Contemporary Engineering Sciences, Vol. 9, 2016, no. 11, 547 - 552
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2016.6440

Human Detection in Top-View Depth Image

Tae-Won Choi

University of Science&Technology (UST)
217, Gajeong-ro, Yuseong-gu, Daejeon, 34129, Korea

Electronics and Telecommunications Research Institute (ETRI)
218, Gajeong-ro, Yuseong-gu, Daejeon, 34113, Korea

Dae-Hwan Kim

Electronics and Telecommunications Research Institute (ETRI)
218, Gajeong-ro, Yuseong-gu, Daejeon, 34113, Korea

Ki-Hong Kim*

University of Science&Technology (UST)
217, Gajeong-ro, Yuseong-gu, Daejeon, 34129, Korea
&
Electronics and Telecommunications Research Institute (ETRI)
218, Gajeong-ro, Yuseong-gu, Daejeon, 34113, Korea
*Corresponding author

Copyright © 2016 Tae-Won Choi, Dae-Hwan Kim and Ki-Hong Kim. This article is distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

The paper presents an efficient and reliable feature descriptor for human detection in a top-view depth image that uses two statistical values of mean and standard deviation. Human detection performance of our descriptor outperform Rauter that use mean value and Simplified Local Ternary Pattern (SLTP). To evaluate the human detection performance using our descriptor, we capture 559 positive and 2382 negative top-view depth images. We get 98.1% precision and 94.2% recall rates. Run-time performance of our descriptor is reasonably fast. Human detection using the proposed descriptor can be applicable in detecting multiple human and tracking real time in top-view depth images.

Keywords: depth image, feature descriptor, human detection

1 Introduction

Human detection or recognition has been very important in various applications including entertainment, robotics, and surveillance so that many researchers have studied the problem. Histogram of Oriented Gradients (HOG) descriptor [1] is very commonly used in human detection in color images. It divides detection window into many cells and calculates orientation of each pixel using difference of x- and y- directions. Human detection in a color image has been a challenging problem because of complex backgrounds, illumination changes, background clutter, intra-class variation, and so on.

Recently, there have been several approaches for human detection using depth information to solve those problems. Background separation can be easily done by using difference of distance. Influence of illumination changes is small and intra-class variance caused by difference in color is not a serious problem in a depth image. Histogram of Normal Vector (HONV) descriptor [2] calculates zenith and azimuthal orientation of normal vector of each pixel and accumulates them into histogram of divided cells. Simplified Local Ternary Pattern (SLTP) descriptor [3] computes differences of depth values along x- and y-directions, respectively, and quantizes them into 0, +1 and -1, and accumulates histogram of 9 different patterns per each pixel.

In some interaction systems, cameras are mounted over human head. Shape of human body looks quite different according to positions, especially in a top-view image, as shown in Figure 1. We can sometimes see only head and shoulder of human, and sometimes see whole body. Various hair styles make this problem much harder. Therefore it is not easy to recognize human in a top-view image. Rauter[4] searched local maxima as the center point and divided nearby area into many blocks calculating mean value per each of them. Blocks are accumulated into a histogram depending on distance from the center point.

In this paper, we present a new descriptor for human detection in a top-view depth image. Our descriptor shows the best detection performance compared to

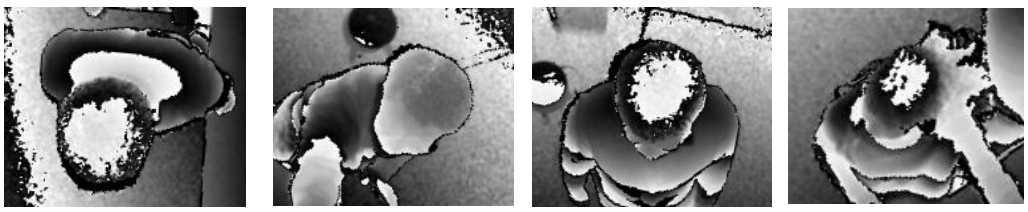

Figure 1. Variance of human shape in a top-view image

the existing ones. Training time and decision time of the classifier is also very fast, since our descriptor uses small memory and has simple calculating process.

