2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops

A Novel Human Detection Approach Based on Depth Map via Kinect

Yujie Shen, Zhonghua Hao, Pengfei Wang, Shiwei Ma
School of Mechatronics Engineering & Automation,
Shanghai University
No.149,Yanchang Rd.200072 Shanghai, China
masw@shu.edu.cn

Wanquan Liu
Department of Computing, Curtin University
WA, 6102 Australia
W.Liu@curtin.edu.au

Abstract

In this paper, a new method of human detection based on depth map from 3D sensor Kinect is proposed. First, the pixel filtering and context filtering are employed to roughly repair defects on the depth map due to information inaccuracy captured by Kinect. Second, a dataset consisting of depth maps with various indoor human poses is constructed as benchmark. Finally, by introducing Kirsch mask and three-value codes to Local Binary Pattern, a novel Local Ternary Direction Pattern (LTDP) feature descriptor is extracted and is used for human detection with SVM as classifier. The performance for the proposed approach is evaluated and compared with other five existing feature descriptors using the same SVM classifier. Experiment results manifest the effectiveness of the proposed approach.

1. Introduction

There has been a growing effort in the development of intelligent video surveillance systems based on automatic detection and understanding of human activity in video images. Different methods have been brought forward to detect human in still image and have achieved high accuracy. These methods can be roughly divided into three different categories; human model based methods [1], template matching based methods [2] and statistical classification methods [3-5]. The approaches in the last category, in which the choice of suitable feature descriptors is critical to the design of a detector, have demonstrated promising results since they are more robust than the other two categories of methods.

Over past few years, several feature descriptors for visible light two dimensional images have been proposed, such as HOG [3], PHOG [6], et al [7-9]. Among them, the LBP (local binary pattern) feature, which is a string of bits obtained by binarizing local neighborhood of pixels with respect to the brightness of central pixel, was recently proposed to capture microscopic local image texture and was applied for human detection successfully [7]. Wang [8] put forward the HOG-LBP human detector with partial occlusion handling, where HOG as a shape feature is complemented with LBP as a texture feature. Other variants from LBP, such as the LTP (local ternary patterns) [9] and CENTRIST (census transform of histograms) [4] also have been attracted many attentions. Although lots of works have reported that these features could be used to obtain accurate results in human detection, they encountered many difficulties in perceiving the shapes of human objects with articulated poses and cluttered background [10,11].

Since depth map represents an object's space information which is an important cue for human to recognize objects, while visible map includes color and illumination mainly. The application of depth map has attracted much more research interests in recent years. For example, Plagemann [12] used local shape features to identify body parts in depth map for human detection. Ikemura [10] proposed a window-based human detection method by using relational depth similarity feature based on depth information. Lu [11] proposed a method of human detection approach based on depth map by using a two-stage model containing a 2-D head contour model and a 3-D head surface mode.

However, since most of the existing depth map sensors, such as TOF camera and binocular cameras, are expensive and lack of friendly application interface, human detection on depth map is rarely applied in practice. Recently, the Microsoft's Kinect provides an easy way to capture real time depth map due to its low cost, simple operation and friendly application programming interface. It has been used in many applications, such as face recognition, pose estimation [1,2,19] etc. Its application on object detection should be an attractive research topic. Unfortunately, since the Kinect mainly depends on speckle method [13], the depth map captured by Kinect often contains much noise. Furthermore, detectors trained from the existing image feature descriptors that are demonstrated very successful on visible images cannot achieve promising results on depth map of Kinect due to its unstable quality and inherent defect. Since there's no significant progress to overcome the defects of depth map currently, human detection based on such depth map is still a challenging task. In fact, effective preliminary filter process and a

978-0-7695-4990-3/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPRW.2013.85

suitable feature descriptor are quite necessary for the designing of human detector based on such depth map.

In this paper we aim to build an integrated human classifier which would perform well based on depth map collected by Kinect. For this purpose, two filtering methods are adopted to process the depth map and a normalized dataset composed of depth maps with various indoor human poses is constructed. A novel feature named LTDP (local ternary direction pattern) with strong noise resistance derived from LBP feature is proposed. Its various performances are evaluated and compared with some existing features based on the created dataset. A SVM classifying algorithm is utilized to generate the integrated human classifier. Furthermore, related theoretical analysis, experimental work and discussion, as well as future works are presented.

