Bag-of-Visual-Words Expansion Using Visual Relatedness for Video Indexing

Yu-Gang Jiang
Department of Computer Science
City University of Hong Kong
yjiang@cs.cityu.edu.hk

Chong-Wah Ngo
Department of Computer Science
City University of Hong Kong
cwngo@cs.cityu.edu.hk

ABSTRACT

Bag-of-visual-words (BoW) has been popular for visual classification in recent years. In this paper, we propose a novel BoW expansion method to alleviate the effect of visual word correlation problem. We achieve this by diffusing the weights of visual words in BoW based on visual word relatedness, which is rigorously defined within a visual ontology. The proposed method is tested in video indexing experiment on TRECVID-2006 video retrieval benchmark, and an improvement of 7% over the traditional BoW is reported.

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing

General Terms: Algorithms, Experimentation.

Keywords: Bag-of-visual-words, visual relatedness, expansion, video indexing.

1. INTRODUCTION

Recently, bag-of-visual-words (BoW) deriving from local keypoints have shown remarkable performance for image and video classification [4, 6]. Keypoints are salient image patches containing rich local information about an image, which can be automatically detected and represented by various detectors and descriptors. Keypoints are then grouped into a number of clusters and each cluster is treated as a visual word. With its keypoints mapped into the visual words, an image can be represented as a feature vector according to the presence or count of each visual word, which forms the basic visual cue in the classification task. This BoW representation is analogous to the bag-of-words representation for text document.

Under the textual bag-of-words model, a document will not be retrieved if it does not contain the terms that are in a query. This will result in poor recall of a retrieval system when the document contains synonymous terms to the query terms. In visual word based image/video classification, the problem may be even more serious. This is due to the fact that visual words are the outputs of clustering algorithms, and can be correlated to each other due to the quantization effect. Motivated by the textual query expansion method which can be used to effectively alleviate the synonym problem, we present a novel method for BoW expansion to remedy the effect of visual word correlation. Figure 1 shows an overview of our approach. Firstly, based on a vocabulary of visual words, a *visual ontology* is constructed to model the hyponym (is-a) relationship of visual words. Within the visual ontology, *visual relatedness* of visual words is rigorously defined, similar to estimating semantic relatedness of textual words using general purpose resources such as WordNet ontology. Finally, the visual relatedness is cleverly incorporated into BoW for visual word expansion.

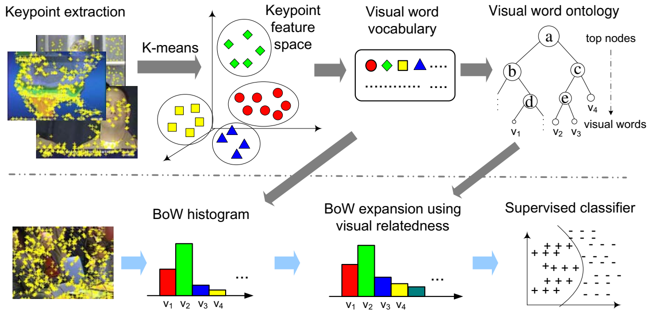

Figure 1: Overview of our approach. The visual word relatedness is rigorously defined in a visual ontology and then used for BoW expansion.

2. VISUAL WORD RELATEDNESS

In this section, we describe our approach for constructing visual ontology and estimating visual relatedness within the ontology. Given a set of keypoints, we first construct a visual vocabulary through clustering the keypoints by k-means algorithm. With the visual vocabulary, a visual ontology is further generated by adopting agglomerative clustering to hierarchically group two visual words at a time in the bottom-up manner. Consequently, the visual words in the vocabulary are represented in a hierarchical tree, namely visual ontology, where the leaves are the visual words and the internal nodes are ancestors modeling the is-a relationship of visual words. An example of the visual ontology is shown on the upper right of Figure 1. In the visual ontology, each node is a hyperball in the keypoint feature space. The size (number of keypoints) of the hyperballs increases when traversing the tree from leaves to root.

With the visual ontology, similar with the semantic relatedness measurements of textual words, the visual relatedness of visual words can also be explored by considering several popular ontological factors such as path length, or information content (IC). In this paper, we adopt JCN [1] to estimate the visual relatedness. Denote v_i and v_j as two visual words, JCN considers the ICs of their common ancestor

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

and the two compared words, defined as:

$$\mathrm{JCN}(v_i, v_j) = \frac{1}{\mathrm{IC}(v_i) + \mathrm{IC}(v_j) - 2 \cdot \mathrm{IC}(\mathrm{LCA}(v_i, v_j))}, \quad (1)$$

where LCA is the lowest common ancestor of visual words v_i and v_j in the visual ontology. IC is quantified as the negative log likelihood of the word probability. The probability is estimated by the percentage of the keypoints in a visual hyperball. For example, the top node "a" in Figure 1 has $\mathrm{IC}(a) = 0$ since $p(a) = 1$.