2 Proposed method

We propose a descriptor for human detection in a top-view depth image. Our descriptor is based on sliding detection window. We divide a detection window into

many blocks and calculate mean and standard deviation per each block.

As shown in Figure 2, block mean values in a window describe the shape of an object and standard deviation highlights boundary of an object. We calculate mean and standard deviation per each block except for invalid pixels having zero depth value. We compute mean of each block, $\overline{D}_B$, as

$$\overline{D}_B = \frac{\sum_{\mathbf{x}\in B,\mathbf{x}\notin E} D(\mathbf{x})}{N_B - N_E}. \quad (1)$$

Here, **x** denotes a pixel position (*x*, *y*) in a block and *D*(**x**) denotes a depth value at **x**. N_B and N_E are the whole number of pixel in a block and the number of error pixel in a block, respectively. We calculate standard deviation of each pixel as:

$$\overline{S}_B = \sqrt{\frac{\sum_{\mathbf{x}\in B,\mathbf{x}\notin E} D^2(\mathbf{x})}{N_B - N_E} - \overline{D}_B^{\,2}}. \quad (2)$$

We make a descriptor by normalizing block means and standard deviations in a window, respectively, and vectorizing them. Note that when both of mean and standard deviations are normalized together, the descriptor rather confuses classifier. To normalize, we use l2-norm that shows the best performance among experimented normalization methods. The size of our descriptor is twice the number of a block size. We use linear support vector machine to train our descriptor classifier.

Figure 2. Depth image(left)), mean(center), and standard deviation

3 Experimental results

We compare our descriptor to SLTP and Rauter. We test various block sizes of 4x4, 8x8, and 12x12 with half size stride of block overlapping or non-overlapping. The proposed method is evaluated in terms of detection performance and run-time performance.

3.1 Detection performance

For the evaluation of detection performance, we build a dataset of 559 positive and 2382 negative top-view depth images. Among the dataset, 504 positive and 2144 negative images are used to train classifier and 55 positive and 238 negative images are used to test detection performance. Table 1 shows precision and recall rates of three different descriptors.

Precision rate of our descriptor is much higher than that of SLTP in all condition. Especially, precision rate of our descriptor looks uniform regardless of block size as shown in the left side of Figure 3. Our descriptor outperforms other descriptors in recall rates as shown in the right side of Figure 3.

	Precision			Recall		
block size / stride	SLTP	Rauter	our descriptor	SLTP	Rauter	our descriptor
12/12	0.834483	0.95069	0.959677	0.88	0.876364	0.865455
12/6	0.834483	0.960861	0.971209	0.88	0.892727	0.92
8/8	0.913761	0.974952	0.979087	0.905455	0.92	0.936364
8/4	0.913761	0.973025	**0.98088**	0.905455	0.918182	0.932727
4/4	0.925094	0.974806	0.979127	0.898182	0.914545	0.938182
4/2	0.925094	0.974806	0.973684	0.898182	0.914545	**0.941818**