The remaining of this paper is organized as follows. Section 2 briefly describes LBP feature and its improved algorithm. The proposed LTDP feature and human detection approach on depth are illustrated in detail in Section 3. In Section 4, experiments are presented and results are analyzed. Some concluding remarks and suggestions for future work are given in Section 5.

2. Overview of LBP-related features

Initially derived from texture analysis, the LBP feature is created as a gray-level texture measure to model texture images [14]. Later, it showed excellent performance in many other fields in terms of speed and discrimination capability [7]. Mathematically, it marks each pixel I_c of an image as a decimal number $LBP_{P,R}(I_c)$, which is formed by comparing the P equally spaced neighboring pixels $I_{p,R}(p = 0, \cdots, P-1)$ on a circle of radius R with the center pixel I_c and concatenating the results binomially with factor 2^p, as

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(I_{p,R} - I_c)2^p \tag{1}$$

where the threshold function is defined as

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{2}$$

When is (0,0), the coordinates of can be given by . The values of gray level of $I_{p,R}$ that do not fall exactly in the center of pixels can be estimated by interpolation. By defining the number of spatial transitions (0/1 changes) in LBP pattern with a U value defined as below in (3), the uniformity of LBP patterns, which refers to the patterns having limited transition or discontinuities (U≤2) in the circular binary presentation, can be determined, where the U value is given as

$$U(LBP_{P,R}) = \left| s(I_{p-1} - I_c) - s(I_0 - I_c) \right| + \sum_{p=1}^{P-1} \left| s(I_p - I_c) - s(I_{p-1} - I_c) \right| \tag{3}$$

The uniform LBP only has 59 bins, one each for 58 possible uniform patterns and one for all of the non-uniform ones.

Later, the LTP [9] feature was proposed to extend LBP to 3-valued codes, in which the gray levels within a zone of width ±t around I_c are quantized to zero and those above this zone are quantized to +1 and below to -1, i.e., the indicator is replaced with a 3-valued function s′ as

$$s'(u, i_c, t) = \begin{cases} 1, & u \geq i_c + t \\ 0, & |u - i_c| < t \\ -1, & u \leq i_c - t \end{cases} \tag{4}$$

Hence, the binary LBP code can be replaced by a ternary LTP code. Usually, LTP uses a coding scheme that splits each ternary pattern into its positive and negative halves. Although the introduction of such a user-specified threshold in LTP breaks the monotonic illumination invariance of LBP feature, it helps suppress the noise that dominates LBP responses in near-uniform regions and provides an additional parameter that can be tuned to extract complementary information.

Recently, the CENTRIST [4] was proposed. Originally, census transform is a non-parametric local transform designed for establishing correspondence between local patches and equivalent (modulo a difference in bit ordering) to the LBP code. Therefore, the values of CENTRIST for an image or image patch can be easily computed. It adopts a spatial pyramid structure by dividing an image into sub-regions and integrating the corresponding results within these regions. The spatial pyramid encodes rough global structure of an image and usually improves the quality of recognition.

In summary, the uniform-LBP reduces the dimensions of LBP, while LTP extends LBP to three-valued codes and therefore enhances its anti-noise performance. The CENTRIST introduces a pyramid structure to LBP and makes a multi-scale observation. Since all of the LBP variants create a model of the image from the comparison between individual pixels, they consequently lack the capability of anti-noise especially for some particular noises.

3. Proposed Human Detection Method

The procedure of the proposed method for human detection based on depth map has four steps. Firstly, the depth map is processed with two filtering methods, i.e. pixel filtering and context filtering. Secondly, a normalized dataset with various indoor human poses is established. Thirdly, the proposed LTDP feature is employed to encode the depth map into feature vectors. Lastly, classification algorithm is utilized to generate the human classifier. The details are given as follows.

3.1. Noise Reduction Filters to Depth Map

Anti-noise approaches based on mean filtering or Gaussian techniques by using appropriate noise models for TOF depth map have been widely studied [3,5,8,12,14]. Compared with TOF data, the depth map captured by Kinect has mounts of null-value areas, which present as 'white holes' in depth map (Fig.2(b)). Moreover, it is obvious that these 'white holes' always exist at boundaries of objects where depth changes sharply. The purpose of filtering is to give the truth depth values for the pixels in that null-value area. However, traditional filters, e.i., mean filtering, Gaussian filtering, usually are utilized to remove salt and pepper noises. They have poor performance in the case discussed in this paper. Fig.2(c) shows the Gaussian filtering result. It is obvious that the 'white holes' cannot be filled up properly.

