2014 3rd International Conference on Informatics, Environment, Energy and Applications
IPCBEE vol.66 (2014) © (2014) IACSIT Press, Singapore
DOI: 10.7763/IPCBEE. 2014. V66. 13

A Research on Position Shift Pedestrian Image Detection using Visual Words Selected Method

Xingguo Zhang [+], Kazuki Saruta, Yuki Terata and Guoyue Chen

Graduate School of Systems Science and Technology, Akita Prefectural University, Akita, Japan

Abstract. Pedestrian detection is an important area in computer vision with key applications in intelligent vehicles. The objective of this paper is to detect position shift pedestrian using Visual Words selection method which is based on Bag-of-Features. It calculates the difference of the total appearance frequency for each Visual Word of pedestrian and non-pedestrian images. The Visual Word which exhibit greater absolute value is more efficient for pedestrian detection and can be selected. The effectiveness of Visual Words selection method has been validated by the distribution analysis of selected feature points. It is show that the features which located in lower body area are more effective for the detection.

Keywords: Bag-of-Features, Visual Word, pedestrian detection, feature points distribution.

1. Introduction

This Pedestrian detection has attracted an extensive amount of interest from the computer vision community over the past few years. Many techniques have been proposed in terms of features, models, and general architectures. However, there still exist many challenges, especially for detection to pedestrians by an on-board camera within a vehicle in the night [1].

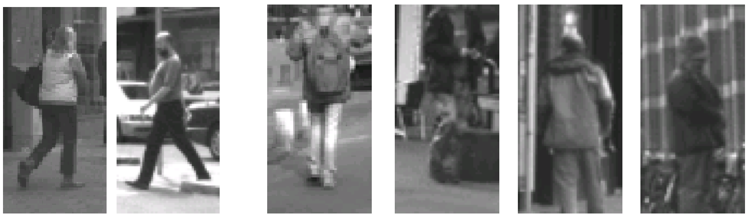
The simplest technique to obtain initial object location hypotheses is the sliding window technique, where detector windows at various scales and locations are shifted over the image. The computational costs of high-precision detection approach for every sliding window are often too high to allow for real-time processing [2], [3], [4]. To speed up the detection process, we decompose pedestrian detection into the potential target detection and classification (model matching). Firstly, the system defines a region of interest (ROI), which is associated possibly with a potential pedestrian. The Haar wavelet-based cascade framework [5] have been shown to be powerful features in the domain of ROI selection, in combination with AdaBoost to construct a classifier. Secondly, detection is validated by high-precision identifying method for the ROI.Some popular pedestrian detection methods are based on histogram of oriented gradients (HOG) feature descriptors [2]. An extension of edgelet features to detect pedestrian has also shown good performance [3]. However, these features are sometimes ineffective for classification performance. Especially it can be problematic when the pedestrian position is not in the middle of the ROI, as shown in Fig. 1 (b). Therefore this paper focus on pedestrian detection not only normal images (as shown in Fig. 1(a)) but also the images pedestrian positions shift inside a ROI existed.

To deal with the classification problem for deformable objects, Bag-of-Features (BoF) have been proposed by G. Csurka et al. In [6], all features are clustered into a visual vocabulary, and the frequency histogram of each Visual Word is used as the input to train a classifier. One advantage of this method is that the frequency histogram is irrelevant with the location of the local feature and is very useful to detect the image of pedestrian with position shift. However, the size of all Visual Words will increase and bring more complexity. The computational costs are often too high to allow for real-time processing. A simple and efficient visual vocabulary is expected to speed up learning and classification. In [7] we proposed a method

[+] Corresponding author. Tel.: + (81-184-272087).
E-mail address: (d14s001@akita-pu.ac.jp).

to reduce the dimension of the classifier by setting a limit value to remove irrelevant and redundant Visual Words which are not useful to classification. The experiment showed that the time in learning and detection process with nearly same precision can reduce about 50% by proposed method even if only 40% Visual Word be selected. However, we have not discussed the detection precisions with different position shift condition.

In this paper, our objective is to detect position shift pedestrian using Visual Words selection method which is based on Bag-of-Features. Firstly, the Visual Words selection method will be done a brief retrospect. Secondly, the selected feature points will be visualized to distinguish each pedestrian position shift condition.



