Pedestrian Density Estimation by a Weighted Bag of Visual Words Model

Shilin Zhang and Xunyuan Zhang

***Abstract*—Pedestrian density estimation is very useful and important under transportation environment. In this paper, we present a novel weighting scheme of "bag of visual words model" for pedestrian density estimation, which characterizes both the weight and the relative spatial arrangement aspects of all visual words in depicting an image. We firstly analyze the visual words generation process. By counting the number of images through which each visual word is clustered and computing the cluster radius of each visual word, we can give each visual word a weight. Specially, the co-occurrences of visual words are computed with respect to spatial predicates over a hierarchical spatial partitioning of an image. The representation captures both the absolute and relative spatial arrangement of the words and, through the choice and combination of the predicates, can characterize a variety of spatial relationships. We validate this hypothesis using a challenging ground truth pedestrian dataset. Our approach is shown to result in higher classification accuracy rates than a non-weighting bag-of-visual-words approach. The time used to generate the visual words of our approach is only 1/20 to 1/30 compared to the time of the traditional image feature cluster process.**

***Index Terms*—Pedestrian detection, spatial relationship, visual words, SIFT.**

I. Introduction

Pedestrian density estimation in crowded scenes is a crucial component for a wide range of applications including transportation surveillance, group behavior modeling and crowd disaster prevention. The reliable person detection and tracking in crowds, however, is a challenging task due to view variations and varying density of people as well as the ambiguous appearance of body parts. High-density pedestrian crowds, such as illustrated in Fig. 1, present particular challenges for the difficulty of isolating individual person with standard low-level methods of background process typically applied in low-density transportation surveillance scenes.

Local invariant features have proven effective for a range of computer vision problems over the last decade. These features characterize the photometric aspects of an image while allowing for robustness against variations in illumination and noise. The geometric aspects of an image can further be characterized by considering the spatial arrangement of the local features. This paper proposes a novel image representation termed bag of visual words integrating weighting scheme and spatial pyramid co-occurrence, which characterizes both the photometric and geometric aspects of an image. Specifically, the co-occurrences of visual words quantized local invariant features are computed with respect to spatial predicates over a hierarchical spatial partitioning of an image. The local co-occurrences combined with the global partitioning allow the proposed approach to capture both the relative and absolute layout of an image. This is the main contribution of our work. Another salient aspect of the proposed approach is that it is general enough to characterize a variety of spatial arrangements.

Fig. 1. High density pedestrian crows.

In recent years significant progress has been made in the field of object detection and recognition [1]. While standard "scanning-window" methods attempt to localize objects independently, several recent approaches extend this work and exploit scene context as well as relations among objects for improved object detection [2]. Related ideas have been investigated for human motion analysis where incorporating scene-level and behavioral factors effecting the spatial arrangement and movement of people have been shown beneficial for achieving improved detection and tracking accuracy. Examples of explored cues include: the destination of a pedestrian within the scene, repulsion from near-by agents due to the preservation of personal space and social grouping behavior, as well as the speed of an agent in the group

The rest of the paper is organized as follows. We put our method in the context of related work in Section II. Section III describes our model. The proposed model is experimentally evaluated in Section IV. In the last section we concluded our paper.

II. Related Works

The broader context of our work is bag-of-visual-words (BOVW) [3] approaches to image classification. These approaches quantize local invariant image descriptors using a

Manuscript received July 23, 2014; revised December 20, 2014. This work was supported in part by the National Natural Science Foundation of China under Grant No.61403004 and the Beijing Higher Education Young Elite Teacher Project under Grant No.YETP1421.

The authors are with North China University of Technology, Beijing, 100144, China (e-mail: zhangshilin@126.com).

visual dictionary typically constructed through k-means clustering. The set of visual words is then used to represent an image regardless of their spatial arrangement similar to how documents can be represented as an unordered set of words in text analysis. The quantization of the often high-dimensional local descriptors provides two important merits: it provides further invariance to photometric image transformations, and it allows compact representation of the image such as through a histogram of visual word counts and/or efficient indexing through inverted files. The size of the visual dictionary used to quantize the descriptors controls the tradeoff between invariance/efficiency and discriminability.

Lazebnik *et al.* was one of the first works to address the lack of spatial information in the BOVW representation. Their spatial pyramid representation was motivated by earlier work termed pyramid matching. The fundamental idea behind pyramid matching is to partition the feature space into a sequence of increasingly coarser grids and then compute a weighted sum over the number of matches that occur at each level of resolution. Two points are considered to match if they fall into the same grid cell and matched points at finer resolutions are given more weight than those at coarser resolutions. The spatial pyramid representation of Lazebnik *et al.* applies this approach in the two-dimensional image space instead of the feature space; it finds approximate spatial correspondences between sets of visual words in two images.