In this paper, two filters, pixel filter and context filter, are adopted to process the noisy depth map. The pixel filter is designed to compensate the 'holes', while the context filter is employed to further reduce noise in general.

The procedure of pixel filtering is given as below

$$N = \sum_{p=1}^{L^2} f(I_p), \quad f(I) = \begin{cases} 0 & \text{otherwise} \\ 1 & I = 0 \end{cases} \tag{5}$$

$$V = \frac{\sum_{p=1}^{L^2} I_p}{N}, \quad \text{as } I_p > 0, \ N > T \tag{6}$$

Firstly, find a zero pixel I and create a filter window with the candidate pixel as its center. Then, count the non-zero pixels in filter window with size of the L^2 pixels. When the counted number exceeds a user-defined threshold T, the value of candidates will change to the mean value V of non-zero pixels; otherwise left it unchanged. An example of 5×5 pixels filer window with threshold T=3 is shown in Fig.1. The number of non-zero pixels in this window is set to be 7, which exceeds the threshold. Thus, the filtering candidate pixel is reset to 45 that is the mean value of the 7 non-zero pixels.

Since the Kinect depth maps are unstable even for the same scene, several frames could be fused together to overcome this defect. This can be done by using context filter. The procedure of context filtering is given as follows. Firstly, find a zero pixel in the depth map and filter it using the pixel filer as described above. Then, retrieve the value on the same location of previous frames. The latest non-zero pixel will be adopted to filter the candidate. If all of the previous values are zero, the candidate will not change. It should be noted that the frames waiting to be retrieved are limited, because the method ignores the scene change between frames. In the experiment of this paper, we fused four frames for waiting to be retrieved.

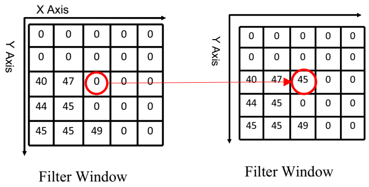


Figure 1: An example of 5×5 pixels filer window with threshold 3

These two methods can either be used to a depth map separately or in series to produce a smoothed result. Fig.2 gives an experiment result by using various filtering methods. In this example, the number of zero pixels in raw depth map, pixel filtering, context filtering, and combined filtering are 1605, 622, 1241 and 498, respectively. We evaluated the performance of the combined filter for an indoor environment. Statistical results by 2000 filtering samples show that this filter can reduce in average 60% of zero pixels. Although the solutions cannot completely remove all noises, they do achieve an appreciable result. It is clear that the depth map after smoothing is more appropriate for extracting accurate feature descriptors.

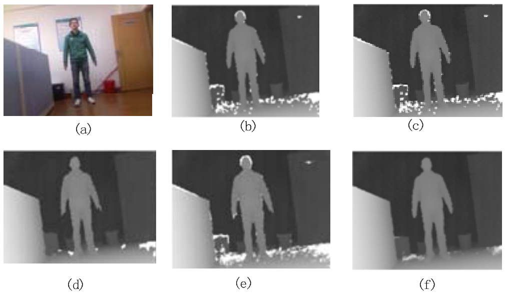


Figure 2: An example of filtering result to a depth map: (a)visible map of the scene; (b)raw depth map; (c)smoothed with Gaussian filter; (d)smoothed with pixel filter;(e) smoothed with context filter; (f)smoothed with two filters combined

3.2. The LTDP Feature

Although there exist various techniques to describe local image regions by image features, searching for an ideal feature descriptor having desired properties is still a tough task with few theoretical guidelines. Alternatively, the problem can be solved by combining multiple complementary features based on different aspects e.g. HOG-LBP. However, the combination of features usually causes a mass dimension of features. Hence, in the proposed method, only one single type of features is adopted and a fast and efficient feature descriptor named LTDP (local ternary direction pattern) is derived from the LBP-related feature descriptors by plugging a specific Kirsch mask [15] and three-valued codes into it. This process is described as follows.

