ONTOLOGY-BASED VISUAL WORD MATCHING FOR NEAR-DUPLICATE RETRIEVAL

Yu-Gang Jiang and Chong-Wah Ngo

Department of Computer Science, City University of Hong Kong

ABSTRACT

This paper proposes a novel approach to exploit the ontological relationship of visual words by linguistic reasoning. A visual word ontology is constructed to facilitate the rigorous evaluation of linguistic similarity across visual words. The linguistic similarity measurement enables cross-bin matching of visual words, compromising the effectiveness and speed of conventional keypoint matching and bag-of-word approaches. A constraint EMD is proposed and experimented to efficiently match visual words. Empirical findings indicate that the proposed approach offers satisfactory performance to near-duplicate retrieval, while still enjoying the merit of speed efficiency compared with other techniques.

1. INTRODUCTION

Near-duplicate keyframe retrieval has recently attracted numerous research attentions for its potential on large-scale video search and summarization. The retrieval task is generally challenging as the degree of near-duplicate can extend to similar keyframes with variations in viewpoint, lighting, editing and acquisition time [1]. Popular approaches for near-duplicate retrieval are mostly based on keypoint features [1, 2, 3], for their tolerance to geometric and photometric variations. Keypoints are salient regions extracted locally over image scales, which were proven to be excellent for tasks like object and texture categorization [4].

Based on the use of keypoints, we can broadly categorize the recent works on near-duplicate retrieval as: keypoint matching, and bin-to-bin comparison with vocabulary generation. The former adopts schemes such as nearest neighbor search [2] and one-to-one symmetric (OOS) mapping [1] to match keypoints across keyframes. This category of approaches is normally slow since the amount of keypoints can range from tens to thousands on average per keyframe. For speed consideration, the second category adopts bin-to-bin word comparison through the offline quantization (or clustering) of keypoints [3]. The quantization generates a vocabulary composing of visual words generally referred as bag-of-words (BoW). By assigning keypoints to the nearest words in vocabulary, direct bin-to-bin comparison can be performed without exhaustive keypoint matching. While superior in speed, the signal loss in quantization and the heuristic in determining cluster setting and assigning words to keypoints, nevertheless, sacrifice the discriminative ability and generalization power of visual words.

In this paper, we propose a novel approach based on the ontology of visual words for near-duplicate retrieval. Figure 1(b) illustrates the main idea. An ontology is generated to model the word-to-word relationship on top of BoW. Analogue to the text-based ontology such as WordNet, the visual ontology captures the is-a relationship of visual words. By traversing the ontology, the linguistic similarity of different words (e.g., v_2 and v_4) can be rigorously defined based on the distance travelled ($v_2 \rightarrow e \rightarrow c \rightarrow v_4$), depth of their ancestor (node c at depth 1) in the ontology, and the probability of words seen. With the ontology, the quantization loss in generating BoW as in Figure 1(c) can be remedied since the similarity of words can be modeled not only through bin-to-bin but also cross-bin comparison. Furthermore, due to the use of BoW, there are normally less words to match, compared to Figure 1(a) where there are thousand of points available for matching. The proposed approach can be viewed as an *efficient* version of keypoint matching and an *extension* of visual vocabulary, where cross-bin matching of words, instead of keypoints, is enabled.

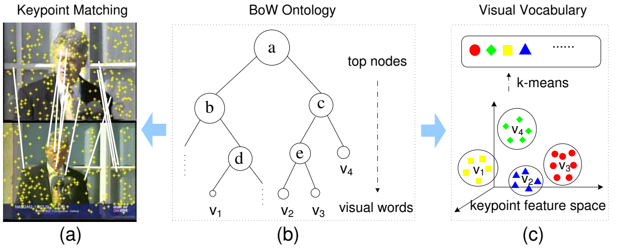


Fig. 1. Visual word ontology (b) as a bridge between expensive keypoint matching (a) and simple bin-to-bin word comparison (c).

There exist several works such as visual phrases [5] and proximity distribution kernels (PDK) [6] on modeling the relationship of visual words. Visual phrases capture the co-occurrence of words through pattern mining, while PDK models the geometric distribution of visual words. Our work in this paper focuses on the exploitation of ontological relationship among visual words for near-duplicate retrieval. Different from visual phrases and PDK, linguistic reasoning attempts to exploit the hyponym of words rather than co-occurrence or geometric relationships.

