Robust Human Detection using Bag-of-Words and Segmentation

Yuta Tani and Kazuhiro Hotta
Meijo University, 1-501 Shiogamaguchi, Tenpaku, Nagoya, 468-0051, Japan

Keywords: Human Detection, Bag-of-Words, Pose Variation, Occlusion, GrabCut and Color Names.

Abstract: It is reported that Bag-of-Words (BoW) is effective to detect humans with large pose changes and occlusions in still images. BoW can make consistent representation even if a human has pose changes and occlusions. However, the conventional method represents all information within a bounding box as positive data. Since the bounding box is the rectangle including a human, background region is also included in BoW representation. The background region affects BoW representation and the detection accuracy decreases. Thus, in this paper, we propose to segment the region by GrabCut or Color Names, and the influence of background is reduced and we can obtain BoW histogram from only human region. By the comparison with the deformable part model (DPM) and conventional method using BoW, the effectiveness of our method is demonstrated.

1 INTRODUCTION

In recent years, people deal with a large number of images as spread of digital cameras and cell phones with a camera. Therefore, it is desired that computers recognize the semantic contents of images automatically. In general, there are many pictures in which humans are doing actions. Thus, for image understanding, recognition of action in still images is important. It is reported that the usage of both the foreground and background regions is effective for object recognition in still images (Russakovsky et al., 2012). For action recognition in still images, the detection of human with actions is required to represent foreground and background independently.

Human with actions in real images often has large pose changes and occlusions. In the case, the detection is difficult even if we use Deformable Parts Model (DPM) (Felzenszwalb et al., 2010). In the conventional method (Tani and Hotta, 2014), BoW representation is used to cope with large pose changes and occlusions, and it gave superior accuracy to DPM. However, the conventional method represents all information within a bounding box as positive data. Some bounding boxes are shown with red rectangle in Figure 1, which also include background region. In addition, the background area is large when human raises his hand or opens his leg. Thus, the conventional method is influenced by background region.

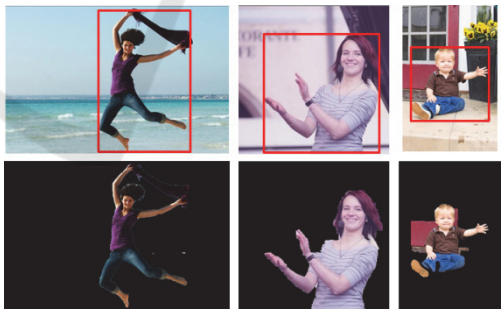


Figure 1: Top row shows positive training images. The red rectangle is the bounding box. It also includes background region. Bottom row is the segmentation results using the GrabCut. If we represent only the segmented human by BoW, the bad influence by background is reduced and the accuracy will be improved further.

Furthermore, in the test step of the method, the test image is divided into grid, and the combination of the divided regions is represented by BoW. The background region also affects the detection result.

In this paper, we use only human region segmented by GrabCut (Rother et al., 2004) as positive training data in order to reduce influence of background. GrabCut does not segment a human region perfectly but it gives good segmentation result as shown in Figure 1. We see that background is removed well. For negative training data, we crop the regions randomly from the outside of the bounding box and represent the region by BoW. We train the SVM using those positive and negative data.

Tani Y. and Hotta K..
Robust Human Detection using Bag-of-Words and Segmentation.
DOI: 10.5220/0005354705040509
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 504-509
ISBN: 978-989-758-090-1

However, GrabCut requires the bounding box including a human. In the training phase, we have a bounding box but we cannot use the bounding box in test phase. Thus, we use the Color Names (Weijer and Schmid, 2007) instead of GrabCut. First, we convert RGB image into Color Names image. If the adjacent pixels are classified as the same color, we put the same label on the region. We can represent various combination of the labeled regions by BoW histogram, and we feed them into SVM. We detect a human as the combination of labeled regions with the maximum score. The details are described in Section 2.