Table 1. Precision and recall

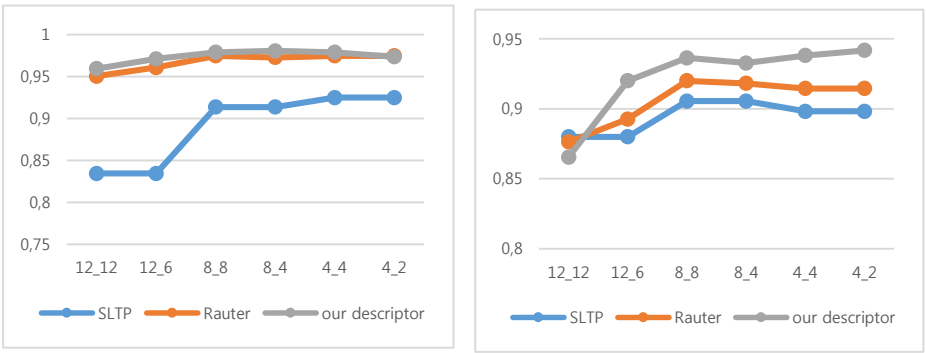

Figure 3. Precision and recall

	Description			Training and Detection		
block size / stride	SLTP	Rauter	our descriptor	SLTP	Rauter	our descriptor
12/12	2.759	2.786	2.784	0.0788	0.0183	0.0232
12/6	2.719	2.835	2.979	0.0796	0.0558	0.0755
8/8	2.729	3.083	3.087	0.2223	0.039	0.0584
8/4	2.916	3.23	3.153	0.2296	0.1683	0.2206
4/4	2.839	2.879	3.089	1.1833	0.1782	0.2347
4/2	2.752	3.253	3.482	1.2174	0.658	0.9411

Table 2. Run-time performance

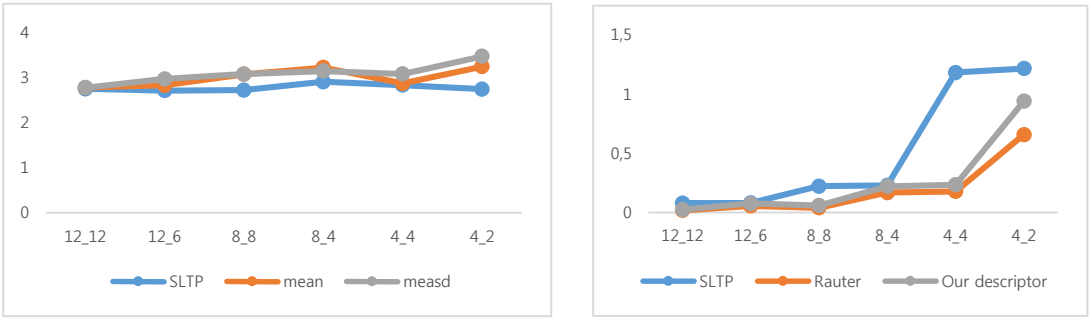

Figure 4. Run-time performance of description(left), training and detection

3.2 Run-time performance

We evaluate run-time performance including description time, training time, and classification time. Our descriptor is not the fastest but shows reasonable run-time performance. As shown in the left graph of Figure 4, description times are similar among tested descriptors. Refer to the right side of Figure 4, Training and detection times increase slowly when the block sizes are bigger than 4.

4 Conclusion and future work

We presented a new descriptor for human detection in a top-view depth image that use statistical value of mean and standard deviation. Experimental results showed that our descriptor provides the best detection performance and reasonable run-time performance, compared to SLTP and Rauter. Therefore, the human detection with the proposed descriptor can be applicable in embedded system. We plan to detect multiple human in a depth image and tracking them using our method.

Acknowledgements. This work was supported by the ICT R&D program of MSIP/IITP. (15501-14-1016, Instant 3D object-based Join & Joy content technology supporting simultaneous participation of users in remote places and enabling realistic experience).

References

[1] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (2005), 886-893. http://dx.doi.org/10.1109/cvpr.2005.177

[2] S. Tang, X. Wang, X. Lv, T.X. Han, J. Keller, Z. He, M. Skubic and S. Lao, Histogram of oriented normal vectors for object recognition with a depth sensor, Chapter in *Computer Vision-ACCV 2012*, Springer Berlin Heidelberg, 2012, 525-538. http://dx.doi.org/10.1007/978-3-642-37444-9_41

[3] S. Yu, S. Wu and L. Wang, Sltp: A fast descriptor for people detection in depth images, *Proc. of 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, (2012), 43-47. http://dx.doi.org/10.1109/avss.2012.67

[4] M. Rauter, Reliable human detection and tracking in top-view depth images, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (2013), 529-534. http://dx.doi.org/10.1109/cvprw.2013.84

Received: April 11, 2016; Published: June 2, 2016