3. BoW EXPANSION WITH VISUAL RELATEDNESS

The visual relatedness is used directly to expand BoW. Let V be a vocabulary of n visual words: $V = (v_1, v_2, \ldots, v_n)$. With the vocabulary, an image I can be represented as a feature vector $F_I = (w_{v_1}, w_{v_2}, \ldots, w_{v_n})$, where w_{v_i} denotes the weight of word v_i in the image. Based on the visual relatedness calculated by JCN, we perform visual word expansion by diffusing weight w_{v_i} of word v_i to another word v_j:

$$w_{v_j} = w_{v_j} + w_{v_i} \times \mathrm{JCN}(v_i, v_j) \times \alpha, \quad (2)$$

where α is a parameter to control the degree of influence of the JCN relatedness. The aim of the word expansion is to alleviate the problem of visual word correlation. More specifically, the weight of a word is diffused by the influence of other ontologically related words. The diffusion inherently results in the expansion of words in an image to facilitate the utilization of word-to-word correlation for image comparison. For instance, assume that we have a vocabulary of only two visual words. Given two images each contains one word, say v_1 and v_2 respectively. With the traditional BoW representation, the L1 distance of the two images is $w_{v_1} + w_{v_2}$. After applying the BoW expansion, the distance will be $|w_{v_1} - \alpha \times w_{v_2} \times \mathrm{JCN}(v_1, v_2)| + |w_{v_2} - \alpha \times w_{v_1} \times \mathrm{JCN}(v_1, v_2)|$, which is smaller if v_1 and v_2 are highly related to each other.

While the idea of using BoW expansion appears intuitive, expanding all the words will sacrifice the sparse property of the original BoW representation. Hence, a compromising scheme is to firstly sort the JCN relatedness of all word pairs, and then only perform expansion for word pairs with higher relatedness.

4. EXPERIMENTS

To verify the performance of our approach, we conduct video indexing experiments on TRECVID-2006 dataset where the training and testing sets consist of 61,901 and 79,484 video shots respectively. The aim of video indexing is to rank the video shots according to the presence of semantic concepts. We use the 20 semantic concepts which were selected in TRECVID-2006 evaluation [5]. We select one key frame per shot, and the keypoints are detected by DoG [3] and described by SIFT [3]. A recent study in [2] showed that the performances of BoW are similar on this dataset for vocabulary sizes ranging from 500 to 10,000. We thus generate a vocabulary of 500 visual words for efficiency. For all the key frames, the BoW features are calculated based on term frequency (*tf*). For each semantic concept, two SVM classifiers are trained respectively using the original BoW features, and the new features after expansion (BoW-JCN). Predictions of the SVMs on the testing set are converted into posterior probabilities, and the performance evaluation follows TRECVID's standard using average precision (AP) computed over top 2,000 ranked shots.

In our experiments, the parameter α in Eq. 2 is empirically chosen as 0.5. Expansion is performed among word pairs with relatedness ranking among the top 1%. As a result, about half of the words are involved in the expansion process. Figure 2 contrasts the average precisions of BoW and BoW-JCN. BoW-JCN can improve the performance for 15 out of the 20 concepts, while the performance of the other 5 remains no change or slightly drops. In term of mean average precision (MAP) over the 20 concepts, the improvement of BoW-JCN (0.094) over BoW (0.088) is 7%. The results demonstrate that the visual words are indeed correlated to each other and using visual relatedness for BoW expansion is promising to alleviate this problem. Note that a MAP of 0.094 of a single feature approach is already comparable to the state-of-the-art results on this dataset.

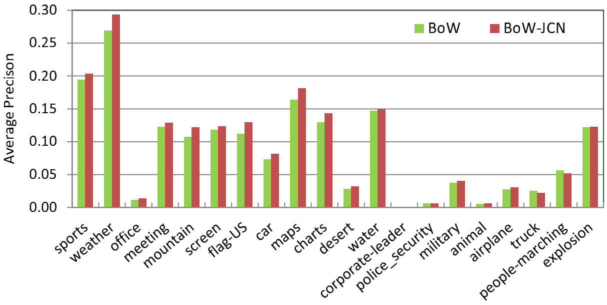

Figure 2: Performances of BoW and BoW-JCN for 20 semantic concepts.

5. CONCLUSION

We have presented a novel method for estimating visual word relatedness based on a visual ontology. We showed that the visual relatedness can be used for BoW expansion, in order to alleviate the problem of visual word correlation. Results of large-scale video indexing experiments verified the facts that the visual words are correlated to each other, and using visual relatedness for BoW expansion can lead to better performance.

Acknowledgements

The work described in this paper was fully supported by a grant from City University of Hong Kong (Project No. 7002241) and a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 118906).

6. REFERENCES

[1] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of ROCLING X*, 1997.

[2] Y. G. Jiang, C. W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *ACM CIVR*, 2007.

[3] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[4] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[5] TREC Video Retrieval Evaluation (TRECVID). *http://www-nlpir.nist.gov/projects/trecvid/*.

[6] J. Zhang et al. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 73(2):213–238, 2007.