At first, the LTDP is calculated by comparing the relative edge response value of a pixel in different directions. The edge response value (S0~S7) of a particular pixel is calculated with the Kirsch mask at eight different directions. The masks (M0~M7) are shown in Fig.3.

$$\underset{\text{East } (M_0)}{\begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix}} \underset{\text{North East } (M_1)}{\begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix}} \underset{\text{North } (M_2)}{\begin{bmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix}} \underset{\text{North Wast } (M_3)}{\begin{bmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix}}$$

$$\underset{\text{West } (M_4)}{\begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix}} \underset{\text{South West } (M_5)}{\begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix}} \underset{\text{South } (M_6)}{\begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix}} \underset{\text{South East } (M_7)}{\begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix}}$$

Figure 3: Kirsch mask used in LTDP

Second, the 3-valued codes for the eight directions with threshold t are defined as follows.

$$s_i(I, M_i, t) = \begin{cases} 1, & I \bullet M_i \geq t \\ 0, & \left| I \bullet M_i \right| < t, \qquad i = 0 \cdots 7 \\ -1, & I \bullet M_i \leq -t \end{cases} \tag{7}$$

where, denotes the neighboring pixels around the center pixel. A sample of LTDP code is shown in Fig.4. When the user-defined threshold t is set too large, there will be many zeros in the 8-bit LTDP codes; alternatively there is no zero when t is set too small. Both cases will reduce discriminative capability. Therefore, the parameter must be set to an appropriate value to distribute 1, 0, -1 homogeneously. In the experimental study of this paper, it is set as 12.

In the experiment of this paper, a uniform pattern argument is designed and a coding scheme is used to split each ternary pattern into its positive and negative halves and subsequently treating them as two separate channels of LTDP features for which separate histograms are computed by combining the results only at the end of the computation. As a result, after these improvements, the LTDP possesses stronger anti-noise capability. For instance, as shown in Fig.4, as the pixel 53 turns to 48 due to noise, the LBP code is changed while LTDP code did not change.

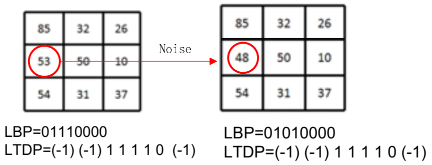


Figure 4: A sample of LTDP code shows stronger noise tolerance than LBP

3.3. Histogram of LTDP Feature for Depth Map

Since the histogram of LTDP features contains the distribution information of local features in an image, in order to preserve spatial information, a depth map should be divided into several non-overlapping rectangular blocks. By letting denote the histogram of LTDP patterns extracted from block $R_i\,|_{i=1,\cdots,N}$, the spatial histogram of an image can be represented as . A spatial histogram, concatenating the histograms of all blocks can be employed to represent the whole image. The spatial histogram encodes both the appearance and the spatial relationships of an image. For instance, given a 64×128 detection window, an image can be divided into 32 blocks. In each block, a 118-dimensional LTDP feature vector with 59-dimension for negative and the other for positive can be extracted. Consequently, a 3776-dimensional feature vector can be extracted from the whole detection window.

In the application of various image features, some parameters, such as the number and size of block, block overlapping, dimension and pyramid space should be considered carefully. Block is the smallest image patch to calculate the feature vector. Some features will be extracted from overlap of blocks, alternatively one can calculate feature vector from pyramid space (i.e., to shrink image in various size). Some features make a more detailed description, but will increase their dimensions at the same time. To illustrate this, with a given 64×128 pixels image, all the aforementioned features are implemented, and the parameters were list in table 1, in which we try to keep the block in similar size as it is a key factor for implementation.

Feature	HOG	LBP	PHOG	LTP	CENTRIST	LTDP
Block size	16×16	16×16	8×16	16×16	16×32	16×16
Overlapping	YES	NO	NO	NO	YES	NO
Spatial pyramid	NO	NO	YES	NO	YES	NO
Block number	105	32	21	32	31	32
Description dimension	3780	1888	1260	3776	3968	3776

Table 1: Parameters of features for a 64×128 image

3.4. Construction of Dataset on Human Poses

There are some public available benchmarks of visible map dataset for human detection, such as INRIA dataset [3] and Caltech Pedestrian Dataset [7]. However, they do not have a depth map dataset. The only one we found up to now is the Shenzhen University Depth Map in [18]. However, this simple dataset only contains a limited range of scale, occlusion and pose variation of human body, which is too small to assess real performance of different features.

