RELATIVE-DISTANCE-BASED SOFT VOTING FOR FEATURE REPRESENTATION AND ITS APPLICATION TO HUMAN ATTRIBUTE ANALYSIS

Toshihiko Yamasaki [1,2,3]

[1] School of Electrical and Computer Engineering, Cornell University,
[2] Dept. of Information and Communication Eng., The University of Tokyo,
[3] JSPS Postdoctoral Fellow for Research Abroad

ABSTRACT

This paper proposes a soft voting based bag-of-features (BoF) model considering relative distance of the feature vectors to the nearest-neighbor codeword. Whereas state-of-the-art kernel distance based soft voting methods require brute force parameter optimization, which is time consuming, the proposed method does not require any optimization. The proposed algorithm was applied to human attribute analysis using top-view images. The experimental results have demonstrated 100% of accuracy for both gender classification and baggage possession classification. It has also been demonstrated that discriminative ability is comparable to that of the fine-tuned codeword uncertainty (UNC) model.

Index Terms— Bag of features, soft voting, human attribute analysis

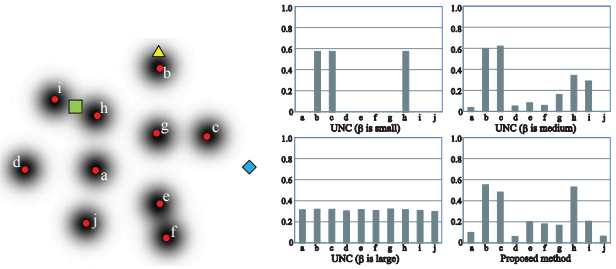
1. INTRODUCTION

A bag-of-features (BoF) model [1] is one of the most successful approaches for efficient and effective feature representation for image/video classification and retrieval. The basic idea of the BoF is to represent the input data as a histogram of code indices of the features regardless of their spatial and temporal orders. In order to quantize the feature vectors and to assign index values to them, k-means-based clustering is commonly used. The representative vectors (codewords) obtained in the clustering are called "visual words" in analogous to the bag-of-words model [2], which is used in text retrieval. Namely, a collection of features extracted from the input data is clustered to assign an index value to each feature vector and a histogram of such indices is used as a new feature vector for further processing. For instance, in [1] which first introduced the concept of BoF in image retrieval, local feature descriptors using scale invariant feature transform (SIFT) [3] were used to form a histogram for each video frame. The same concept can also be extended to the temporal domain [4][5].

After the great success of the BoF model, a number of techniques to improve the performance have been proposed [6][7][8]. Philbin et al. discussed the optimal number of clusters [6]. In [7], the spatial distribution of the local feature descriptors were taken into consideration by dividing the input images into sub-regions in a pyramid manner. A space-time pyramid with adaptive multiple kernel learning was also proposed for robust video event recognition [8].

Another significant progress in generating the feature vector was soft voting. In [1], for instance, only a single codeword was assigned to each feature vector. Recent studies have shown that soft voting that assign two or more codewords with weight values contributes to improving the image/video classification and retrieval performance [9][10][11]. In [9], the weight assigned to each feature vector was an exponential function of the distance to the codewords. In the codeword uncertainty (UNC) model [10], although the weights were also based on the exponential function of the distance as in [9], the weights were further normalized by the sum of the distances to all the codewords. In [11], the graphical structure among codewords was investigated. Refs. [12][13] tried to assign a few soft codewords to each feature vector using reconstruction based coding instead of using soft voting.

Although the soft voting method yields better performance than the conventional hard voting method, an extra parameter optimization for the weight value calculation is needed [9][10] in addition to the codebook size optimization. This paper proposes a soft voting method which considers relative distance. The distance to each codeword normalized by the distance to the nearest neighbor is used as a weight value for the soft voting. The proposed algorithm has been applied to human attribute analysis using surveillance video data, which has been a difficult problem in previous literatures [14][15]. The experimental results demonstrated that the performance for both gender classification and bag possession classification was 100% and its performance and robustness is comparable to that of the UCN model. The efficiency in terms of the parameter optimization was also investigated. Namely, the contributions of this paper are twofold. One is soft-voting-based feature representation considering the relative distance between the



(a) (b)

Fig. 1. (a) An example of the weight distribution of a codebook with a Gaussian kernel with three input feature vectors. (b) Generated feature vectors using UNC model with different βs and the proposed algorithm.

input and the codewords and the other is accurate human attribute analysis.

The rest of this paper is organized as follows. Section 2 briefly reviews the conventional algorithms on soft voting based BoF and describes the proposed algorithm. The experimental conditions and preprocessing are explained in Section 3. The experimental results are demonstrated in Section 4 followed by concluding remarks in Section 5.