(a) full images (b) position shift images

Fig. 1: Some pedestrian images under different conditions.

2. Pedestrian Detection Process

2.1. Basic pedestrian detection outline

Fig. 2 shows the outline of pedestrian detection in our approach, which consists of feature extraction, building of visual vocabulary, building a frequency histogram, screen out of Visual Words, building a new frequency histogram, and training the classifier.

At first, local features are extracted from training samples, and are clustered into X groups with the K-means algorithm. After clustering, the visual vocabulary is built, and frequency histogram of each Visual Word, which records the number of it occurrence, is calculated. A compact and efficient new visual vocabulary is reconstructed by screening out some redundant Visual Words and more details are discussed later. New frequency histogram based on new visual vocabulary is considered as the input of classifier, which is trained with the SVM algorithm, and then the classifier will make decision based on these remained features.

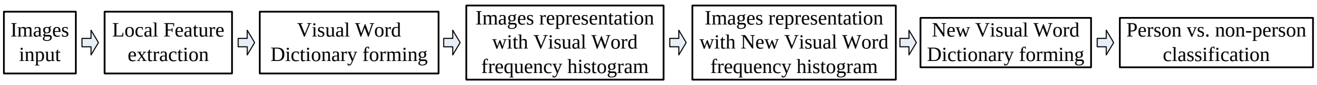

Images input
Local Feature extraction
Visual Word Dictionary forming
Images representation with Visual Word frequency histogram
Images representation with New Visual Word frequency histogram
New Visual Word Dictionary forming
Person vs. non-person classification

Fig. 2: The outline of proposed pedestrian detection method.

2.2. Visual Words selection method

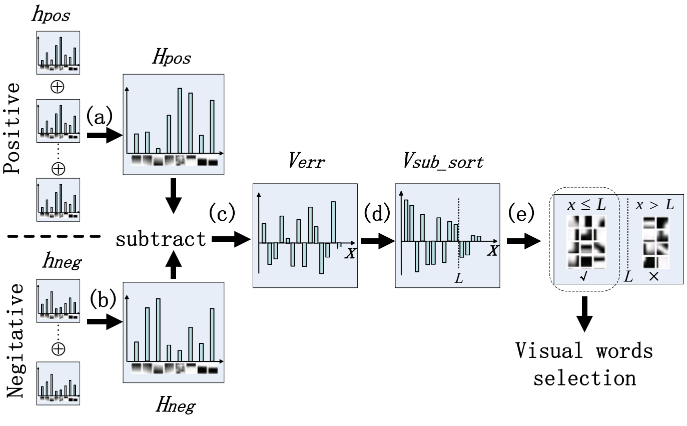


Fig. 3: Flowchart of proposed method.

In this section, we present the process to select the efficient Visual Words. One of disadvantages of dense regular grid is that large redundant features are included into the visual vocabulary, and it will spend more time on feature extraction and classification during the training and recognition stage. A simple and efficient Visual Word is expected to speed up learning and classification. Here we propose a method to reduce the dimension of the classifier by setting a limit value to remove irrelevant and redundant Visual Words which are not useful to classification.

A brief overview of this approach is given in Fig. 3. Firstly, image samples for pedestrian and non-pedestrian are labelled as positive sample and negative sample, respectively. All features in one image are clustered into X Visual Words with the K-means algorithm, and their frequency histogram $h_{pos}(x)$ or $h_{neg}(x)$ for each Visual Word x is calculated. After all training samples are analyzed, the total frequency histogram for positive sample and negative sample is computed as shown in Fig. 3 (a, b).

$$\begin{cases} H_{pos}(x) = \sum_{m=1}^{M} h_{pos}(x,m) \\ H_{neg}(x) = \sum_{n=1}^{N} h_{neg}(x,n) \end{cases} \quad x \in [1, X] \tag{1}$$

where M and N are the number of positive samples and negative samples, respectively, and X is the chosen size of Visual Words in a dictionary. $H_{pos}(x)$ represents the total number of Visual Word x extracted from all positive samples.

