Object Discovery by Clustering Correlated Visual Word Sets

Gibran Fuentes Pineda, Hisashi Koga and Toshinori Watanabe
Graduate School of Information Systems
The University of Electro-Communications
Tokyo, Japan
Email: {gibranfp,koga,watanabe}@sd.is.uec.ac.jp

***Abstract*—This paper presents a novel approach to discovering particular objects from a set of unannotated images. We aim to find discriminative feature sets that can effectively represent particular object classes (as opposed to object categories). We achieve this by mining correlated visual word sets from the bag-of-features model. Specifically, we consider that a visual word set belongs to the same object class if all its visual words consistently occur together in the same image. To efficiently find such sets we apply Min-LSH to the occurrence vector of the each visual word. An agglomerative hierarchical clustering is further performed to eliminate redundancy and obtain more representative sets. We also propose a simple and efficient strategy for quantizing the feature descriptors based on locality-sensitive hashing. By experiment, we show that our approach can efficiently discover objects against cluster and slight viewpoint variations.**

***Keywords*-hashing; Min-LSH; object discovery; correlated itemset mining;**

I. Introduction

Modern feature detectors and descriptors have boosted the development of effective models to represent collections of images and videos. In particular, the bag-of-features approach [1] has been widely adopted due to its simplicity, flexibility and excellent performance. Many methods based on this approach can efficiently classify similar scenes robustly against occlusion, illumination and viewpoint changes. However, their ability to classify the same object class from cluttered scenes is limited because they judge the similarity among images globally. To overcome this problem, methods based on unsupervised object discovery have been recently proposed.

Latent variable models such as PLSA and LDA have been successfully applied to the discovery of particular object classes [2], [3] as well as object categories [4]. These models represent each image as a mixture of K topics where each topic corresponds to a single object class. The main limitation of such models is that the number of topics K must be given by the user and it is not always obvious. This is particularly true for large and diverse sets of images where the number of classes can be tied to subjective judgments. In addition, as it is very time consuming to estimate the model parameters, latent variable models are not easily scalable to large databases.

In this paper, we propose a novel approach to automatically discovering particular object classes (as opposed to object categories) in cluttered scenes. The basic idea is to find discriminative visual word sets that can effectively represent particular object classes. We achieve this by mining sets of visual words with highly correlated occurrences from the bag-of-features model. To efficiently mine such correlated sets, we rely on Min-LSH. A pattern summarization based on hierarchical agglomerative clustering is further used to eliminate redundancy and obtain more representative sets. We show that the discovered visual word sets can effectively represent a particular object class for recognition. Different from [2], [4], our approach works despite the number of object classes is not given. In a previous paper [5], we also proposed a hash-based approach that can efficiently discover objects from images without supervision. However, [5] assumes that objects are isolated from the background and that they don't overlap each other and consequently can not be applied to cluttered scenes. In this work we get rid of the above strong assumptions.

Chum et al. [6] have previously used Min-LSH to image retrieval based on the bag-of-features model. However, this approach judges the similarity between images globally and hence is not suitable for object discovery in cluttered scenes. Geometric min-hash [7] partially addresses this problem by considering local spatial information for computing the hash values but it can only discover small objects.

The remainder of the paper is organized as follows. Section II reviews the hashing schemes used for similarity searching. We then describe the bag-of-features approach in Sect. III. Our method is introduced in Sect. IV. Sect. V reports the experimental results and finally, Sect. VI gives the conclusions and future plans.

II. Similarity Search by Hashing

To cope with the high dimensionality of the image representation and scale to large image databases, we rely on two hashing schemes for efficient similarity search in high dimensions, namely *locality-sensitive hashing* (LSH) [8] and *Min-LSH* [9]. Next, we briefly review these hashing schemes.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.189

IEEE computer society

A. Locality Sensitive Hashing

LSH is a randomized method for approximate similarity search in high dimensional spaces. Our approach relies on the LSH algorithm proposed by Gionis et al. [8]. We describe this algorithm from now on. Let P be a set of points in a d-dimensional space and C be the maximum coordinate value of any point in P. Every $p \in P$ is transformed to a Cd-dimensional vector by concatenating unary expressions for every coordinate, i.e.,

$$f(p) = \text{Unary}(x_1)\text{Unary}(x_2) \cdots \text{Unary}(x_d), \qquad (1)$$

where Unary(x) is a sequence of x ones followed by $C - x$ zeros. A hash function is computed by picking 'up k bits uniformly at random from these Cd bits and concatenating them. This corresponds to partitioning the d-dimensional space into cells of different sizes by k hyperplanes so that close points will have the same hash value with high probability. As k becomes large, points that are far from each other are less likely to take the same hash value because the size of generated cells becomes small. By contrast, depending on the result of space division, close points may take different hash values. To exclude this failure, LSH defines multiple l hash functions $h_1, h_2, \cdots h_l$ expecting that close points will take the same hash value at least for one hash function.