The spatial pyramid representation characterizes the absolute location of the visual words in an image. Saverese *et al.* [4] propose a model which instead characterizes the relative locations. Motivated by earlier work [5] on using correlograms of quantized colors for indexing and classifying images [5], [6], they use correlograms of visual words to model the spatial correlations between quantized local descriptors. The correlograms are three dimensional structures which in essence record the number of times two visual words appear at a particular distance from each other. Correlogram elements corresponding to a particular pair of words are quantized to form correlations. Finally, images are represented as histograms of correlations and classified using nearest neighbor search against exemplar images. One challenge of this approach is that the quantization of correlograms to correlations can discard the identities of associated visual word pairs and thus may diminish the discriminability of the local image features.

Some methods also characterize the relative locations of visual words. Their proximity distribution representation is a three dimensional structure which records the number of times a visual word appears within a particular number of nearest neighbors of another word. It thus captures the distances between words based on ranking and not absolute units. A corresponding proximity distribution kernel is used for classification in a support vector machine (SVM) framework. However, since proximity kernels are applied to the whole image, distinctive local spatial distributions of visual words may be overshadowed by global distributions.

Collectiveness measuring methods describe the degree of individuals acting as a union in collective motion. It depends on multiple factors, such as the decision making of individuals and crowd density. Quantitatively measuring this universal property is important in order to understand the general principles of various crowd behaviors. In this area, some works [7]-[9] achieved fruitful outcomes.

But most existing crowd surveillance technologies lack universal classification methods with which to characterize pedestrian density. Existing works simply measured the average velocity of all the individuals to indicate the collectiveness of the whole crowd, which is neither accurate nor robust.

III. Method Presentation

We assume each image I contains a set of *N* visual words c_i at pixel locations (x_i, y_i) where each word has been assigned a discrete label $c_i \in [1...M]$ from a visual dictionary containing M words. The locations of the visual words could either be determined using a dense grid or an interest-point detector. We use Lowe's scale invariant feature transform detector [8] in the experiments below. Local invariant features are extracted at these locations and quantized into a discrete set of labels using a codebook typically generated by applying k-means clustering to a large, random set of features. We also use Lowe's SIFT descriptor [8] in the experiments below.

A. Bag of Visual Words Representation

The non-spatial BOVW representation simply records the visual word occurrences in an image. It is typically represented as a histogram:

$$BOVW = [t_1, t_2, t_3 ..... t_M] \quad (1)$$

where t_M is the number of occurrences of visual word *m*. To account for the difference in the number of visual words between images, the BOVW histogram is typically normalized to have unit L1 norm.

$$I(H1_l, H2_l) = \sum_{k=1}^{D} \sum_{m=1}^{M} \min(H1_l(k,m), H2_l(k,m)). \quad (2)$$

A BOVW representation can be used in kernel based learning algorithms, such as non-linear support vector machines, by computing the intersection between histograms. Given BOVW1 and BOVW2 corresponding to two images, the BOVW kernel is computed as:

$$K_{BOVW}(BOVW1, BOVW2) = \sum_{m=1}^{M} \min(BOVW1(m), BOVW2(m)). \quad (3)$$

The intersection kernel is a Mercer kernel which guarantees an optimal solution to kernel-based algorithms based on convex optimization such as nonlinear SVMs.

Our contribution propose the question, "do the most frequent visual words function like stop words"? We approach this problem by examining the classification performance using vocabularies without the most frequent visual words. After removal the most frequent visual words, the classification rate is declined. So the problem is not so simple. We examine the image feature process, and find an important phenomenon. The good visual words have more images when they were generated during the cluster process and the image radius is smaller than the noise words'. As can be seen in Fig. 2, the big circle is a word which radius is

longer than the other ones'. If a smaller word contains more images than the other ones', then it is more like a good word. Through the above standard, we can filter out some noise words. We put out the following standard:

The num_i means the number of images that the i_{th} word has and the r_i is the radius of the i_{th} word. If the above formula qualified, then we filter out the word.

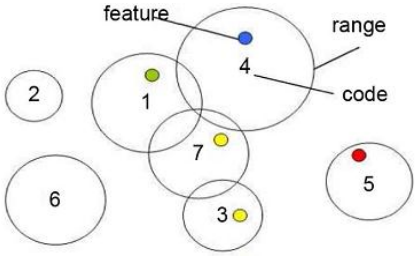

Fig. 2. Visual Words' weighing scheme.