2. ONTOLOGY-BASED VISUAL LINGUISTICS

2.1. Ontology of Visual Words

In textual information retrieval (IR), linguistic reasoning has been known as a useful feature for word disambiguation. Take the words "car" and "truck" as an example. They are not matched by comparing characters, but can be semantically linked by "motor vehicle" through the is-a relationship in ontology. Building such an ontology for visual words enlightens the possibility of modeling word-to-word similarity, and meanwhile, reduces the signal loss during

The work described in this paper was fully supported by a grant from City University of Hong Kong (Project No. 7002112) and a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 118906).

quantization. Since visual words are the outcome of clustering, apparently the is-a relationship can be explicitly mined by considering the proximity among clusters.

Given a set of keypoints, we first construct a visual vocabulary through the clustering of keypoints by k-means algorithm. Each keypoint cluster is treated as a "visual word" in the vocabulary, and thus forms the BoW for describing visual content. With BoW, a visual ontology is further generated by adopting agglomerative clustering to hierarchically group two words at a time in the bottom-up manner. Consequently, the visual words in a vocabulary are represented in a hierarchical tree (ontology), where the leaves are the words and the internal nodes are ancestors modeling the hyponym (is-a relationship) of words. Figure 1(b) shows an example of the visual ontology. Each node is a hyperball in the keypoint feature space. The size (number of keypoints) of the hyperballs increases when climbing the tree upward.

2.2. Linguistic Similarity of Visual Words

With the BoW ontology, the linguistic similarity can be explored by considering the specificity, path length and information content (IC) of words. The specificity refers to the depth of a word in the tree. The deeper a word, the more specific the word. Path length means the minimum number of links to traverse from one word to the other. IC is inversely proportional to the probability of a word being seen, analogue to the inverse document frequency which can be computed when generating the ontology. We adopt three popular text linguistic measures to exploit the BoW ontology.

2.2.1. Resnik

Resnik considers the IC of common ancestor for similarity measure [7]. Denote v_i and v_j as two visual words, Resnik is defined as

$$sim(v_i, v_j) = \mathrm{IC}(\mathrm{LCA}(v_i, v_j)), \tag{1}$$

where LCA is the lowest common ancestor of v_i and v_j. IC is quantified as the negative log likelihood of word/node probability: $IC(v) = -\log p(v)$, where the probability $p(v)$ is estimated by the percentage of keypoints in the visual hyperball v. For instance, the top node "a" in Figure 1(b) has IC $= 0$ since $p(a) = 1$.

2.2.2. JCN

Resnik has the disadvantage that all words sharing one LCA have the same similarity, despite how far the distances between them. JCN deals with this problem by also considering the ICs of the compared words, defined as [8]:

$$sim(v_i, v_j) = \frac{1}{\mathrm{IC}(v_i) + \mathrm{IC}(v_j) - 2 \cdot \mathrm{IC}(\mathrm{LCA}(v_i, v_j))}. \tag{2}$$

2.2.3. WUP

In addition to IC, WUP considers the path length and the depth of words to measure the linguistic similarity [9]:

$$sim(v_i, v_j) = \frac{2 \cdot \mathrm{depth}(\mathrm{LCA}(v_i, v_j))}{\mathrm{len}(v_i, v_j) + 2 \cdot \mathrm{depth}(\mathrm{LCA}(v_i, v_j))}, \tag{3}$$

where $\mathrm{len}(v_i, v_j)$ represents the minimum path length between word v_i and v_j, and $\mathrm{depth}(\mathrm{LCA}(v_i, v_j))$ is the depth of node $\mathrm{LCA}(v_i, v_j)$ in the BoW ontology.

3. LINGUISTIC MATCHING WITH EMD

Based on the BoW ontology, two different visual words can always be matched by measuring their linguistic similarity. Consequently, given m words in a keyframe, there is $O(m^2)$ possible matching of words for comparing a keyframe pair. Because the bin-to-bin measures such as cosine similarity and Euclidean distance cannot characterize the linguistic similarity, we adopt earth mover's distance (EMD) [10] as the underlying framework for matching two sets of visual words across bins. Specifically, the ground distance of EMD is based on the linguistic measure, while the weight for a word is characterized, for instance, by the frequency of its appearance in a keyframe.