In experiments, we use the Stanford 40 dataset (Yao et al., 2011). The dataset contains images in which people appear with large pose changes and occlusions. In evaluation, the accuracy is measured by the average overlapping rate of the area between the detected region and the ground truth attached to test images. As the baseline, we evaluate DPM. It achieves 28.42% while our method achieves 52.07%. The method without segmentation by GrabCut or Color Names achieves 48.45%. These results demonstrate that our method is effective for human detection with partial occlusion and pose changes.

This paper is constructed as follows. Our human detection method is explained in Section 2. The experimental results using the Stanford 40 dataset are shown in Section 3. Comparison with DPM and the conventional method is also shown. Finally, conclusion and future work are described in Section 4.

2 DETECTION USING BAG-OF-WORDS OF SEGMENTED REGIONS

In the conventional method (Tani and Hotta, 2014), BoW representation is used to cope with large pose changes and occlusions. A codebook is made by k-means of RootSIFT (Arandjelovi'c and Zisserman, 2012). The method was superior to the DPM. However, the conventional method represented all information within the bounding box by BoW. Thus, the conventional method was influenced by background region. Furthermore, in the test step, the image is divided into grid and the combination of grid is represented by BoW. Thus, the background region also affects the detection result.

In this paper, GrabCut is used to segment a human in training. In test phase, we segment the region using the Color Names, and the combination of segmented regions is fed into SVM. By using only the segmented region without background, the accuracy is improved.

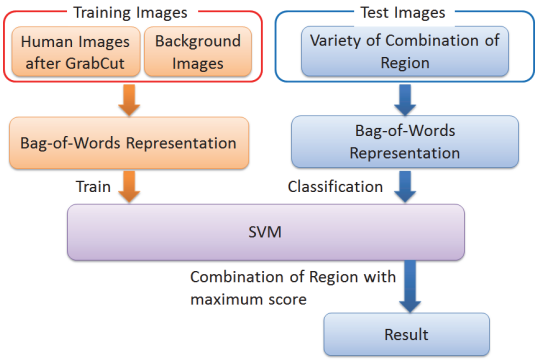


Figure 2: Overview of the proposed method.

An overview of our detection method using BoW of segmented regions is shown in Figure 2. As described previously, we use the GrabCut result for the bounding box as the positive training data. The negative samples are generated by cropping the region randomly except for human regions. We train the SVM using BoW of those positive and negative samples.

In the test phase, the bounding box is unavailable. Thus, we use Color Names instead of GrabCut. We segment the region based on Color Names, and BoW histograms of various combination of the segmented regions are computed. We feed them into the SVM, and the combination of region with the maximum output is used as the detection result.

2.1 Bag-of-Words Representation

We use a standard BoW representation. A codebook is made by k-means of RootSIFT similar to the conventional method. In general, the location information of local features is helpful in object categorization (Lazebnik et al., 2006). However, when a human has occlusions or pose changes, it makes the feature vector inconsistent. Thus, we use the standard BoW representation to cope with such cases.

In the experiments, in order to be independent of the image size, we extract RootSIFT features with grid spacing of 2% and patch size of 1%, 3% and 5% for the smaller width or height of each image. The number of visual words is set as 1000. These parameters are the same as the conventional method (Tani and Hotta, 2014). For fair comparison, we use the same parameters. We train a SVM with a Hellinger kernel (Vedaldi and Zisserman, 2010). By using linear SVM after taking root of elements in a

BoW histogram, the nonlinearity can be used without increasing the computation time.

