International Journal of Multimedia and Ubiquitous Engineering
Vol. 10, No. 3 (2015), pp. 99-108
http://dx.doi.org/10.14257/ijmue.2015.10.3.10

Crowded Pedestrian Detection and Density Estimation by Visual Words Analysis

Shilin Zhang and Xunyuan Zhang

North China University of Technology
zhangshilin@126.com

Abstract

Crowded pedestrian detection and density estimation are very useful and important under transportation environment. In this paper, we present a novel method for crowded pedestrian detection and density estimation through a weighting scheme of bag of visual words model which characterizes both the weight and the relative spatial arrangement aspects of visual words in depicting an image. Firstly, we analyze the visual words generation process. We give each visual word a weight by counting the number of images through which each visual word is clustered and computing the cluster radius of each visual word. To be more specifically, the co-occurrences of visual words are computed with respect to spatial predicates over a hierarchical spatial partitioning of an image. We validate this method using a challenging ground truth pedestrian dataset Pascal VOC 2007. Our approach is shown to be more accuracy than a non-weighting bag-of-visual-words one. The algorithm's cost is also more efficient than the competing pairs.

Keywords: *crowded pedestrian detection, pedestrian density estimation, automotive safety*

1. Introduction

Significant progress has been made in crowed pedestrian detection and density estimation in the last few decades. It has been widely applied to automotive safety, robotics and intelligent video surveillance. 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 Figure 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.

SIFT is a local invariant feature and has been proven to be effective for a range of computer vision problems over the last few years. These features characterize the photometric aspects of an image 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.

ISSN: 1975-0080 IJMUE

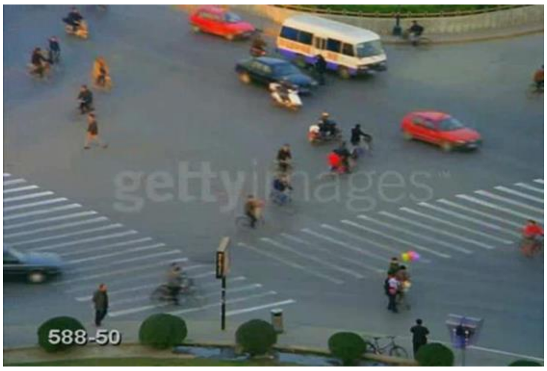

Figure 1. Pedestrian Crowd in Crossing Road

Thanks to the significant progress that has been made in the field of object detection and recognition [1], we borrow the idea, namely, the visual words model, for crowded pedestrian detection and density estimation. While traditional 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 pedestrian detection [2]. Related works have been investigated human motion analysis which incorporates scene-level and behavioral factors for pedestrian detection. The spatial arrangement and people movements have been shown beneficial for achieving improved object 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 paper is organized as follows. We review the context of related work in Section 2. In Section 3 we describe our model in detail. The proposed model is experimentally evaluated and remarked in Section 4. In the last section we concluded our paper.

2. Related Works

The progress on object detection has been achieved by the investigation on classification approaches, features and articulation handling approaches. The broader context of our work is pedestrian detection and bag-of-visual-words [3] approaches. Bag-of-visual-words quantize local invariant image descriptors using a 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 efficiency and discriminability.

The 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 model defines the absolute location of the visual words in an image. The work [4] proposes a model which instead characterizes the relative locations. Motivated by earlier work on using relate histogram of quantized colors for indexing and classifying images [6], they use relate histogram of visual words to model the spatial correlations between quantized local descriptors. The relate histogram are three dimensional structures which in essence record the number of times two visual words appear at a particular distance from each other. Relate histogram 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 relate histogram 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 framework. However, since proximity kernels are applied to the whole image, distinctive local spatial distributions of visual words may be overshadowed by global distributions.

To detect crowded persons and depict the pedestrian density in transport environment, 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, 8] 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 pedestrians to indicate the collectiveness of the whole crowd, which is neither accurate nor robust.

3. Pedestrian Density Detection and Estimation

Each image I contains a set of 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 visual 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 in the experiments below.

