Exploring Semantic Concept Using Local Invariant Features

Yu-Gang Jiang, Wan-Lei Zhao and Chong-Wah Ngo

Department of Computer Science
City University of Hong Kong, Kowloon, Hong Kong

{yjiang,wzhao2,cwngo}@cs.cityu.edu.hk

Abstract

This paper studies the role and performance of local invariant features arisen from interest points in describing and sketching semantic concepts. Both the local description and spatial location of interest points are exploited, separately and jointly, for concept-based retrieval. In concept description, a visual dictionary is generated with each keyframe being depicted as a vector of keywords. Semantic concepts are learnt and then spotted in this vector space model. In concept sketching, the location distribution of interest points, which outlines the basic shape of concepts, is novelly modelled with embedded Earth Mover's Distance. Experimental results with TRECVID-2005 corpus show that by incorporating both properties of interest points with baseline features, an improvement of 70% (over color) and 26% (over color and texture) in concept retrieval is reported.

Keywords: Semantic Concept Retrieval, Local Interest Point (LIP), Visual Keyword, LIP Distribution.

1 Introduction

Shot (or keyframe) retrieval through the filtering of semantic concepts has recently attracted numerous research attentions. Generally speaking, the concept filtering serves as a pioneering step in bridging the semantic gap of low-level features and high-level concepts. Previous evaluation in TRECVID (TREC video retrieval evaluation website n.d.) indicates that the state-of-the-art performance in concept-based retrieval is jointly or partly attributed to the advanced models in feature description, machine learning and multi-modality fusion (Amir et al 2005, Chang et al 2005). In this paper, we investigate the role of local invariant features, specifically the local interest point (LIP) and the related descriptors, in boosting the performance of concept retrieval from the view of feature-level analysis. Figure 1 illustrates the idea of our work. The middle column shows a group of LIPs overlaid on two keyframes with a concept *mountain*, and the right column sketches the basic outline of *mountain* in the keyframes with LIPs. Intuitively, both examples indicate the expressive and delineative power of LIPs respectively in locating key parts and describing the shape of a concept. This paper explores the potential of LIPs in these two aspects: 1) generate LIPs as visual keywords to describe concepts, 2) model the location distribution of LIPs to sketch concepts.

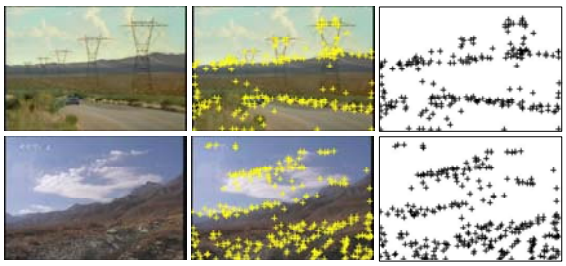


Figure 1: Keyframes with detected LIPs. Both the description and spatial location of the LIPs are utilized for semantic concept retrieval in our approach.

LIP-based representation has been studied for concept retrieval (Chang et al 2005), near-duplicate keyframe retrieval (Ke, Sukthankar, and Huston 2004, Zhang and Chang 2004) and image matching (Grauman and Darrell 2005). In (Chang et al 2005, Zhang and Chang 2004), a part-based random attributed relational graph model (RARG) is proposed to capture the statistical attributes and topological relationship of interest points. As reported in (Chang et al 2005), when incorporating this model with other global features such as color and texture, the performance of retrieval is increased by about 10%. Similar in spirit, we study the performance of LIPs, but we do not specifically cast LIPs to a learning model to infer the matching and topology of interest points as in (Chang et al 2005, Zhang and Chang 2004). The model could be expensive to learn and involves a couple of empirical parameters that are not easy to be optimized. Instead, we use a relatively simple learning platform to investigate how LIPs describe and sketch concepts purely by their features.