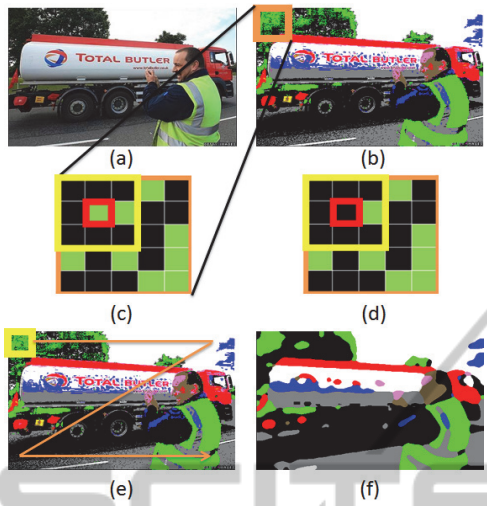


Figure 3: The details of filtering. We convert input image (a) into Color Names (b). (c) is the enlarged region of orange rectangle region in (b). We pay attention to local 3×3 region shown as yellow, and the center pixel shown as red is replaced by the most frequent Color Name in the 3×3 region as (d). We carry out this process to all pixels in the image shown in (e). The result after applying some filters is shown in (f).

2.2 Segmentation using GrabCut

It is reported that GrabCut gave good segmentation result (Gavves et al., 2013). Furthermore, it can segment the region from the bounding box without human interaction. Since the Stanford 40 dataset used in experiments contains the bounding box including a human, we segment human region by GrabCut from the bounding box. Some segmentation results are shown in Figure 1. Segmentation result is not perfect but GrabCut segments human region roughly.

Figure 4: Left image is the result by applying a 3×3 filter ten times. Applying the small filter many times does not decrease the number of regions much. Right image is the result by applying a 21×21 filter once. The large filter loses the fine edge. For example, sky and the window of the truck become together.

2.3 Segmentation using Color Names

Since the bounding box is unavailable in test phase, we cannot use GrabCut. Thus we use the Color Names instead of using GrabCut in order to segment regions. Since the 11 basic colors are fixed in Color Names, it is suited to rough segmentation. First, we convert RGB (Figure 3 (a)) image into Color Names (Figure 3 (b)) and divide the image into regions based on the label assigned by Color Names. If the adjacent pixels has the same color, the same label is attached to the region. However, when the number of regions is large, the number of combination becomes huge and the computational cost is also high. Thus, we apply a filter to reduce the number of regions and the computation time. Here we pay attention to 3×3 region as shown in Figure 3 (c), and the center pixel is replaced by the most frequent Color Name in the 3×3 region as shown in Figure 3 (b). This process is carried out for all pixels in the image as shown in Figure 3 (e).

If a small filter is applied to the image many times, the number of region does not decrease as shown in Figure 4. If a large filter is used, the fine edges are lost. To avoid those cases, we apply the filters while changing the filter size as 3×3, 5×5, 7×7, 9×9, 11×11 pixels. The result after applying these filters is shown in Figure 3 (f). By applying the filter with different sizes, we can decrease the number of region while maintaining fine edges.

Next, we reduce the number of regions further. In order to simplify the description, we treat the image of Figure 5 (a) as the result after applying filters. We search the smallest region (the V region shown in (a)) and the region is merged to the smallest adjacent region of V in the image. In the example, the region is III, and III and V are merged. We search the smallest region again and the same process is repeated till the number of regions becomes threshold or less. In the experiment, we set the threshold value as 20.

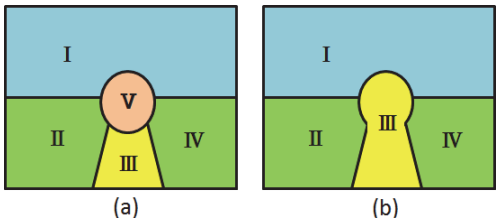


Figure 5: We search the smallest region in (a). In this example, it is V. The region is merged to the smallest adjacent region III as shown in (b).

2.4 Detection Method

As described previously, in the test step, we segment the region using Color Names. After that, we compute the score of all combinations of the labeled

regions. In the case of Figure 5, all combinations are shown in Figure 6 (a). The non-adjacent combinations such as [II, IV] are not used. We represent all combinations by BoW. For example, in the case of combination [I, IV], the shaded region shown in Figure 6 (b) is represented by BoW. We feed those representations into SVM. Finally, we detect the combined region with the maximum score as a human.

