

SLTP: A Fast Descriptor for People Detection in Depth Images

Shiqi Yu, Shengyin Wu

School of Computer Science and Software Engineering,
Shenzhen University, Shenzhen 518060, China

{shiqi.yu@, wushengyin@email.}szu.edu.cn

Liang Wang

National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, China

wangliang@nlpr.ia.ac.cn

Abstract

This paper presents a new feature descriptor for real-time people detection in depth images. The shape cue in depth images can reduce negative impacts of variations of clothing, lighting conditions and the complexity of backgrounds. The proposed Simplified Local Ternary Patterns (SLTP) can take advantage of depth images to describe human body shape with low computational cost. To evaluate the SLTP feature, we establish a dataset with 7260 positive samples. A series of experiments are carried out on this dataset, and the results show that the SLTP feature can achieve a high detection rate with a low false positive rate. Besides, SLTP is easy to implement, and performs fast (over 80 frames per second) on a standard desktop computer.

1. Introduction

In some time-critical applications such as robots, games and driver assistance, real-time people detection is of great importance. People detection in images has achieved a great progress in recent years [5, 7], but it remains challenging due to bad illumination, complex background and human body pose variation.

There has been some work on people detection in the recent literature. In [4], Dalal and Triggs proposed a method that takes Histograms of Oriented Gradients (HOG) as the feature and the SVM as the classifier. Their method gives encouraging results. It is a milestone contribution to people detection, and many newer methods are inspired by HOG. Zhu *et al.* [15] adopted AdaBoost and cascades to accelerate HOG feature selection (and thus people detection). To solve the problem of partial occlusion, Wang *et al.* [13] combined HOG and LBP (Local Binary Pattern) [3] fea-

tures, and achieved good performance. Enzweiler *et al.* [7] presented a survey on people detection, in which the experimental results have shown that when compared with wavelet-based AdaBoost cascade and Neural Network/LRF, the HOG/LinSVM combination performs better at higher resolutions. Dollár *et al.* [5] introduced the state of the art people detection system, and their results and analysis indicated that HOG remains competitive.

Even though great progress has been achieved in people detection, due to bad illumination, varying clothing texture, complex background, different human postures and occlusion, detecting human bodies in images is still a challenging task. In the recent years, a new kind of cameras, named depth cameras (or range cameras), have been developed. In particular, Kinect [1] released by Microsoft measures distance based on stereo triangulation with the help of infra-red patterns [8].

Using the depth images, a human body can be detected by a motion detection method combined with image segmentation. But this kind of method can not distinguish human bodies with other moving objects. They also can not work when the sensor is moving, such as the sensor is carried by a moving robot. So we proposed a method which can detect people in only one static image.

With the shape cue in depth images, the negative impact of bad illumination and complex background on people detection would be greatly reduced. Some previous work [14, 10] has proved this. The Histograms of Depth Differences (HDD) feature in [14] and the Histograms of Oriented Depths (HOD) feature in [10] are very similar with HOG. So they have equivalent computational cost as HOG, and the computational cost of HOG is relatively high. In [14], the proposed method was designed based on depth images from a TOF camera [2]. The images contain more noise than those captured from a Kinect depth camera.

In this paper, a new kind of feature is accordingly proposed for people detection in high-precision depth images collected from Kinect, leading to a higher detection rate and lower computational cost. The contribution of this paper is twofold: (1) a new feature descriptor, SLTP, for real-time people detection in depth images, (2) and a large evaluation dataset for people detection which contains depth images and the corresponding intensity ones.

The remaining part of this paper is organized as follows. Our proposed SLTP feature descriptor is presented in Section 2. Then in Section 3 we introduce our evaluation dataset. The experimental results and analysis are given in Section 4. The last section concludes the paper.

2. Feature Extraction

People detection is based on the sliding window method. In this paper, a type of feature is effectively designed to describe human body in depth images. Such feature is named as Simplified Local Ternary Patterns (SLTP), which can describe the local depth variance.