B. Spatial Pyramid

The spatial pyramid representation of Lazebnik *et al.* [1] partitions an image into a sequence of spatial grids at resolutions 0,...,L such that the grid at level l has 2^l cells along each dimension for a total of D =4^l cells. A BOVW histogram is then computed separately for each cell in the multi-resolution grid. $H_l(k, m)$ is the count of visual word m contained in grid cell k at level l. This representation is summarized in Fig 3.

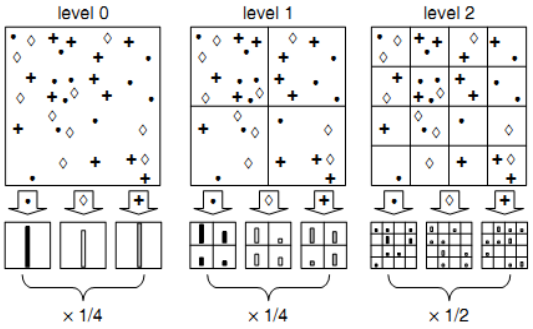

Fig. 3. A three-level spatial pyramid.

A spatial pyramid match kernel (SPMK) is derived as follows. Let H_{1l} and H_{2l} be the histograms of two images at resolution l. Then, the number of matches at level l is computed as the histogram intersection:

$$\mathrm{K}_{SCK} = (\mathrm{VWCM1}_{\rho}, \mathrm{VWCM\,2}_{\rho}) = \sum_{u,v \in M} \min(\mathrm{VWCM1}_{\rho}(\mathrm{u,v}), \mathrm{VWCM\,2}_{\rho}(\mathrm{u,v})) \quad (4)$$

Abbreviate $I(H1_l, H2_l)$ to I_l. Since the number of matches at level l includes all matches at the finer level l+1, the number of new matches found at level l is $I_l - I_{l+1}$ for l =0,..., L−1. Further, the weight associated with level l is set to 1/ (2^{L-l}) which is inversely proportional to the cell size and thus penalizes matches found in larger cells. Finally, the SPMK for two images is given by:

$$K_{SPMK} = I_L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}} (I_l - I_{l+1}) \quad (5)$$

The SPMK is a Mercer kernel [1].

C. Spatial Co-occurrence

Spatial co-occurrence of visual words is motivated by Yang *et al.*'s seminal work [9] on gray level co-occurrence matrices which is some of the work on image texture. A GLCM provides a straightforward way to characterize the spatial dependence of pixel values in an image. We extend this to the spatial dependence of visual words.

Formally, given an image I containing a set of N visual words c_i at pixel locations (x_i, y_i) and a binary spatial predicate ρ where $c_i \rho c_j \in$ {T,F}, we define the visual word co-occurrence matrix (VWCM) as:

$$VMCW_{\rho}(u,v) = \| (c_i, c_j) \,|\, (c_i = u) \cap (c_j = v) \cap c_i \rho c_j \| \quad (6)$$

That is, the VWCM is a count of the number of times two visual words satisfy the spatial constrains. The choice of the predicate ρ determines the nature of the spatial dependencies. This framework can support a variety of dependencies such as the two visual words needing to be within a certain distance of each other, to have the same orientation, etc. We describe a number of predicates in the experiments section.

We derive a spatial co-occurrence kernel (SCK) as follows. Given two visual co-occurrence matrices $\mathrm{VWCM1}_{\rho}$ and $\mathrm{VWCM2}_{\rho}$ corresponding to two images, the SCK is computed as the intersection between the matrices:

$$K_{SPCK}(\mathrm{VWCM1}^l_{\rho}, \mathrm{VWCM\,2}^l_{\rho}) = \sum_{l=1}^{L} w_l \sum_{k=1} \sum_{u,v \in M} \min(\mathrm{VWCM1}\rho(\mathrm{u,v}), \mathrm{VWCM\,2}\rho) \quad (7)$$

where the weights w_l are chosen so that the sum of intersections has the same maximum achievable value for each level; e.g., $w_l = 1/4^l$. As a sum of intersections, the SPCK is a Mercer kernel.

To account for differences in the number of pairs of code words satisfying the spatial predicate between images, the matrices are normalized to have an L1 norm of one. The SCK, as an intersection of two multidimensional counts, is also Mercer kernel. A spatial pyramid co-occurrence kernel (SPCK) corresponding to the spatial pyramid co-occurrences for two images $\mathrm{VWCM1}_{\rho}$ and $\mathrm{VWCM2}_{\rho}$ is then computed.