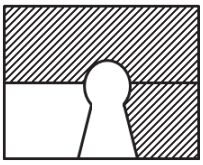
3 EXPERIMENTAL RESULTS

In the experiments, we use the Stanford 40 dataset (Yao et al., 2011). The dataset is originally made for action recognition in still images and contains 40 daily human actions. The bounding boxes including a human are already given, and we use them for training and evaluating the detection results.

For training, we made positive images by using GrabCut from training images and made negative images by randomly cropping the regions except for the bounding box. Consequently, the training set consists of 3,136 positive and 15,703 negative images. We represent them by BoW and train SVM by LIBLINEAR (Fan et al, 2008).

Ⅰ	Ⅰ,Ⅱ	Ⅰ,Ⅱ,Ⅲ	Ⅰ,Ⅱ,Ⅲ,Ⅳ
Ⅱ	Ⅰ,Ⅲ	Ⅰ,Ⅱ,Ⅳ	
Ⅲ	Ⅰ,Ⅳ	Ⅰ,Ⅲ,Ⅳ	
Ⅳ	Ⅱ,Ⅲ	Ⅱ,Ⅲ,Ⅳ	
	Ⅲ,Ⅳ		

(a)



(b)

Figure 6: We compute the score of all combinations of the labeled regions. In the case of Figure 5, all combinations are shown in (a). We do not use the non-adjacent combination such as [II, IV]. We represent all combinations by BoW. For example, in the case of combination [I, IV], the shaded region shown in (b) is represented by BoW, and we compute the SVM score. The combination of region with the maximum score is detected as a human.

We use 4,345 test images from the Stanford 40 dataset. Test images also include humans with various poses and occlusions.

The proposed method is compared with DPM and the conventional method using BoW (Tani and Hotta, 2014). We explain each method.

(A) Deformable Part Model:

As the baseline method, we detect humans by DPM. The available annotation on the Stanford 40 dataset is only the bounding boxes indicating a human. Thus it is difficult to train the DPM using only those annotations, and we use the available source code with pre-trained model obtained from (http://cs.brown.edu/~pff/latent-release4/). The model is trained with the INRIA Person Dataset. Of course, since the training samples are different from our method, the direct comparison with DPM is difficult but we show the result as a reference.

(B) Conventional method using BoW:

This method represents all information within the bounding box by BoW. Thus the method is influenced by background region contained in the bounding box. Furthermore, in test step, the image is divided into grid and the combination of divided region is represented by BoW. The background region also affects the detection result.

(C) Our proposed method:

Our method segments the region by GrabCut or Color Names to reduce the influence of background. We compute BoW histograms of the combination of the segmented regions, and the BoW histograms are fed into SVM. We detect the combination of the segmented region with the maximum SVM score.

We evaluate the detection results by overlapping rate R defined as

$$\mathrm{R} = \frac{T \cap D}{T \cup D} \times 100[\%] \quad (1)$$

where T is the area of the ground truth and D is the area of the detection result. The value of R increases as the overlapping area of the T and D becomes large. If they match perfectly, the overlapping rate becomes 100%. Since we want to evaluate how much detection result matches to the ground truth, we use this evaluation measure.

Table 1: Average overlapping rate for each method.

	(A)	(B)	(C)
Overlapping rate	28.42%	48.45%	52.07%

Table 1 shows the average overlapping rate for test images in each method. The DPM (A) is inferior to other two methods. By using BoW instead of DPM, it becomes robust to partial occlusion and pose changes. This result shows that BoW is effective for detection tasks when the appearance of a human is much different.

By the comparison with the methods (B), we see that the proposed method (C) is better than the conventional method (B). The accuracy improvement is 3.62%. This result shows that the background region influences BoW histograms and the combination of segmented regions is effective for improving the accuracy.

Figure 7: Comparison results. The blue rectangle is the ground truth and the red rectangle is the detection result.