EMD measures the distance between two weighted point sets as a transportation problem [10]. A point set is normally referred to as a signature. EMD strives to find the minimum amount of "work" to transport the weights from one signature to the other. In BoW, a keyframe P is represented as a signature $P = \{(p_1, w_{p_1}), ..., (p_m, w_{p_m})\}$ of m words, where p_i indexes the p_ith visual word in the vocabulary, and w_{p_i} is the corresponding weight. To match P with another keyframe Q of n words, the EMD is computed as

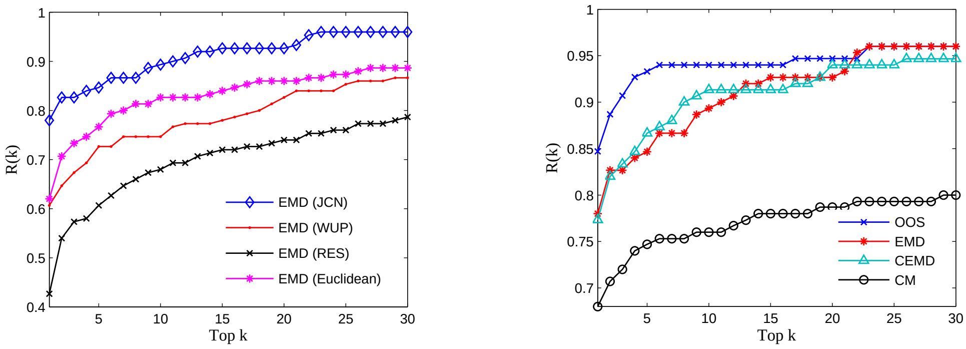
$$EMD(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}}, \tag{4}$$

where the ground distance d_{ij} between words v_{p_i} and v_{q_j} is measured via the linguistic similarity such as JCN. The flow f_{ij}, representing the amount of weight transferred from v_{p_i} to v_{q_j}, is optimized during the transportation.

3.1. Constraint Matching

While the idea of adopting EMD for exploring linguistic similarity appears intuitive, the approach suffers from speed inefficiency. Suppose the number of visual words in two signatures are m, the complexity of EMD is $O(m^3 \log m)$. Considering that there are generally tens to hundreds of visual words in a keyframe, the matching could be computationally intensive. Here we propose a novel constraint matching by dividing the visual vocabularies into k groups of words, namely visual chapters, and consequently enforcing EMD not to match words across chapters. This is equivalent to "distributive matching" where there are k EMDs being performed for each chapter, and then merged as a whole. The idea is based on the fact that visual words are clusters in nature. Certain categories of visual words (e.g., people and building) are seldom matched, and thus can be ignored from EMD matching when speed is an issue to consider.

To learn the visual chapters of a vocabulary, we compute a flow matrix $\mathbf{F}$ by observing the accumulated flows (f_{ij}) of EMD over a set of training examples. EMD will basically transport flows among similar words. The matrix $\mathbf{F}$ hints the correlation among visual words, and each entry F_{ij} indicates the total sum of flows between two words v_i and v_j. By treating $\mathbf{F}$ as a similarity matrix, an undirected fully-connected graph is constructed over $\mathbf{F}$, where nodes are words and edges represent similarities based on EMD flows. The normalized cut algorithm [11] is then employed to partition the graph into disjoint k sub-graphs. Each sub-graph is treated as a visual chapter of the vocabulary. The visual words with lower amount of flows are expected to stay in different chapters.



(a)

(b)

Fig. 2. Experimental results on Columbia dataset: (a) Comparison of linguistic measures; (b) Performance of EMD versus CEMD.

With k chapters of words, the constraint EMD, namely CEMD, of two keyframes P and Q is performed by running EMD separately in each chapter and then combined as

$$CEMD(P,Q)=\sum_{c=1}^{k}(\frac{S_{P_c}}{S_P}+\frac{S_{Q_c}}{S_Q})EMD(P_c,Q_c), \qquad (5)$$

where S_{P_c} is the number of visual words that P has in chapter c, and similarly for S_{Q_c}. Note that $S_P = \sum_{c=1}^{k} S_{P_c}$ and $S_Q = \sum_{c=1}^{k} S_{Q_c}$. For two keyframes with m visual words, the speed of CEMD is improved to $O(k\times(\frac{m}{k})^3\log(\frac{m}{k}))$. While the level of complexity in terms of big-O is the same as the original EMD, CEMD is practically more efficient (c.f. Table 2 in the experiments).