IV. Experiments

We evaluate our proposed weighting scheme integrated spatial pyramid co-occurrence representation on PASCAL VOC 2007 and a pedestrian dataset.

A. Weighting Scheme and Spatial Predicate

We revised the traditional BOVW method by introducing a new visual word generation during the traditional image feature cluster process. In our scheme, we cluster the visual words on every single class of images instead of on all the images. And we remove the stop words by our standard formula.

We consider two types of spatial predicates: proximity predicates which characterize the distance between pairs of visual words, and orientation predicates which characterize the relative orientations of pairs of visual words. Since our primary goal is to analyze overhead imagery, and according to

Tobler's first law of geography, all things on the surface of the earth are related but nearby things are more related than distant things.

The SIFT detector provides the orientation of the interest points used to derive the visual words. We postulate that these orientations are indicative of the local shape of image regions and thus derive orientation predicates which consider the relative orientations of pairs of visual words.

B. Experiment Setup

The PASCAL corpus was used for the PASCAL Visual Object Classes Challenge. It has 10000 labeled images from multiple sources. PASCAL images are less noisy and cluttered. We choose it since it has been frequently used as a benchmark for evaluating key point-based features. Using a second and very different corpus also makes the conclusions in this paper more convincing.

The classification is conducted in a "one-against-all" manner. Using the Support Vector Machines (SVM), we build 20 binary classifiers for the 20 semantic concepts in Pascal VOC dataset, where each classifier is for determining the presence of a specific concept or object. We use average precision (AP) to evaluate the result of a single classifier, and mean average precision (MAP) to aggregate the performance of multiple classifiers. Note that the state-of-the-art classification performance on Pascal VOC is about 0.7 in MAP since the classification is very difficult on this challenging corpus.

C. Experiment Results

We compare our methods with the traditional BOVW, Spatial Pyramid Matching Method, and our weighting and integrated Spatial Visual words Co-occurrence Method. The result is shown in Fig. 4.

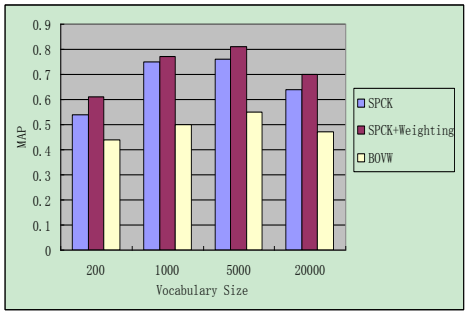

Fig. 4. Comparison of 3 methods.

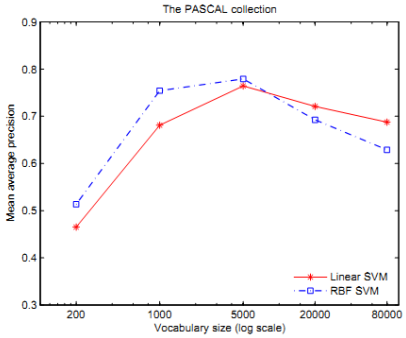

Fig. 5. The influence of different SVM kernel.

Apparently, when the vocabulary size is 5000, the three methods' result is better. Our method outperforms the other two methods by a large margin. We also compared different SVM kernel's influences on the classification rates. It is depicted in Fig. 5.

As for the weighing scheme, we used the Term Frequency (T), Term Frequency with Inversed Document Frequency (TI), and our weighting scheme (WTI). Our weighing scheme achieves the best accuracy when compared with the other two weighting schemes.

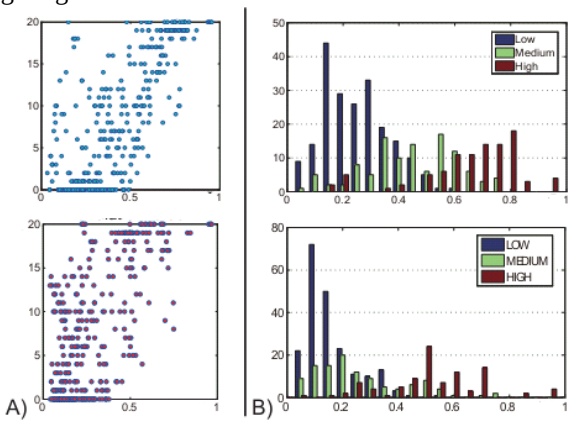

Fig. 6. Pedestrian density classification comparisons.