SLTP is inspired by Local Ternary Patterns (LTP) [11] and Histogram of Oriented Gradients (HOG) [4]. LTP is introduced by Tan and Triggs for face recognition. It is a 3-valued coding. For a pixel g_c at position (x, y) , it is coded by the differences between its neighbors $g_i (i = 1, 2, \dots, 8)$ and itself,

$$t_i(x, y) = \begin{cases} 1, & g_i \geq g_c + T_g \\ 0, & |g_i - g_c| < T_g \\ -1, & g_i \leq g_c - T_g \end{cases} \quad (i = 1, 2, \dots, 8) \quad (1)$$

where T_g is a user-specified threshold. For 8-neighbors coding, there are $3^8 = 6561$ patterns. LTP can be treated as two separate channels of LBP descriptors [3], and only uniform patterns are considered. So the number of patterns are reduced to 118 [11].

The proposed SLTP is coded only by 4-connected pixels. For a reference pixel at (x, y) , we first compute its depth differences Δ_x and Δ_y in the x - and y -direction respectively. And then we quantize the depth differences into 0, +1 and -1. In detail, SLTP is defined as follows:

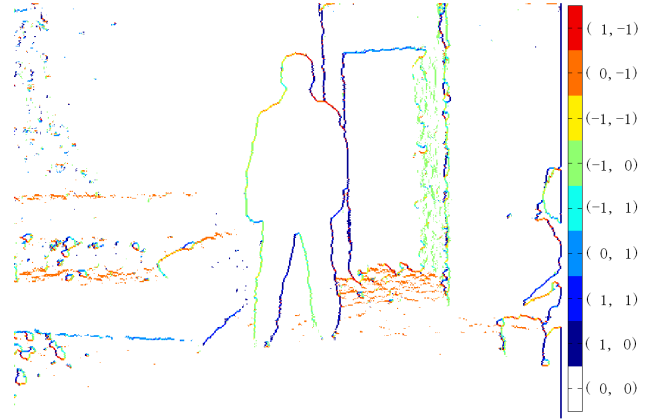
$$SLTP(x, y) = (t_x, t_y) \quad (2)$$

$$t_x = \begin{cases} 1, & \Delta_x \geq T_d \\ 0, & |\Delta_x| < T_d \\ -1, & \Delta_x \leq -T_d \end{cases} \quad (3)$$

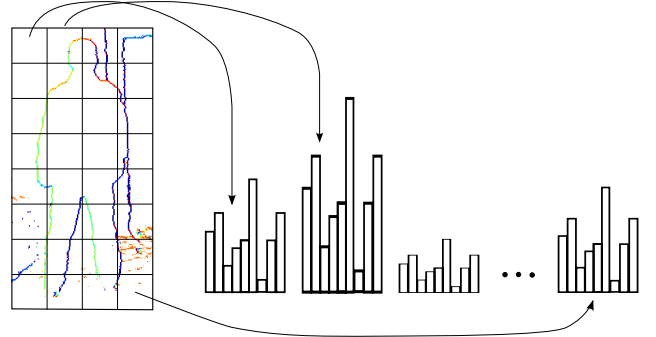
$$t_y = \begin{cases} 1, & \Delta_y \geq T_d \\ 0, & |\Delta_y| < T_d \\ -1, & \Delta_y \leq -T_d \end{cases} \quad (4)$$



(a) A depth image. The darker pixels are closer to the camera than the lighter pixels.



(b) The SLTP extracted from the depth image. Different colors indicate different patterns.



(c) The SLTP histogram extracted from a detection window.

Figure 1. The SLTP description.

where $\Delta_x = d(x+1, y) - d(x-1, y)$, $\Delta_y = d(x, y+1) - d(x, y-1)$, $d(x, y)$ is the depth value at position (x, y) , and T_d is a user-specified threshold.