3.1. Feature Representation

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

$$BOVW = [t_1, t_2, t_3 \ldots\ldots t_M] \tag{1}$$

The 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 bag-of-visual-words histogram is typically normalized to have unit L1 norm.

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

The visual words 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:

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

The 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 inspect 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 Figure 2, the big circle is a visual 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 better visual word. Through the above standard, we can filter out some noise words. We put out the following measure 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.

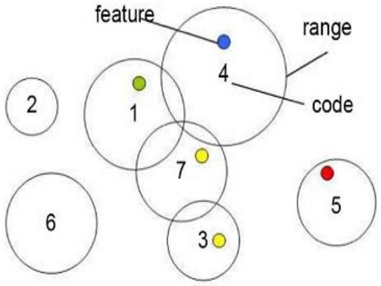

Figure 2. Visual Words' Feature Representation

3.2. Visual Words Location Presentation

The visual words location presentation, *i.e.*, the spatial pyramid model of Lazebnik *et al.*, 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 Figure 3.

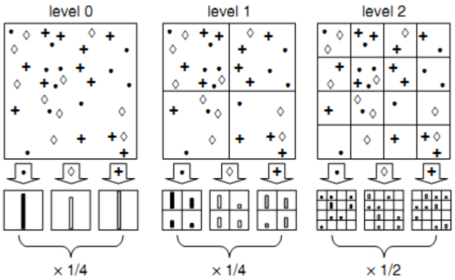

Figure 3. A Three-level Spatial Pyramid

The spatial pyramid match kernel is derived as follows. Let $H1_l$ and $H2_l$ be the histograms of two images at resolution l. Then, the number of matches at level l is computed as the histogram intersection:

(4)

We abbreviate $I(H1_l, H2_l)$ to I_l. As 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.

3.3. Spatial Definition

Spatial definition of visual words is motivated by Yang et al.'s seminal work 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 ci 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 as:

(5)

The VWCM is the 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 as follows. Given two visual co-occurrence matrices $VWCM^1_\rho$ and $VWCM^2_\rho$ corresponding to two images, the SCK is computed as the intersection between the matrices:

(6)

The weight w_l is 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 corresponding to the spatial pyramid co-occurrences for two images $VWCM^1_\rho$ and $VWCM^2_\rho$ is then computed.

4. Experiments and Comparison

We extensively evaluated our proposed weighting scheme integrated spatial pyramid co-occurrence representation on PASCAL VOC 2007.

4.1. Weighting Scheme and Spatial Predicate

During the traditional image feature cluster process, we revised the traditional visual words method by introducing a new visual word generation procedure. In our scheme, we cluster the visual words on every single class of images instead of on all the images. Then we remove the stop words by our measures of standard formula presented above.

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 are considered. 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.

4.2. Experiment Setup

The PASCAL VOC 2007 image database is 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 jobs are conducted in a one-against-all manner. Using the Support Vector Machines, we build binary classifiers for the semantic concepts in Pascal VOC dataset, where each classifier is for determining the presence of a specific concept or object. In our experiment, the person corpus is within our comparison. We use average precision to evaluate the result of a single classifier, and mean average precision to aggregate the performance of multiple classifiers. Note that the state-of-the-art classification performance on Pascal VOC 2007 is about 0.7 in MAP since the classification is difficult on this challenging database.

4.3. 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 Figure 4.

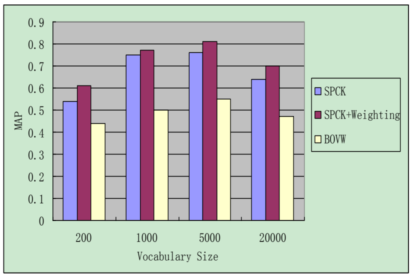

Figure 4. Comparison of Detection Methods

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 Figure 5.

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

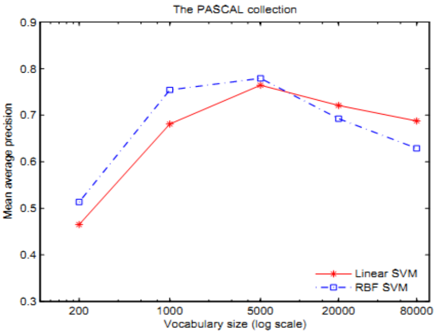

Figure 5. The Impact of Different SVM Kernel

The second one is the classification accuracy based on the collectiveness density. We divide all the videos into three categories by majority voting of subjects, and then evaluate how the proposed collectiveness descriptor can classify them. Figure 6 plots 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 pedestrian density descriptor can delicately measure the density of pedestrian crowds. Figure 7 shows the extended crowd density classification results in Beijing crossing street.