4. EXPERIMENTS

To verify the performance of the proposed approach, we conduct various near-duplicate keyframe retrieval experiments. We use the Columbia dataset [12] which contains 600 keyframes from TRECVID-2004 benchmark. There are 150 near-duplicate pairs in this dataset, and we use all of them (300 duplicates) as queries for assessing retrieval performance. The evaluation is based on the probability of successful top-k retrieval [12], defined as $R(k)=N_c/N_a$ where N_c is the number of queries that find its duplicate in the top k list, and N_a is the total number of queries. To further strengthen our claim, we also use a larger dataset - TRECVID-2006 test set containing a total of 79,484 keyframes in the experiments.

Throughout the experiments, we use DoG [13] as the keypoint detector and SIFT [13] of 128-dimensional feature as the descriptor. For Columbia dataset, a visual vocabulary of 1,000 words is built, associated with an ontology of 32 levels. For TRECVID-2006 test set, a smaller vocabulary of 500 words is constructed due to speed reason considering the large amount of keyframes in the dataset. The depth of the associated ontology is 23.

4.1. Comparison of Linguistic Measures

First, let us experiment and compare three linguistic measures: JCN, Resnik (RES) and WUP, with EMD as the distance measure. We use term frequency (TF) as the weighting scheme to generate the BoW feature vectors (signatures) for EMD matching. The TF weights the importance of a word using the frequency of its appearance in a keyframe. Figure 2(a) shows the performance comparison of the three measures on Columbia dataset. Among them, JCN demonstrates the best performance for accounting the ICs of visual words and their ancestor. Resnik, considering only the IC of the lowest common ancestor (LCA), loses the discriminative power as assigning equal similarity to words sharing the same LCA. WUP, utilizing path length and depth, does not show apparent advantage over JCN, while still performing better than Resnik. We investigate the results and find that this is mainly because the similarity of some words is set close to 0 as long as their LCA near to root (where depth(LCA)=0), despite the distance between two words. Our finding indeed indicates that the ancestor relationship and ICs of words are the best pieces of resources to use in near-duplicate retrieval. To further justify the usefulness of the linguistic measure, we also compare the performances with EMD which uses Euclidean distance between words (cluster centroids) as the ground distance. As shown in Figure 2(a), Euclidean is not better than JCN, but still outperforms WUP and Resnik. This probably indicates that word distance is an important factor that should not be ignored as in Resnik. While JCN and WUP do not account word distance, the information can be indirectly inferred from the ICs and path length of words. JCN, when considering ICs of three parties (words and their LCA), shows better performance because of the additional consideration of IC (cluster size) and LCA (global view of inter-cluster distance and density).

4.2. CEMD vs. EMD

Next, we compare the performance of CEMD and EMD with JCN as the linguistic measure. In CEMD, a total of 300 keyframes from TRECVID-2005 benchmark are randomly selected for generating the visual chapters. This training set is independent of the Columbia and TRECVID-2006 test sets. In our experiment, the vocabulary is empirically divided into 8 chapters. Figure 2(b) shows the performances of CEMD and EMD, in comparing with OOS [1] and block-based color moment (CM). OOS, in contrast to our approach, adopts keypoint matching (without vocabulary) and thus is computationally slow. Nevertheless, since no quantization loss is involved, OOS can offer upper limit performance. CM, on the other hand, serves as a baseline. As shown in Fig 2(b), the performance of CEMD is highly competitive with EMD. CEMD offers better retrieval rate for the top-

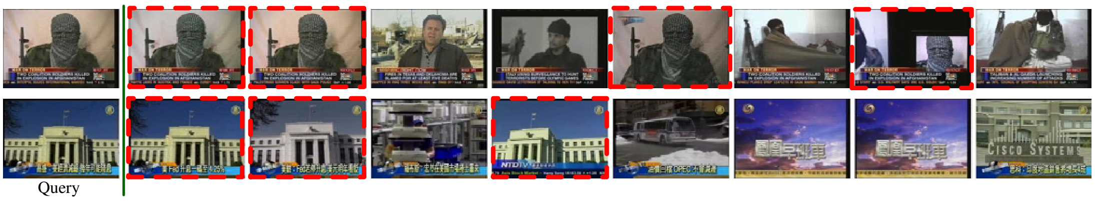


Fig. 3. Examples of near-duplicate keyframes retrieved by CEMD. The query keyframes are shown on the left most column, followed by the most similar retrieved keyframes. The true positives are marked in red boxes.