There are $3 \times 3 = 9$ patterns to describe the local depth differences, and each pattern describes one kind of depth difference, such as pattern (1,-1) describing the depth difference whose direction is top-right in an image coordination

system, and (0, 0) describing pixels in a flat area. In a depth image the human body contour can be extracted by SLTP. So SLTP can describe human body shapes. An example of the SLTP extracted from a depth image is shown in Fig. 1.

A histogram of SLTP can be used as the feature descriptor of a depth image region. The histogram contains information about the distribution of different kinds of depth differences. For retaining spatial information, we separate a scanning window into non-overlapped blocks as shown in Fig. 1.(c). The block size in our experiments is 8×8 pixels. In some HOG-based methods [4, 13], the blocks are overlapped to achieve higher performance. Because the gradient magnitude sum varies with different blocks, histogram normalization and overlapped blocks are employed. For the SLTP histogram, normalization is not needed since the sum of a histogram from blocks with the same size is equal.

3. Dataset

We have established a large dataset, SZU Kinect People Dataset¹, for people detection in Kinect depth images. The images in the SZU Kinect People Dataset were collected by a Kinect [1] sensor. The subjects in images were manually labelled by a bounding box. There are 7260 positive samples labelled in the dataset. The positive samples are randomly divided into two sets. The first set is the training set and contains 4268 samples. The second set, the test set, contains the rest 2992 samples. The negative samples are from B3DO (Berkeley 3-D Object Dataset) [9]. B3DO is designed for household object detection, and human bodies are not included. We created 442,308 negative samples (59,508 for training and 382,800 for test) from B3DO. In order to evaluate approaches of people detection in both depth images and color (intensity) images, the corresponding color images were also collected. The depth images and color images are calibrated. The same object locates in the same position both in a depth image and a color one. The example images including the subjects from the SZU dataset performing different postures are shown in Fig. 2. The dataset is very challenging for people detection due to the diversity of human postures.

The depth information is recovered by the structured-light method [8]. Some depth pixels can be missing when the object is occluded or the object surface is infrared-absorbing, glass, etc. The missed pixels are shown in black color in Fig. 2. As described above, the proposed SLTP measures the depth difference. Usually the persons are in front of other objects, and most missed pixels do not lie on the human body surface. So in our experiments, the missed pixels are all assigned with the maximum depth value.

¹The dataset will be publicly available for research purpose after obtaining all subjects' authority.



Figure 2. Some depth images and the corresponding color ones from the SZU Kinect People Dataset.

4. Experimental Results

To evaluate the proposed SLTP feature descriptor, we design several sets of experiments. In these experiments, we choose the popular classifier, namely linear SVMs [12], as the learning paradigm. To evaluate different kinds of features, Detection Error Tradeoff (DET) curves [4] are employed. DET can describe the miss rates against false positives per window (FPPW). In [5], Dollár *et al.* suggest using false positives per image (FPPI) because FPPI provide a natural measure of error of a detection system. Since the main propose of this paper is to evaluate the SLTP descriptor, not to evaluate the whole detection system, we choose FPPW in our experiments.

4.1. Different thresholds

In Equation (3) and (4), the threshold T_d is user-specified. Its value would affect the detection performance. What value is the best one for people detection is what we concern. Besides, when the threshold varies, how the detection rate is affected is also investigated.

The first set of experiments are carried out with different T_d . The detection rates at FPPW= 10^{-4} are plotted in Fig.3, from which we can find that the detection rate reaches the peak when $T_d = 3.0$ centimeters. And then the detection rate decreases as T_d increases. So we choose $T_d = 3.0$ cm for SLTP feature extraction in our experiments.