In this study, we constructed a depth map dataset in our laboratory so as to provide a fair benchmark to compare the performance of different features on depth map. In this dataset, nearly 3×104 frames were collected in various indoor scenes such as meeting room, living room, kitchen, and office, by using a Kinect with 11-bit and 640x480 resolution depth map video generated at 30Hz. In order to make a comparable quality with still image, the device is moved slowly. About 40% of the frames had no pedestrians, and about 60% of the frames contained one or two persons.

All frames containing human body are annotated with totally 2×104 labeled bounding boxes and are split into training part and testing part. Fig.5 gives some samples in this dataset, where Fig.5 (a) are positive samples of walking, occlusion, strolling and running in sequence from left to right, and Fig.5 (b) are negative samples of drinking fountains, air-conditioner and chair.

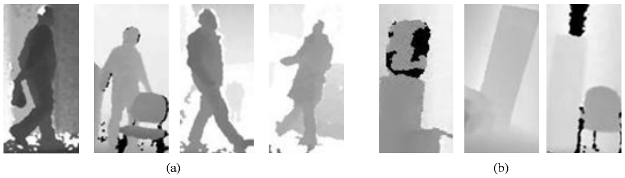


Figure 5: Samples in the constructed depth map dataset

3.5. Classification Algorithm

For binary classifications, there are many classification methods such as the nearest neighbor, neural networks, decision tree, and support vector machine (SVM). Among these algorithms, SVM is relatively robust and easy to be implemented. In this study, both linear kernel SVM and nonlinear kernel SVM are used as classification algorithm. Commonly, the polynomial kernel, radial basis function kernel and sigmoid kernel are used as nonlinear kernels in SVM. In each nonlinear kernel, the optimal value of gamma parameter can be estimated by using the method introduced in [16], where the inverse of the mean value obtained from a distance matrix of the feature vectors is adopted to reduce computational cost by further doing cross validation. In the experiments of this paper, the feature vectors of the training depth map dataset are sent to SVM for generating a classification model.

4. Experiments and Discussions

In the experiments, the performance of the proposed LTDP feature is evaluated and compared with five existing features, i.e. LBP, LTP, CENTRIST, HOG and PHOG, on the constructed depth map dataset. And the Performance of LTDP based SVM classifiers are investigated.

4.1. Feature Evaluation

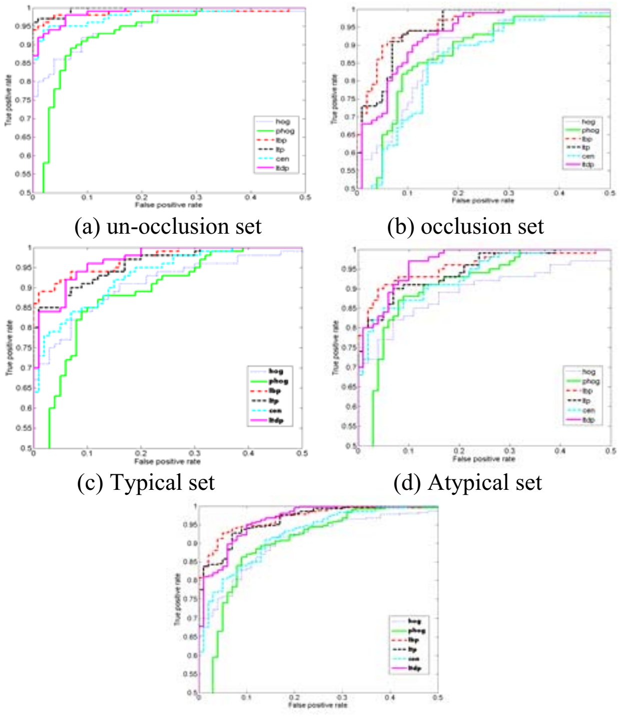
Considering that the ROC (receiver operating characteristic) curve concerns the true positive rate and the false one, we adopt it to demonstrate the detailed performance of detectors using different features in the experiments. With this criterion, by tuning the threshold, the corresponding ROC space points and the value of AUC (area under the ROC curve) can be readily obtained according to the classification results. Typically, the values of AUC range from 0.5 to 1.0, the larger it is the better the performance is.

