pose and lighting. In: CVPR. 1. (2000) 826–833
16. Deselaers, T., Criminisi, A., Winn, J., Agarwal, A.: Incorporating on-demand stereo for real time recognition. In: CVPR. (2007)
17. Birchfield, S., Tomasi, C.: Depth discontinuities by pixel-to-pixel stereo. IJCV **35** (1999) 269–293
18. Song, Y., Demirdjian, D., Davis, R.: Tracking body and hands for gesture recognition: Natops aircraft handling signals database. In: FG. (2011) 500–506
19. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. PAMI **24** (2002) 971–987
20. Pugeault, N., Bowden, R.: Spelling it out: Real-time asl fingerspelling recognition. In: ICCVW on Consumer Depth Cameras for Computer Vision. (2011) 1114–1119