4.2. Different window sizes

A smaller detection window size means lower computational cost and higher detection speed. But it can also cause a lower detection rate and a higher miss rate. According to our experimental results in Fig.4, we can find that the 64×128 window can only achieve slightly better perfor-

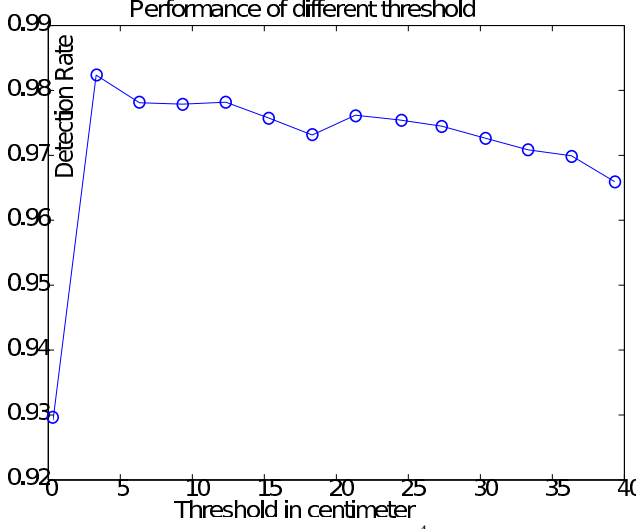


Figure 3. The detection rates at $FPPW=10^{-4}$ with different thresholds.

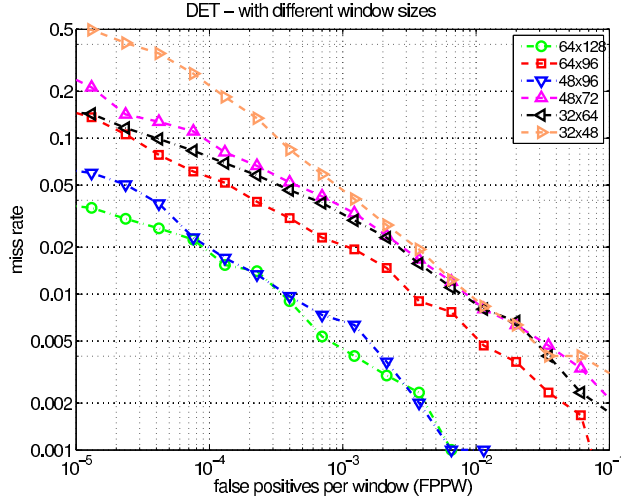


Figure 4. The miss rate vs. FPPW in the case of different detection window sizes.

mance than the 48×96 one. But the miss rate increases from 2% to 7% when the window size decreases from 48×96 to 32×64 . So the 48×96 detection window has better trade-off between the computational cost and the detection rate.

Because the Kinect sensor has usually better performance when the distance is less than 5 meters, the subjects are often relatively close to the sensor in some real applications. This means the the lower legs sometimes might not appear in the depth images. The window sizes, 64×96 , 48×72 and 32×48 , are employed. The feet and lower legs area is not included in those windows. The miss rates are higher than those with the feet and lower legs area (Fig.4).

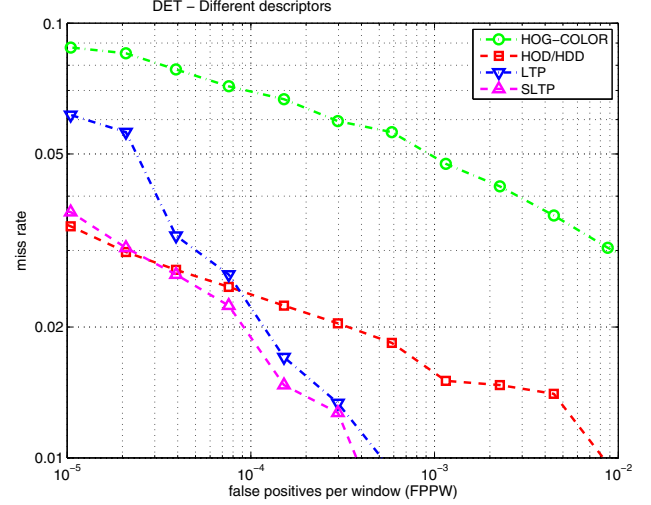


Figure 5. The comparison between SLTP and other features.