2. PROPOSED ALGORITHM

In the hard voting approaches [1], a single codeword is assigned to each input feature vector. Then, a histogram of word frequencies that describes the probability density over the codewords is calculated as in the following equation.

$$CB(w)=\frac{1}{n}\sum_{i=1}^{n}\begin{cases}1 \; if \; w=\arg\min_{v\in V}\left(D(v,r_i)\right)\\ 0 \; otherwise\end{cases} \quad (1)$$

Here, n is the number of feature vectors, r_i is ith feature vector, and $D(v;r_i)$ is the distance between a codeword v and r_i. V is the codebook. L2 norm is employed for the distance metric in this paper.

The UNC [10] is a soft voting model which considers "relevancy" determined by the ratio of the kernel values for all codewords v in the codebook V:

$$UNC(w)=\frac{1}{n}\sum_{i=1}^{n}\frac{K_\sigma\left(D(w,r_i)\right)}{\sum_{j=1}^{|V|}K_\sigma\left(D(v_j,r_i)\right)}, \quad (2)$$

where K_σ is the Gaussian-like kernel defined as

$$K_\sigma(x)=\exp\left(-\beta x^2\right).$$

The β value needs to be decided by empirical study because the distance distribution depends on the feature representation algorithm.

The proposed method is soft voting which considers the relative distance to the nearest codeword:

$$REL(w)=\frac{1}{n}\sum_{i=1}^{n}\frac{\arg\min_{v\in V}D(v,r_i)}{D(w,r_i)}. \quad (3)$$

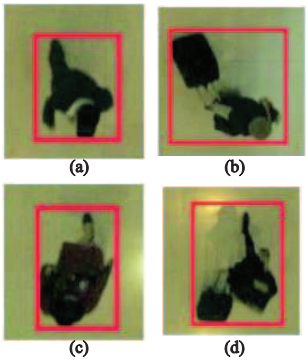


Fig. 2. Examples of pedestrians: (a) male w/o bag, (b) male with bag, (c) female w/o bag, (d) female with bag.

Table 1. Summary of pedestrians' attributes

	With bag	W/o bag	Total
Male	272	187	459
Female	179	150	329
Total	451	337	788

Whereas the UNC model considers the distribution of the distances from the input feature vector to all the codewords, the proposed model only cares the relative distance to the closest codeword. Since there is no magic number in eq. (3), one has to optimize only the codebook size. The conceptual difference of the proposed algorithm with the conventional methods is illustrated in Fig. 1. As shown in Fig. 1(b), if β in the UNC model is too large, it is nothing more than conventional BoF. On the other hand, if β is too small, the generated feature vector is flat and there is no discrimination power. In addition, the variance in each dimension is not always uniform, making it more difficult to decide the optimal β.

3. EXPERIMENTAL SETUP: APPLICATION TO HUMAN ATTRIBUTE ANALYSIS

The proposed algorithm was applied to the pedestrians' attribute analysis: gender classification and bag possession classification. Such human attribute analysis is important not only for safety and security purposes but also better services such as digital signage. One of the challenging aspects in our study is that we use only top-view images to protect privacy. Our previous works demonstrated that the classification performance was 69% and 76% respectively [14] by the p-type Fourier descriptor with Gaussian mixture model based bag-of-frames representation, and 96% and 97% respectively [15] by the hard voting using HoG features [16]. In the previous works, the performance was also sensitive to the codebook size (please see Section 4), therefore the feasibility of the algorithms was still questionable.