B. Min-LSH

Min-LSH is an efficient technique to find similar items from binary dyadic data (e.g. document-term matrix). Let $X_1, X_2, \ldots, X_N$ be a set of N binary vectors with coordinates $x_1, x_2, \ldots, x_M$. Min-LSH generates a random permutation π of the coordinates $x_1, x_2, \ldots x_M$ and assigns to each binary vector $X_i, i = 1, \ldots, N$ its minimum nonzero coordinate over the permutation, i.e.,

$$h(X_i) = min(\pi(X_i)), \qquad (2)$$

The probability that two binary vectors X_i, X_j have the same min-hash value (i.e. their first nonzero coordinate over the permutation is the same) is equal to their Jaccard coefficient, that is,

$$P[h(X_i) = h(X_j)] = \frac{\mid X_i \cap X_j \mid}{\mid X_i \cup X_j \mid} = sim(X_i, X_j). \qquad (3)$$

where $P[E]$ denotes the probability of the event E. Thus, similar binary vectors have high probability of having the same min-hash value while dissimilar ones have low probability. To estimate the degree of similarity, k different permutations $\pi_1, \ldots, \pi_k$ are generated and k min-hash values $min\{\pi_1(X_i)\}, \ldots, min\{\pi_k(X_i)\}$ are computed for each binary vector.

To retrieve similar binary vectors, the min-hash values are grouped into l tuples $g_1, \ldots, g_l$ of r different min-hash values. Two binary vectors with an identical tuple are regarded as similar. Finding binary vectors with an identical tuple can be easily implemented by hashing: l hash tables are defined (one for each tuple) and binary vectors with identical tuple are mapped to the same hash bucket. As they are expected to agree in several min-hash values, two highly similar binary vectors X_i, X_j have high probability of being stored in the same hash bucket at least in one hash table. The above probability is expressed by the following equation.

$$P_{collision}[X_i, X_j] = 1 - (1 - sim(X_i, X_j)^r)^l. \qquad (4)$$

The selection of the values of r and l is a trade-off between recall and precision.

III. Bag-of-Features

In the following, we present an overview of the steps to generate the bag-of-features representation of a set of images $\Sigma = \{I_1, I_2, \ldots, I_N\}$ and specify the techniques used in each step.

- Each image is described by a set of 128-dimensional SIFT vectors computed from affine covariant regions. Specifically, we extract regions by using the Maximally Stable Extremal Region (MSER) [10] which obtained the highest scores among different types of affine regions [11].
- A vocabulary of *visual words* $\mathcal{V} = \{v_1, \ldots, v_M\}$ is constructed by clustering the SIFT descriptors; each cluster center represents a visual word. This is typically done by clustering the descriptors of a random subset of images with k-means. However, k-means becomes slow for large values of k. Here we propose a more efficient algorithm based on LSH. Given a set of descriptors D_{train}, the idea is to define a voting approach for each descriptor and select as centers the descriptors with the greatest number of votes. For each descriptor $d \in D_{train}$, the voting approach keeps track of the number of ε-nearest neighbors of d. Then the descriptor with greatest number of votes is regarded as a cluster center and all its ε-near neighbors are assigned to it; this process is repeated until all the descriptors are assigned to a cluster. We accelerate the search for near descriptors by resorting to LSH.
- Each descriptor is assigned to the visual word with the nearest center. We search for the nearest center by using LSH. Then, we represent each image as a set of visual words (frequencies are not taken into account).
- Very rare and very common visual words are discarded from the visual vocabulary: visual words that occur in more than 50% or less than 0.2% of the images in the database.

- The set of images is represented by an $N \times M$ binary co-occurrence table T whose elements $t_{ij} \in \{0,1\}$ record the absence or presence of the visual word v_i in the image I_j. Each column j of T is a binary vector $\hat{I_j}$ which defines the bag-of-features representation of the image I_j. Conversely, each row i of T is a binary vector $\hat{v_i}$ which presents the images where the visual word v_i occurs (we call these vectors *occurrence vectors*).

IV. Object Discovery

Given a co-occurrence table T, we formulate the problem of object discovery as a data mining problem. We consider that the occurrence vectors of visual words that belong to the same category are highly correlated (this is true for discriminative visual words). Then, our object discovery approach is divided in two stages: 1) correlated visual word set mining based on Min-LSH and 2) pattern summarization by agglomerative hierarchical clustering. In the following, we describe each of these stages.