During the training procedure, each feature is tested on five sub-datasets, i.e. reasonable set, un-occlusion set, occlusion set, typical set and atypical set. The five sub-datasets are selected from dataset we constructed and are classified by whether the human body is occluded from the view point of camera, i.e. reasonable set, un-occlusion set, occlusion set, typical set and atypical set. The occlusion set was collected by considering a human body with over 15% areas being occluded, while the un-occlusion set contained the others. The typical set and the atypical set contained samples with abnormal view point or human pose divided by different viewing points and poses. While, the reasonable set was collected from dataset randomly. Each subset contains 100 positive samples and 100 negative samples.

The ROC curves obtained from experiments are given in Fig.6. All of them were the result of corresponding classier trained by a linear SVM. It should be noted that although the searching for hard examples in the negative dataset is critical in training, the number of retraining iterations used in experiments is also important. In training procedure, only two bootstrapping rounds are retrained since more rounds of retraining may lead to exaggerated memory requirements for SVMs [17]. The occlusion will degrade the performances for all classes of features tremendously as it can be observed that the ROC curves of each feature are declined in Fig.6 (b). In addition, it can be observed that the changes of human pose and viewing point also slightly decrease the performance.

More detailed analysis of the ROC curve is listed in

Table.2, where the AUC of all features together with linear SVM are given. Because the data of reasonable subset are selected from the four subsets randomly, this subset has more generality in classification task. Therefore, in the table, all features are ranked in the last column in term of the AUC value obtained from reasonable set. It can be observed that the proposed LTDP feature performed better than others. The results also indicate that, even though the LTP extends LBP to 3-valued codes, the improvement is limited, while the LTDP with a Kirsch mask significantly improves the performance of classification.



(a) un-occlusion set

(b) occlusion set

(c) Typical set

(d) Atypical set

(e) Reasonable set

Figure 6: ROC curves obtained from five sub-datasets

	Atypical	Typical	Occlusion	Un-occl usion	Reasonable	Rank
LBP	0.9756	0.9836	0.9649	0.9974	0.98038	3
LTP	0.9725	0.9785	0.9744	0.9979	0.98083	2
HOG	0.9437	0.9525	0.9364	0.9731	0.95143	5
PHOG	0.9438	0.9364	0.9251	0.9604	0.94255	6
CENRIST	0.9664	0.9645	0.9215	0.9894	0.96045	4
LTDP	0.9859	0.9872	0.9680	0.9926	0.98342	1

Table 2 AUC value of ROC curves

4.2. Performance of LTDP with SVM Classifiers

In order to validate the effect of the proposed approach, the performances of LTDP based on SVM classifiers with different kernels are evaluated and compared in this section. The results are given in Fig.7. It shows that linear kernel SVM slightly outperform the nonlinear kernels SVM, which implied that the depth map reduced nonlinear effect on classification. It is not a simple task to simply declare one kernel is better than another in complicated hydrological simulation. Theoretically linear kernel may be considered as a private case of non-linear ones. However, when optimal decision is the border of linear decision, linear kernel will outperform other kernels. Some results of pedestrian detection on depth map by using LTDP based linear kernel SVM are shown in Fig.8, it manifests the effectiveness of the proposed method.