4.3. Comparison with HOG, HOD, HDD and LTP

HOG is now considered to be the best single feature for people detection in intensity/color images [6]. HOD and HDD follow the same procedure with HOG for the depth images [10, 14], and they can be considered as HOG for depth images. SLTP is inspired by LTP [11]. So the comparison between SLTP and HOG, HOD/HDD and LTP is carried out.

Four experiments are performed for feature comparison, i.e., HOG in color images, HOD/HDD in depth images, LTP in depth images, and SLTP in depth images. The experimental results are shown in Fig.5. From the results we can find that higher detection rates can be achieved using depth images than color images. This could be caused by the unique advantages of depth images, i.e., depth images are more robust to illumination and clothing textures. In addition, we find that SLTP can also achieve better performance than HOG, HOD/HDD and LTP.

4.4. Computational efficiency

The proposed SLTP feature extraction contains only simple subtraction, logical operation and pattern statistics, so the computational cost should be much lower than that of HOG. We compare the computational cost of SLTP with that of HOG. The HOG feature extraction procedure is from OpenCV (a high-quality implementation). To detect human body in a 320×240 image, HOG takes 80.6 milliseconds (12.4 FPS) on our desktop computer with an Intel i3 3.0GHz CPU and 1GB memory. The proposed SLTP takes only 11.6 milliseconds (86.2 FPS) on the same computer. In this comparison, the scanning window scale factor is the same value 1.1. SLTP is about 7 times fast than HOG. Another advantage is that SLTP can be run on some embedded systems

with a higher frame rate since there is no floating point operation in feature extraction.

In addition, for each block in a detection window, the LTP histogram contains 118 bins, but the SLTP histogram contains only 9 bins. That is, the dimension of the SLTP histogram is only $9/118 \approx 1/13$ of that of LTP. So SLTP is much faster than LTP. SLTP is a highly efficient feature descriptor with great discriminant capability.

5. Conclusions and Future Work

A new feature, SLTP, for people detection in depth images has been introduced in this paper. Experimental results have shown that SLTP can take advantage of depth images and gives better detection rates than HOG and LTP with low computational cost. Besides, the selection of the threshold in feature extraction and the detection window size is also suggested empirically by a series of experiments.

SLTP is a kind of statistical feature, in which only depth differences within a local region are considered. The human body size and some other body geometries are not yet taken as features here. If the statistical features are combined with some geometric features, higher detection rate with lower false positive rate would be achieved. We will investigate this in future work. Besides, we also will explore the use of SLTP for other detection tasks of different objects, such as chair, cup and bottle detection.

Acknowledgements

This work is partly supported by Shenzhen Fundamental Research Program (Grant No. JC200903170431A, JC201005280432A), the “Hundred Talent Program of CAS” grant and Natural Science Foundation of SZU (Grant No. 201211).

References

- [1] Microsoft Kinect for XBOX 360. <http://www.xbox.com/en-US/kinect>.
- [2] TOF cameras of MESA Imaging AG. <http://www.mesa-imaging.ch>.
- [3] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE TPAMI*, 28(12), 2006.
- [4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.
- [5] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, 2009.
- [6] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE TPAMI, Preprint*, 2012.
- [7] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE TPAMI*, 31(12), 2009.
- [8] B. Freedman, A. Shpunt, M. Machline, and Y. Arieli. Depth mapping using projected patterns. US Patent Application, Pub. No.: US 2010/0118123 A1.
- [9] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3-D object dataset: Putting the Kinect to work. In *ICCV Workshop on CDCCV*, 2011.
- [10] L. Spinello and K. O. Arras. People detection in rgb-d data. In *IROS*, 2011.
- [11] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. In *AMFG*, 2007.
- [12] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.
- [13] X. Wang, T. X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In *ICCV*, 2009.
- [14] S. Wu, S. Yu, and W. Chen. An attempt to pedestrian detection in depth images. In *Proc. of the 3rd Chinese Conf. on Intelligent Visual Surveillance*, 2011.
- [15] Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *CVPR*, 2006.