Some detection results by each method are shown in Figure 7. The blue rectangle shows the ground truth and the red rectangle shows the detection result. We see that the proposed method (C) can detect a human correctly. In contrast, the DPM (A) gives poor result for this dataset. One reason is that another dataset is used for training. Since the INRIA Person Dataset is made for the pedestrian detection, the resolution of human image is low. Another reason is that DPM cannot treat the large pose changes even if whole body appears. This decreases the accuracy. However, when human appears neither large pose changes nor occlusions, DPM (A) can detect humans well.

The conventional method (B) represents all information within the bounding box. Since the background region within the bounding box is also trained as positive data, the method (B) tends to detect a human with large background. By using segmentation, the background is reduced and the proposed method (C) can detect human with higher accuracy

4 CONCLUSION AND FUTURE WORK

In this paper, we proposed the method for detecting a human using Bag-of-Words of the combination of regions segmented by Color Names. BoW is robust to partial occlusions and pose changes. Furthermore, we can reduce the influence of background region by combining the segmented regions. This improves the detection accuracy.

In the proposed method, merging the smallest region with adjacent region in test phase is forcible way. If we use another segmentation method, the

detection accuracy may be improved. It is worth trying to use the SLIC which gives a good segmentation quality (Achanta et al., 2010). This is a subject for future work.

ACKNOWLEDGEMENTS

This work was partially supported by KAKENHI Grant Number 24700178.

REFERENCES

Russakovsky, O., Lin, Y., Yu, K. and Fei-Fei, L., 2012. Object-centric spatial pooling for image classification, European Conference on Computer Vision.

Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D., 2010. Object Detection with Discriminatively Trained Part Based Models, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 9, Sep.

Csurka, G., Dance, C., Fan, L., Willamowski, J. and Bray, C., 2004. Visual Categorization with Bags of Keypoints, Proc. of ECCV Workshop on Statistical Learning in Computer Vision, pp. 59–74.

Arandjelovi'c, R. and Zisserman, A., 2012. Three things everyone should know to improve object retrieval, In IEEE Conference on Computer Vision and Pattern Recognition, pp. 2911-2918.

Discriminatively trained deformable part models. http://cs.brown.edu/~pff/latent-release4/

Lazebnik, S., Schmid, C. and Ponce, J., 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, In IEEE Conference on Computer Vision and Pattern Recognition, pp. 2169-2178.

Fan, R., Chang, K., Hsieh, C., Wang, X. and Lin, C. 2008. LIBLINEAR: A library for large linear classification, Journal of Machine Learning Research 9, pp. 1871-1874.

Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L.J. and Fei-Fei, L., 2011. Human Action Recognition by Learning Bases of Action Attributes and Parts, Internation Conference on Computer Vision.

INRIA Person Dataset http://pascal.inrialpes.fr/data/human/

Vedaldi, A. and Zisserman, A., 2010. Efficient Additive Kernels via Explicit Feature Maps, In IEEE Conference on Computer Vision and Pattern Recognition, Vol. 34, No. 3, pp. 480-492.

Tani, Y. and Hotta, K., 2014. Robust Human Detection to Pose and Occlusion Using Bag-of-Words, International Conference on Pattern Recognition, pp. 4376-4381.

Rother, C., Kolmogorov, V., and Blake, A., 2004. GrabCut: Interactive foreground extraction using iterated graph cuts, The ACM Special Interest Group on Computer Graphics, Vol. 23, pp. 309-314.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and. Susstrunk, S. 2010. SLIC superpixels, Technical report, EPFL.

Weijer J., Schmid C., 2007, Applying Color Names to Image Description, In IEEE Conference on Computer Vision and Pattern Recognition, Vol. 3, pp. 493-496.

Gavves E., Fernando B., Snoek C.G.M., Smeulders A.W.M., and Tuytelaars T, 2013, Fine-Grained Categorization by Alignments, In IEEE International Conference on Computer Vision.