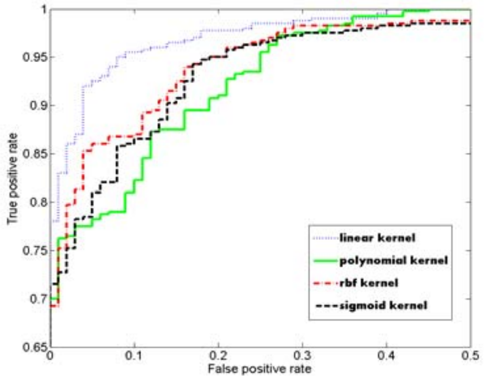


Figure 7: Performances of LTDP based SVM classifiers with different kernels

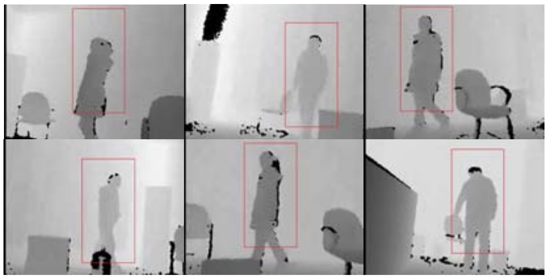


Figure 8: Samples of pedestrian detection results

5. Conclusion and Future works

In this paper, we proposed a new method of human detection using SVM algorithm via extracting the new designed LTDP feature descriptor only based on depth map collected by Kinect sensor. Two effective filtering methods, pixel filtering and context filtering, were employed to smooth the depth map at first. A normalized dataset composed of depth maps with various indoor human poses was constructed as benchmark for evaluating the new feature. The LTDP feature was then derived by simply introducing a Kirsch mask and three-valued codes into the LBP and LTP feature. Experiments results on the collected dataset showed that, not only the LTDP feature outperformed other five existing feature descriptors which were commonly used for visible images, but also the nonlinear effect of SVM classification task were reduced by using the LTDP on depth map. Since the occlusion was challenging for all features including the LTDP feature, future work will focus on the LTDP based human detection on depth map including occlusion case in outdoor environment. And further investigation will put more attention on comparison task.

6. Acknowledgments

The authors should acknowledge the supports to this project from the China National Science Foundation No.61171145, Shanghai Educational Research Foundation No.12ZZ083 and Shanghai University Graduate Students Innovation Foundation.

References

[1] Broggi A, Fascioli A, Grisleri P, Graf T, and Meinecke M. Model-based validation approaches and matching techniques for automotive vision based pedestrian detection. Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA.pp.1-8,2005.

[2] Leibe B, and Seemann E Schiele B. Pedestrian detection in crowded scenes. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. San Diego, CA: pp. 878-885, 2005.

[3] Dalal, N.; and Triggs, B. Histograms of oriented gradients for human detection. Computer Vision and Pattern Recognition, pp.886-893,2005.

[4] Jianxin Wu, and Rehg, J. M. CENTRIST: A Visual Descriptor for Scene Categorization. Journal of Pattern Analysis and Machine Intelligence, 33(8):1489-1501, 2011.

[5] Bo Wu, and Ram Nevatia, Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. Journal of Computer Vision, 75(2): 247-266, 2007.

[6] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. Proceedings of the International Conference on Image and Video Retrieval, pp. 401-408,2007.

[7] Dollar P., Wojek, C, Schiele, and B, Perona, P. Pedestrian detection: A benchmark. Conference of Computer Vision and Pattern Recognition, pp.304-311, 2009.

[8] Wang, Xiaoyu, Han, Tony X., and Yan, Shuicheng. An HOG-LBP human detector with partial occlusion handling. Conference of Computer Vision, pp.32-39, 2009.

[9] Xiaoyang Tan, and Triggs, B. Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions. Journal of Image Processing, 19(6):1635-1650, 2010.

[10] S. Ikemura, and H. Fujiyoshi. Real-Time Human Detection using Relational Depth Similarity Features, ACCV. pp. 25-38, 2010.

[11] Lu Xia, Chia-Chih Chen, and Aggarwal J.K. Human detection using depth information by Kinect. Computer Vision and Pattern Recognition Workshops, pp.15-22,2011.

[12] Plagemann C, Ganapathi V, Koller D, and Thrun S. Real-time identification and localization of body parts from depth images. Conference of Robotics and Automation, pp.3108-3113, 2010.

[13] D. Tao, X. Li, X. Wu, and S. Maybank. Human carrying status in visual surveillance. Conference of Computer Vision and Pattern Recognition, pp. 1670–1677, 2006.

[14] Ojala T, Pietikainen M, and Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Journal of Pattern Analysis and Machine Intelligence, 24(7):971-987, 2002.

[15] M. Sonka, V. Hlavac, and R. Boyle. Image Processing, Analysis, and Machine Vision, 2nd ed. , Brooks/Cole, Pacific Grove, CA, 1999.

[16] Chih-chung Chang, and Chih-Jen Lin. LIBSVM: A library for support vector machines. Journal of ACM Transactions on Intelligent Systems and Technology, 2(3)27:1--27:27, 2011.

[17] Felzenszwalb, P., McAllester, and D., Ramanan D. A discriminatively trained, multiscale, deformable part model. Conference of Computer Vision and Pattern Recognition, pp.1-8, 23-28, 2008.

[18] Shengyin Wu, Shiqi Yu, and Wensheng Chen. An attempt to pedestrian detection in depth images. Conference of Intelligent Visual Surveillance, pp.97-100, 2011.

[19] Demetriou Michael K., Kounalakis Tsampikos, Vidakis Nikolaos, and Triantafyllidis Georgios A.. Fast 3D scene object detection and real size estimation using Microsoft Kinect sensor. Conference of Computer Graphics and Imaging, pp:254-260, 2012.