Then we normalize two frequency histograms as

$$\begin{cases} \overline{H_{pos}(x)} = \dfrac{H_{pos}(x)}{\sum_{i=1}^{X} \left(H_{pos}(i)\right)^2 + \varepsilon} \\ \overline{H_{neg}(x)} = \dfrac{H_{neg}(x)}{\sum_{i=1}^{X} \left(H_{neg}(i)\right)^2 + \varepsilon} \end{cases} \quad (\varepsilon = 0.0001) \tag{2}$$

And the difference between $H_{pos}(x)$ and $H_{neg}(x)$ is calculated to obtain an error vector $V_{err}(x)$=$H_{pos}(x) - H_{neg}(x)$ (Fig. 3 (c)). Plus value in V_{err} means this Visual Word is effective classify positive sample and minus value means effective for negative sample. Larger absolute value of $V_{err}(x)$ means that the x'th feature is more beneficial to classification.

We sort the Visual Words in descending order of absolute value and set a limit value L to determine the expected size of the new visual vocabulary we want to preserve (Fig. 3 (d)). The Visual Words with $V_{err}(x)$ below the limit value L will be considered as redundant Visual Words and are screened out of the original dictionary, and the remained L Visual Words consist of a new visual vocabulary (shown in Fig. 3 (e)). The next we remove corresponding dimensions of original histogram h_m and h_n according to the new visual vocabulary. The new frequency histogram of visual vocabulary is the input of the classifier, which is trained with the SVM algorithm.

As the experimental results of proposed method, the precision variation was very small when L≥200, which means that 200 efficient Visual Words in the original dictionary can result in almost the same performance as it with all 500 Visual Words. The detection precision decreases become more quickly as L decreases for L<200. In addition, the training and classification time by the SVM algorithm at different size of new visual vocabulary is shown in Fig. 5. Both decrease quickly as L decreases. It is shown that the computational time can be saved almost 50% at L=200 and almost the same recognition accuracy is kept as original Visual Words size 500. In other words, the experiment showed that the time in learning and detection process with nearly same precision can reduce about 50% by proposed method even if only 40% Visual Words be selected.

3. Experiments and Results

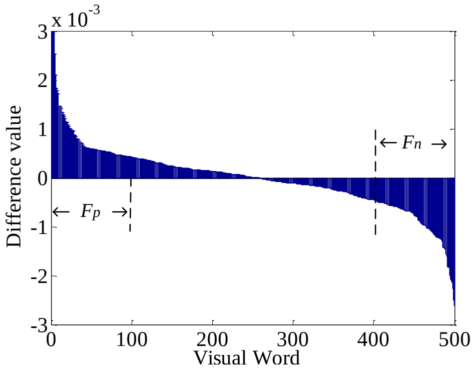
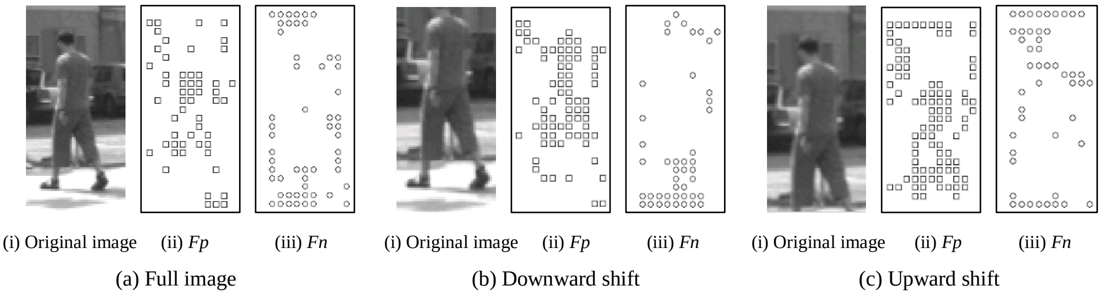


Fig. 6: The justification of *Fp* and *Fn* with Visual Word.

In this section the distributions of selected feature points from pedestrian images and non-pedestrian images will be analyse. And the detection precision and distribution analyse for position shift images also be given.

The result of sorting the difference vector which is obtained by Fig. 3(c) stage with an initial Visual Word dictionary size N=500 as shown in Fig. 6. The horizontal axis represents the Visual Words and the vertical axis represents the difference value of the total appearance frequency by each Visual Word between pedestrian image and non-pedestrian images. The Visual Word which has positive difference value is contributes to the determination of the pedestrian, and negative difference value is contributes to the determination of the non-pedestrian. In Fig. 6, the top 100 Visual Words *Fp* represents the effect on the determination of pedestrian, 100 lower Visual Words *Fn* represents effective in the determination of non-pedestrian.