A. Correlated Visual Word Set Mining

Mining correlated visual word sets with exact algorithms can be prohibitively expensive. We avoid exhaustive pairwise comparisons by relying on Min-LSH. Typically, the binary vectors $\hat{I_1}, \hat{I_2}, \ldots, \hat{I_N}$ of T are input to Min-LSH in order to retrieve similar images (or documents). In contrast, we input the occurrence vectors $\hat{v_1}, \hat{v_2}, \ldots, \hat{v_M}$ of T to mine visual words with correlated occurrences (we call these sets *correlated word sets*). Formally, a correlated word set is defined as follows.

Definition 1: A set of visual words ϕ is a correlated word set if all visual words $v_i \in \phi$ agree in a given tuple g_j.

In other words, a correlated word set consists of the visual words stored in the same hash bucket in a given hash table. We expect that visual words that appear together in several images are mapped to the same hash bucket with high probability. On the other hand, since no minimum support is considered, rare visual words that occur together incidentally will have high probability of being stored in the same hash bucket. However, the stop list diminishes this problem by discarding very rare visual words. In this way, Min-LSH groups discriminative visual words that are likely to belong to the same object class and filter out noisy ones.

B. Pattern Summarization

The number of correlated word sets generated by Min-LSH will be typically large and highly redundant. On the other hand, due to image variations and visual polysemes (visual words belonging to different objects), visual words belonging to the same class may be divided into different correlated word sets. To cope with these problems, we perform a pattern summarization by clustering the discovered sets in a hierarchical agglomerative manner. We consider that two correlated word sets with large overlap are likely to belong to the same object class. The rationale is that highly discriminative visual words will appear in different correlated sets together with other informative visual words. Hence, we merge similar correlated word sets into a single set. The next Rule formalizes this idea.

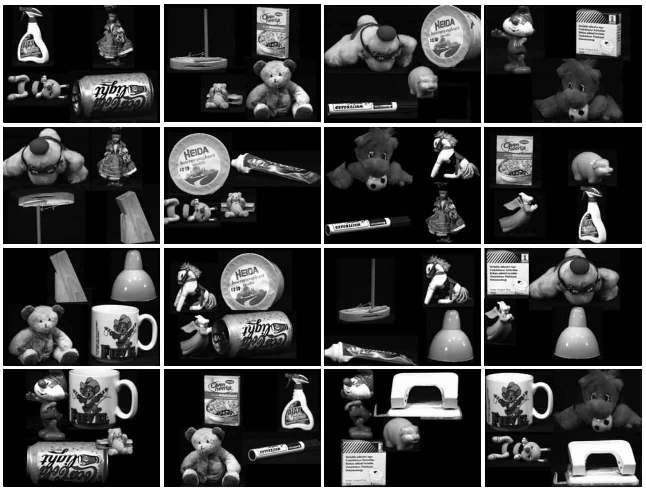

Figure 1. Sample images with multiple objects.

Rule 1: Two correlated visual word sets ϕ_1, ϕ_2 are merged into a single set $\phi_{1,2} = \phi_1 \cup \phi_2$ if $sim(\phi_1, \phi_2) > \mu$.

The similarity measure $sim(\phi_1, \phi_2) \in [0,1]$ in Rule 1 is given by the following equation.

$$sim(\phi_1, \phi_2) = \frac{\mid \phi_1 \cap \phi_2 \mid}{min(\mid \phi_1 \mid, \mid \phi_2 \mid)}. \tag{5}$$

We apply Rule 1 in a hierarchical agglomerative manner in order to cluster multiple correlated word sets that belong to the same class and derive more representative sets. Finally, only clusters with more than τ visual words are regarded as meaningful object classes.

V. Experimental Results

We carried out the experiment on the ALOI database [12]. This database contains images of a thousand different objects with illumination and viewpoint variations. Each image in the database depicts a single object against an homogeneous background. In order to evaluate our approach against clutter, we combined images of different objects into a single image. We used the set of stereo images of 22 different objects to generate a total of 18 images with four different objects each. In Fig. 1 we illustrate some examples of these images.