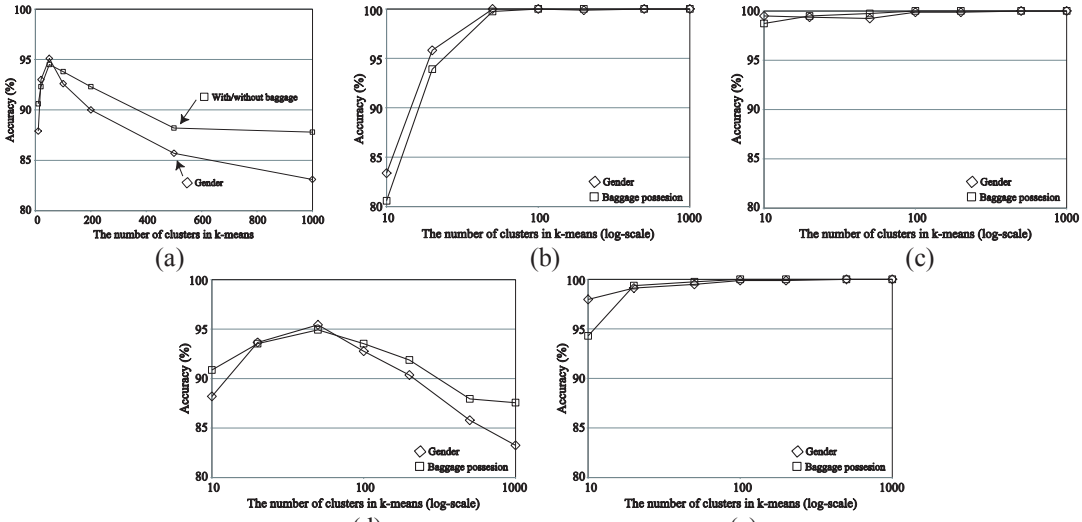


Fig. 3. Mean classification accuracy over the ten-cross validation. The codebook size was changed. (a) hard voting based [15], (b) UNC [10], β=0.001, (c) UNC, β=0.1, (d) UNC, β=10, (e) proposed.

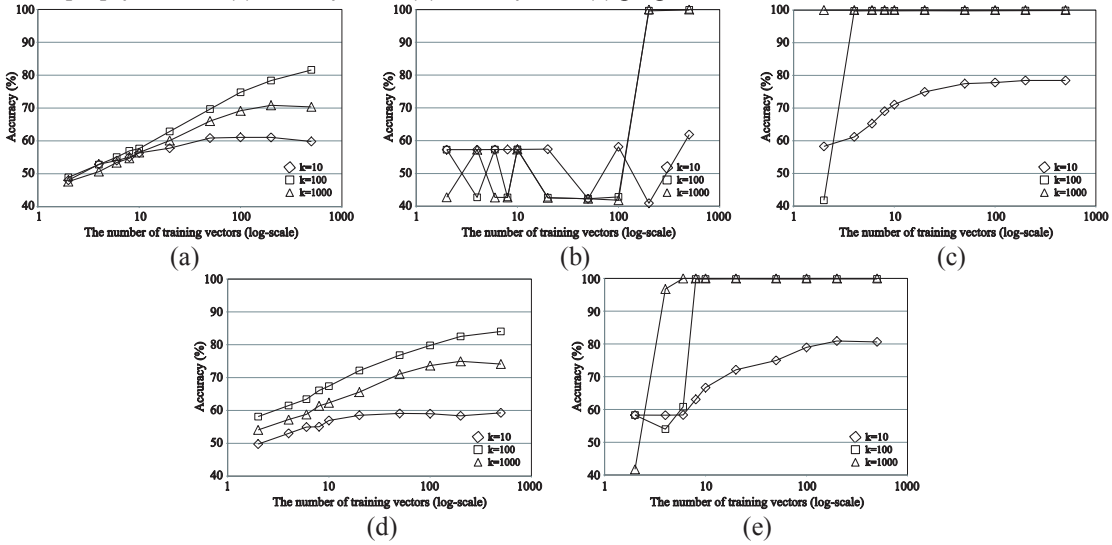


Fig. 4. Gender classification accuracy over the 100 random sampling as a function of the number of training samples; (a) hard voting based [15], (b) UNC [10], β=0.001, (c) UNC, β=0.1, (d) UNC, β=10, (e) proposed.

The test data were captured at Haneda airport (Tokyo International Airport), which is one of the top five busiest airports in the world. The camera was attached on the ceiling at 12m height and looked straightly down the floor. The view area was about 6m x 4.5m. The frame rate was 6.25 frames per second. The image size was 720 x 540. Sample images are shown in Fig. 2. The 60-minute data recorded in the afternoon on Sunday among one-month length data were used for the experiments. The number of detected pedestrians and their attributes are summarized in Table 1. The detected pedestrians' size is typically 100 x 100. We assume that detection, segmentation, and tracking are already done and data with occlusion or erroneously tracked pedestrians are eliminated in advance. For the details on detection, tracking, and feature extraction, please see [15].

The detected pedestrian regions were resized to 60 x 60 regardless of its original aspect ratio and HoG features were extracted. The k-means clustering was applied to a set of HoG features for all the frames of all the pedestrians to form bag-of-frames feature vectors. The underlying idea is representing the appearance of the pedestrian in each frame by the HoG feature and representing the difference how the pedestrian walks depending on gender and bag possession status by the bag-of-frames model. The classification was conducted using a support vector machine (SVM) [17].

4. EXPERIMENTAL RESULTS

Figure 3 shows the classification performance as a function of the codebook size. The graphs are the results of ten-cross

validation. For the UNC algorithm, β was also varied from 0.001 to 10. The parameters for the SVM classification were optimized for each case. In the hard voting approach [15], the performance was sensitive to the codebook size. Although the best accuracy was 96-97%, it went down to 80-90% as soon as the parameters shifted from their optimal values. When the UNC is employed, if the proper β is chosen (β=0.1, Fig. 3(c)), the accuracy is very high and stable. However, if β is not appropriate, the accuracy becomes sensitive to the codebook size (β=0.001, Fig. 3(b)) or becomes even worse than the hard voting case (β=10, Fig. 3(d)). On the other hand, the proposed method shows its robustness to the codebook size and the accuracy is very high.