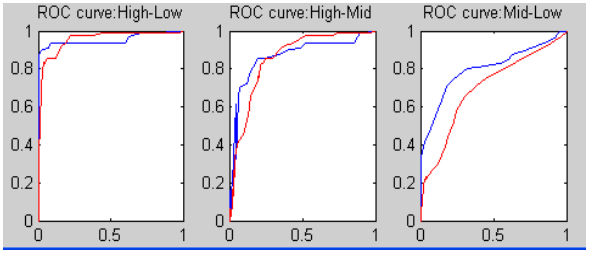

Fig. 7. ROC comparisons.

Fig. 8a. Pedestrian density estimation.

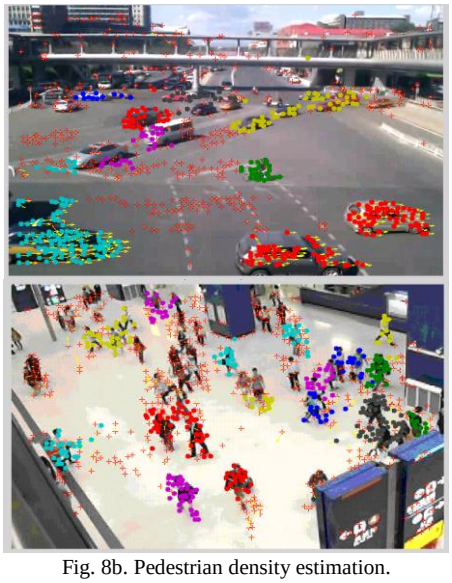

Fig. 8b. Pedestrian density estimation.

The second is the classification accuracy based on the collectiveness density. We divide all the videos into three categories by majority voting of subjects' rating, and then evaluate how the proposed collectiveness descriptor can classify them. Fig. 6 and Fig. 7 plot the ROC curves and the best accuracies which can be achieved with all the possible decision boundaries for binary classification of high and low high and medium, and medium and low categories .It indicates our collectiveness descriptor can delicately measure the density of pedestrian crowds. Fig. 8 shows the pedestrian density classification results.

V. Conclusion

Bag-of-visual-word is an effective image representation in the classification task, but various representation choices w.r.t its dimension, weighting, and word selection has not been thoroughly examined. In this paper, we have applied techniques used in text categorization, including term weighting, stop word removal, feature selection, to generate various visual-word representations, and studied their impact to classification performance on the PASCAL collections. This study provides an empirical basis for designing visual-word representation that is likely to produce superior classification performance.

The enhanced pedestrian classification methods can be applied to measure crowd density under complex transportation environments, which is difficult previously because universal properties of crowd systems could not be well quantitatively measured. This paper is an important starting point in these exciting research directions.

References

[1] Lazebnik, C. Schmid, and J. Ponce, "Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR 2006*, New York, NY, USA, pp. 2169-2178.

[2] J. Zhang, M. Marszalek, and S. Lazebnik, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int. J. Comput*, vol. 2, pp. 213-220, 2007.

[3] V. Bettadapura, G. Schindler, and T. Ploetz, "Augmenting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition," in *Proc. CVPR 2013*, Portland, Oregon, USA.

[4] S. Savarese, J. Winn, and A. Criminisi , "Discriminative object class models of appearance and shape by correlations," in *Proc. CVPR 2006*, New York, NY ,USA.

[5] D. Liu, G. Hua, P. Viola, and T. Chen, "Integrated feature selection and higher-order spatial feature extraction for object categorization," in *Proc. CVPR 2008*, Anchorage, Alaska, USA.

[6] D. Singaraju and R. Vidal, "Using global bag of features models in random fields for joint categorization and segmentation of Obj," in *Proc. CVPR 2011*, Colorado, USA.

[7] W. Zhang, D. Zhao, and X. Wang, "Agglomerative clustering via maximum incremental path integral," *Pattern Recognition*, 2013.

[8] L. Kratz and K. Nishino, "Pedestrian efficiency in crowded scenes," in *Proc. ECCV*, 2012.

[9] B. Zhou, X. Wang, and X. Tang. "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *Proc. CVPR 2012*, USA.

Shilin Zhang was born in Shandong province of China in 1980. He graduated from Chinese Academy of Sciences and received his PhD degree in computer science in 2012 in China. He is now associated with North China University of Technology and his current research interests include image processing, pattern recognition and so on. Dr. Zhang is a member of Chinese Association of Automation.

Xunyuan Zhang was born in Shandong province of China in 1989. He graduated from Qufu Normal University of China and received his bachelor degree in automation science in 2011.

He is now pursuing his master degree in North China University of Technology, and his research interests include image processing, pattern recognition and so on.