Our experiment consisted of two stages: 1) discovery and 2) classification. In the former, the method discovers the object classes and represents each of them as a set of visual words. In the latter, the discovered object classes were used to classify the 18 images. Note that since each image contains 4 different objects, each one should be assigned

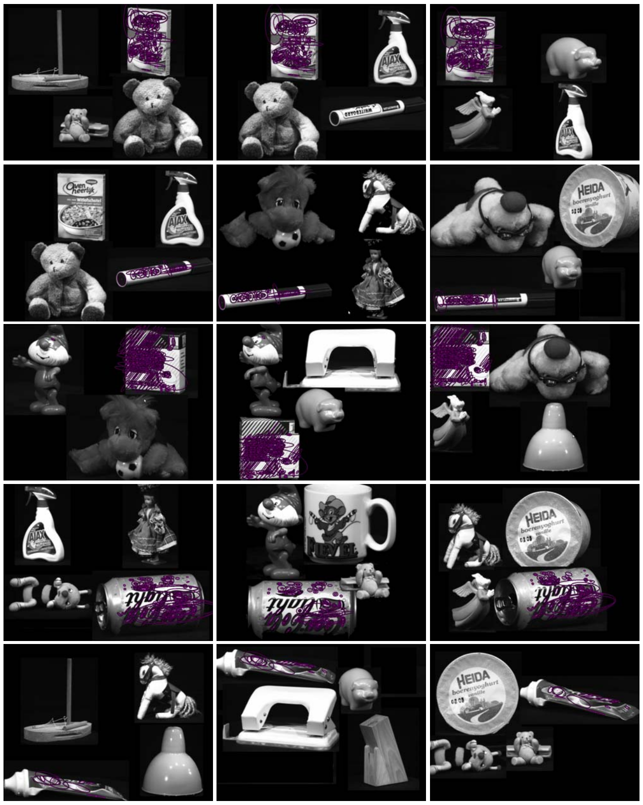

Figure 2. Examples of five discovered object classes. Each row illustrates the matching descriptors of a discovered object class.

to 4 different classes. The values of the parameters of our approach were the following: $k = 120$, $r = 3$, $l = 40$, $\mu = 0.66$ and $\tau = 5$. Our method discovered in total 28 object classes. From the 22 ground truth objects, our method correctly classified all the images that contain 12 of these objects and only 2 images of other 6. In addition, our method assigned all the images of 2 different ground truth objects to a single discovered class. Figure 2 illustrates some representative examples of the discovered objects.

VI. Conclusions

This paper proposed an approach to automatically discovering particular objects from a set of unannotated images. Our approach exploited the correlation of the occurrences of visual words belonging to the same object. We show that a correlated word set can effectively represent a particular object class. The utilization of Min-LSH to mine correlated word sets makes our system scalable to large collections of images and objects. By experiment, we show that our approach can efficiently discover meaningful object classes against clutter and slight viewpoint variations.

As future work, we plan to perform more extensive experiments on large image datasets and study the impact of the parameters r and l of the Min-LSH algorithm in the performance and accuracy of our approach. We are also interested in incorporating some spatial information to improve our results. Applying our approach to the discovery of object categories represents also an interesting future direction.

VII. Acknowledgment

The authors would like to thank anonymous reviewers for many helpful comments. This research was supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan. Grant-in-Aid for Scientific Research (C), 22500122, 2010.

References

[1] J. Sivic, , and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *ICCV*, 2003, pp. 1470–1477.

[2] J. Philbin, J. Sivic, and A. Zisserman, "Geometric lda: A generative model for particular object discovery," in *BMVC*, 2008.

[3] J. Tang and P. H. Lewis, "Non-negative matrix factorisation for object class discovery and image auto-annotation," in *CIVR*, 2008, pp. 105–112.

[4] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *ICCV*, 2005, pp. 370–377.

[5] G. Fuentes Pineda, H. Koga, and T. Watanabe, "Unsupervised object discovery from images by mining local features using hashing," in *CIARP*, 2009, pp. 978–985.

[6] O. Chum, J. Philbin, and A. Zisserman, "Near duplicate image detection: min-hash and tf-idf weighting," in *BMVC*, 2008, pp. 17–24.

[7] O. Chum, M. Perdoch, and J. Matas, "Geometric min-hashing: Finding a (thick) needle in a haystack," in *CVPR*, 2009, pp. 17–24.

[8] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *VLDB*, 1999, pp. 518–528.

[9] E. Cohen, M. Datar, S. Fujiwara, A. Gionis, P. Indyk, R. Motwani, J. D. Ullman, and C. Yang, "Finding interesting associations without support pruning," *TKDE*, vol. 13, no. 1, pp. 64–78, 2001.

[10] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *BMVC*, 2002, pp. 384–393.

[11] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *IJCV*, vol. 65, no. 1-2, pp. 43–72, 2005.

[12] J.-M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, "The Amsterdam library of object images," *IJCV*, vol. 61, no. 1, pp. 103–112, 2005.