Table 1. Improvement of CEMD over BoW on TRECVID-2006 test set.

	BoW	CEMD	Improvement
MAP	0.479	0.549	14.6%

Table 2. Per query retrieval efficiency on TRECVID-2006 test set.

Keypoint-based	Ontology-based		Vocabulary-based	Baseline
OOS	EMD	CEMD	BoW	CM
$5h$ $44m$	$2h$ $5m$	$2m$ $59s$	$50s$	$22s$

k ($k \leq 10$) list, despite the fact that CEMD is about 40 times faster than EMD. Compared with OOS, CEMD offers lower precision but faster speed (about 110 times), considering that there are 202 words (against 340 keypoints) on average in each keyframe for matching.

4.3. Performance on TRECVID-2006 dataset

To further verify the performance, we also conduct experiments on a larger dataset: TRECVID-2006 test set containing 79,484 keyframes. We experiment 110 near-duplicate queries randomly found in the test set. Each approach returns the top-40 ranked keyframes for performance comparison. To evaluate the results, two assessors were invited to label the returned keyframes and then produce ground-truth. We then calculate an average precision over the top-40 list for each query.

We compare the proposed CEMD approach with BoW based on cosine similarity [3]. The mean average precisions (MAP) over the 110 queries are reported in Table 1. As shown in the Table, CEMD outperforms BoW by 14.6%. This again confirms the effectiveness of the proposed visual linguistics for near-duplicate retrieval. Figure 3 shows some examples of near-duplicate keyframes retrieved by the CEMD. Our approach could successfully retrieve near-duplicates with variations such as lighting, color and scale.

Table 2 lists the average response time of a query on TRECVID-2006 test set, including the time of uploading features and saving results. The experiments are conducted on a Pentium-4 3GHz machine. The proposed CEMD is significantly faster than EMD (about 40 times) and OOS (about 110 times). OOS is extremely slow due to the large amount of keypoints available for matching between all keyframe pairs. If the index structure LIP-IS proposed in [1] is used, the speed of OOS matching would be improved to around 30 minutes, which is still significantly slower compared with the proposed CEMD.

5. CONCLUSION

Motivated by the text ontology in IR that is useful for word disambiguation, we have introduced an approach to exploit visual ontology in order to remedy the quantization loss in generating visual words. Our findings indicate that the ontology-based hyponym relationship and information content are useful for modeling the linguistic similarity of visual words. With the proposed constraint EMD matching, we have empirically demonstrated the potential of the visual linguistics on near-duplicate retrieval experiments. However, while ontology is shown to be useful, we only explore the hyponym relationship of visual words. Other aspects such as synonymy and polysemy of visual words could be further studied to extend our current work.

6. REFERENCES

[1] C. W. Ngo, W. L. Zhao, and Y. G. Jiang, "Fast tracking of near-duplicate keyframes in broadcast domain with transitivity propagation," in *ACM Multimedia Conference*, 2006.

[2] Y. Ke, R. Suthankar, and L. Huston, "Efficient near-duplicate detection and sub-image retrieval," in *ACM Multimedia Conference*, 2004, pp. 869–876.

[3] X. Wu, W. L. Zhao, and C. W. Ngo, "Near-duplicate keyframe retrieval with visual keywords and semantic context," in *ACM CIVR*, 2007.

[4] J. Zhang and et al., "Local features and kernels for classification of texture and object categories: A comprehensive study," *IJCV*, vol. 73, no. 2, pp. 213–238, 2007.

[5] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: from visual words to visual phrases," in *CVPR*, 2007.

[6] H. Ling and S. Soatto, "Proximity distribution kernels for geometric context in category recognition," in *ICCV*, 2007.

[7] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *IJCAI*, 1995.

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

[9] Z. Wu and M. Palmer, "Verb semantic and lexical selection," in *Annual Meeting of the ACL*, 1994, pp. 133–138.

[10] Y. Rubner and et al., "The earth mover's distance as a metric for image retrieval," *IJCV*, vol. 40, no. 2, pp. 99–121, 2000.

[11] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. on PAMI*, vol. 22, no. 8, pp. 888–905, 2000.

[12] D-Q. Zhang and S-F. Chang, "Detecting image near-duplicate by stochastic attributed relational graph matching with learning," in *ACM Multimedia Conference*, 2004.

[13] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.