The gender classification performance as a function of the number of training data is shown in Fig. 4. This experiment shows how much discriminative ability the extracted features have. The training samples were picked up randomly and the rest was used for testing. Half of the training samples were positive and the other half were negative. This procedure was repeated 100 times. In this experiment, the cost for the constraints violation was set to 128 and the other parameters were set as default. In the hard voting approach [15], the accuracy improvement over the number of the training samples is rather gentle, showing that the generated feature vectors are not discriminative and a lot of support vectors are needed. For the UNC model, only several samples are necessary when the codebook size and β are optimized. If the parameters are not optimal, it requires more than 100 training samples (β=0.001, Fig. 4(b)) or the performance is almost same as the hard voting (β=10, Fig. 4(d)). The proposed algorithm shows comparable performance to the optimized UNC model. Only about 10 samples are need if the codebook size is 100 or larger. In this point of view, our model is easier to optimize while achieving comparative performance with the UNC model. The performance for the baggage possession classification was almost the same and omitted because of the limited space.

5. CONCLUSIONS

In this paper, a soft-voting-based BoF representation has been proposed. The relative distance to that of the nearest-neighbor codeword was used for the soft voting. The proposed algorithm was compared with the hard voting method and one of the state-of-the-art soft voting methods called UNC model. The experimental results have shown that the proposed algorithm worked better and more robustly than the hard voting method and it was also comparable to the optimized UNC, which required time-consuming parameter optimization. The algorithm has been successfully applied to gender classification and baggage possession classification of the pedestrians in the airport.

6. ACKNOWLEDGEMENTS

The author would like to thank Prof. Kiyoharu Aizawa for valuable discussion. The author also would like to thank Mr. Tomoaki Matsunami for his contribution to data collection and labeling.

7. REFERENCES

[1] J. Sivic and A. Zisserman, “Video Google: a text retrieval approach to object matching in videos,” Proc. IEEE ICCV, pp. 1470-1477, 2003.

[2] T. Joachims, “Text categorization with support vector machines: learning with many relevant features,” Proc. ECML, pp. 137–142, 1998.

[3] D. Lowe, “Distinctive image features from scale-invariant keypoints,” IJCV, 60(2), pp. 91-110, 2004.

[4] S. Poullot, M. Crucianu, and O. Buisson, “Scalable mining of large video databases using copy detection,” Proc. ACMMM, pp. 61-70, 2008.

[5] D. Demirdjian and S. Wang, “Recognition of temporal events using multiscale bags of features,” Proc. IEEE Workshop on CIVI, pp.8-13, 2009.

[6] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” Proc. IEEE CVPR, pp.1-8, 2007.

[7] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: spatial pyramid matching for recognizing natural scene categories,” Proc. IEEE CVPR, pp. 2169-2178, 2006.

[8] L. Duan, D. Xu, I. W. Tsang, and J. Luo, “Visual event recognition in videos by learning from web data,” Proc. IEEE CVPR, pp. 1959-1966, 2010.

[9] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Lost in quantization: improving particular object retrieval in large scale image databases,” Proc. IEEE CVPR, pp. 1-8, 2008.

[10] J. van Gemert, J.-M. Geusebroek, C. Veenman, and A. Smeulders, “Kernel codebooks for scene categorization,” Proc. ECCV, Vol. 5304, pp. 696-709, 2008.

[11] Y. Huang, K. Huang, C. Wang, and T. Tan, “Exploring relations of visual codes for image classification,” Proc. IEEE CVPR, pp.1649-1656, 2011.

[12] J. Yang, K. Yu, Y. Gong, and T. Huang, “Linear spatial pyramid matching using sparse coding for image classification,” Proc. IEEE CVPR, pp. 1794-1801, 2009.

[13] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, “Locality-constrained linear coding for image classification,” Proc. IEEE CVPR, pp. 3360-3367, 2010.

[14] T. Yamasaki and T. Matsunami, “Pedestrian attribute analysis using a top-view camera in a public space,” Proc. MMM, LNCS 7131, pp. 541-550, 2012.

[15] T. Yamasaki and T. Matsunami, “Human attribute analysis using a top-view camera based on multi-stage classification," Proc 5th ACM/IEEE ICDSC2011, USB proceedings, 2011.

[16] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” Proc. IEEE CVPR, pp. 886-893, 2005.

[17] http://www.csie.ntu.edu.tw/~cjlin/libsvm/