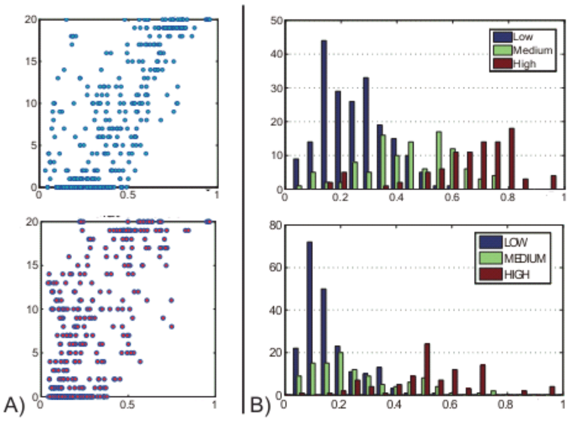

Figure 6. Pedestrian Density Classification Comparisons

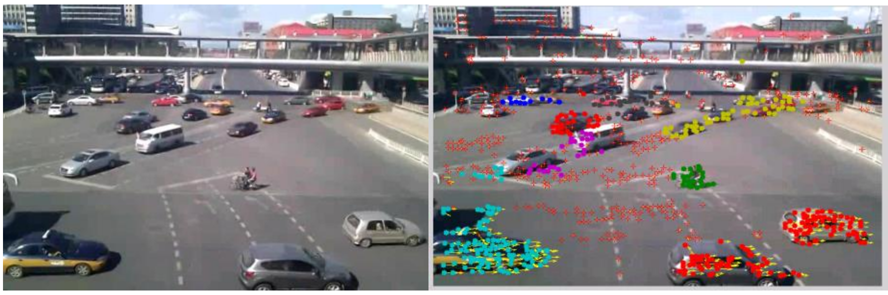

Figure 7. Density Estimation

5. Conclusion

In this paper, we propose a new probabilistic framework for crowded pedestrian detection and density estimation. 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 VOC 2007 image 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.

Acknowledgement

This work was supported 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.

References

[1] C. Schmid, J. Ponce and Lazebnik, “Spatial pyramid matching for recognizing natural scene categories”, Proceedings of computer vision and pattern recognition, **(2006)**; New York, USA.

[2] J. Zhang, M. Marszalek and S. Lazebnik, “Local features and kernels for classification of texture and object categories: A comprehensive study”, International Journal of Computing, vol. 2, no. 23, **(2007)**.

[3] V. Bettadapura, G. Schindler and T. Ploetz, “Augmenting Bag-of-Words: Data-Driven Discovery of Temporal and Structural Information for Activity Recognition”, Proceedings of computer vision and pattern recognition, **(2013)**; Portland, Oregon USA.

[4] S. Savarese, J. Winn, and A. Criminisi, “Discriminative object class models of appearance and shape by correlations”, Proceedings of computer vision and pattern recognition, **(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”, Proceedings of computer vision and pattern recognition**, (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”, Proceedings of computer vision and pattern recognition, **(2011)**; Colorado, USA.

[7] W. Zhang, D. Zhao and X. Wang, “Agglomerative clustering via maximum incremental path integral”, Pattern Recognition, vol. 5, no. 21, **(2013)**.

[8] B. Zhou, X Wang and X. Tan, “Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents”, Proceedings of computer vision and pattern recognition, **(2012)**; Portland, USA.

Authors

Shilin Zhang, he was born in Shandong province of China on 1980 and graduated from Chinese Academy of Sciences and received his PhD degree in computer science on 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. He is a member of Chinese Association of Automation.

Xunyuan Zhang, he was born in Shandong province of China on 1989 and graduated from Qufu normal University of China and received his bachelor degree in Automation Science on 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.