(a) Full image

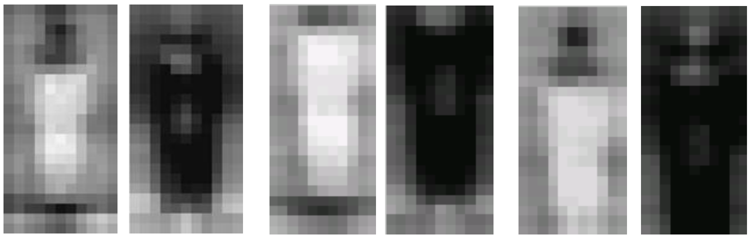
(b) Downward shift

(c) Upward shift

Fig. 7: Visualization examples of selected feature points.

Fig. 7 shows some examples of the distribution of selected feature points from images. (a) is the Full image, (b) is obtained by the image(a) with one fifth downward shift, (c) is obtained by the image(a) with one fifth upward shift. The circles are *Fp* feature points distribution, the square represent the position of *Fn* feature points. It can be seen that *Fp* are mainly distributed to the body region in the pedestrian images, *Fn* are mainly distributed to the background of except body areas. With the pedestrian position shift, the subject positions of feature points are shifted.

Fig. 8 shows average distribution of selected feature points from 3 kinds of pedestrian images 3000 respectively. The white areas represent which areas have further quantity. The black areas represent the quantity are less. It can be seen that *Fp* are mainly distributed to the body region in the pedestrian images, however, more features are distributed to the middle and lower body region. *Fn* are mainly distributed to the background of body region. Fig. 9 shows the relation between position shift and detection precision by DET curve. It can be seen that downward shift has little effect on precision but upward shift will have an impact. It means that the features which located in lower body area are more effective for the detection.



(i) *Fp* (ii) *Fn* (i) *Fp* (ii) *Fn* (i) *Fp* (ii) *Fn*

(a) Full image (b) Downward shift (c)Upward shift

Fig. 8: Average distribution of selected features.

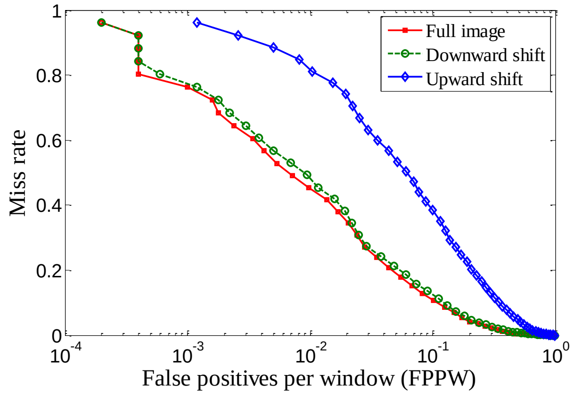


Fig. 9: The relation between position shift and detection precision.

4. Conclusion

In this paper, Visual Words selection Method was been used to detect the position shift pedestrian. The effectiveness of the method has been validated by the distribution analysis of selected feature points. In addition, the detection precisions with different position shift condition have been discussed. The experiment results are show that downward shift has little effect on precision but upward shift will have an impact. It means that the features which located in lower body area are more effective for the detection.

5. References

[1] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol.31, no.12, pp. 2179-2195, 2009.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. vol.2, pp.886-893, 2005.

[3] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *Int'l J. Computer Vision*. vol.75, no.2, pp.247-266, 2007.

[4] P. Sabzmeydani and G. Mori. Detecting Pedestrians by Learning Shapelet Features. *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 2007.

[5] P. Viola, M. Jones, and D. Snow. Detecting Pedestrians Using Patterns of Motion and Appearance. *Int'l J. Computer Vision*. vol. 63, no. 2, pp. 153-161, 2005.

[6] G. Csurka, C. R. Dance, L. Fan, J. Willamowski and C. Bray. Visual categorization with bags of keypoints. *ECCV International Workshop on Statistical Learning in Computer Vision*. pp.1–22, 2004.

[7] X. Zhang, K. Saruta, Y. Terata, and G. Chen. Visual Words Simplified Method for Fast Pedestrian Detection. *Intl. Conf. on Computational Intelligence and Software Engineering*. pp. 47-50, 2012.