In (Ke, Sukthankar, and Huston 2004), LIP is also explored for keyframe and image matching. Due to the large amount of LIPs in a keyframe, locality sensitive hashing (LSH) is proposed for fast search of nearest neighbors. However, a recent study by Zhao, Jiang, and Ngo (2006) shows that the empirical performance of LSH for LIPs is indeed unsatisfactory. This is mainly due to the fact that LSH simulates approximate search and the chance of returning nearest neighbors in general is not high. As a consequence, the outcome of LIP matching often shows faulty and ambiguous matches. In (Grauman and Darrell 2005), the embedded Earth Mover's Distance (e-EMD) is employed to model the distribution of LIPs in feature space. They study the performance of image matching in three datasets with scenes (from the sitcom *Friends*), objects (from ETH-80) and textures (from VisTex), and yield encouraging result. Our work is different from (Ke,

Sukthankar, and Huston 2004) where we do not perform LIP matching and employ LSH respectively for speed and effectiveness reasons. In addition, our approach learns the location distribution of LIPs for concept sketching, rather than the distribution in feature space for matching which results in high dimensional feature representation as presented in (Grauman and Darrell 2005).

2 Approach Overview

This paper investigates the role of visual keywords and their location distribution in high-level concept retrieval. Figure 2 depicts the flow of our framework which is composed of a group of classifiers based on various descriptors including the proposed local invariant features. The color moment and wavelet texture serve as the baseline to judge the improvement of local features formed by LIPs.

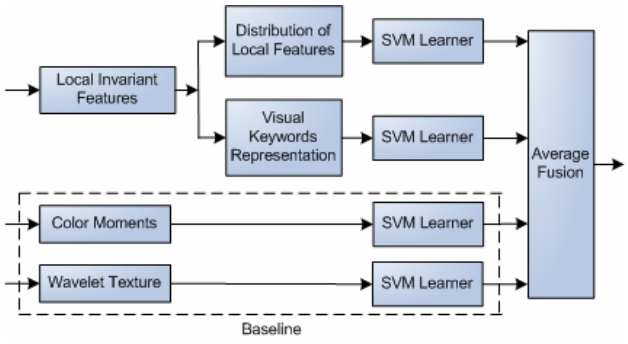


Figure 2: Framework.

In our approach, LIPs are located by the Difference-of-Gaussian (DoG) detector (Lowe 2004) over different scales. The detector is scale invariant and can tolerate certain amount of affine transformation. Each LIP is characterized by a 36-dimensional PCA-SIFT feature descriptor. The descriptor has been demonstrated to be distinctive and robust to color, geometric and photometric changes (Ke and Sukthankar 2004). Generally the number of LIPs in a keyframe can range from several hundreds to few thousands and thus prohibit the efficient matching of LIPs with PCA-SIFT across large amount of keyframes. We generate a visual dictionary as in (Sivic and Zisserman 2003) by offline quantization of LIPs. Subsequently each keyframe is described as a vector of visual keywords that facilitate direct keyframe comparison without point-to-point LIP matching. The local distribution of LIPs, on the other hand, is represented as shape-like features in the multi-resolution grids. The features are then embedded in a space where distance is evaluated with the e-EMD measure.

For each concept, an ensemble of classifiers as in Figure 2 is learnt. The extracted unimodal features are attached respectively to support vector machines (SVM) for discriminative classification in their own feature space. The results of various SVM learners are then re-ranked with average fusion. Since our aim is to investigate the role of LIPs from the feature-level point of view, we do not pay particular attention to the aspects of machine learning and multi-modality fusion. The framework we adopt is one of the commonly used platform for learning and fusion.

3 Local Feature Representation

3.1 Generating Visual Keywords

We generate a visual dictionary of LIPs based on (Sivic and Zisserman 2003). We select 1500 keyframes from TRECVID-2005 development set, with about 80% of them containing the 39 high-level concepts specified in TRECVID. In total, there are about 850,000 LIPs extracted. Empirically we quantize these local points into 5,000 clusters, and each cluster represents a visual keyword. With this visual dictionary, the classical *tf-idf* is used to weight the importance of keywords. A keyframe is then represented as a vector of keywords, analogous to the traditional text-based vector space model.

3.2 Modeling Location Distribution

We describe the distribution of LIPs with multi-resolution grid representation as illustrated in Figure 3. The size of grids varies at different resolutions and thus the granularity of shape information formed by LIP distribution changes according to the scale being considered. We compute the first three moments of grids to describe the shape-like information of LIPs across resolutions. Each grid is physically viewed as a point characterized by moments and weighted according to its level of resolution. With this representation, basically a keyframe is treated as a bag of grid points. The similarity between keyframes is based upon the matching of grid points within and across resolutions depending to their feature distance and transmitted weights that can be evaluated with Earth Mover's Distance (EMD). The complexity of EMD, nevertheless, is expensive and has an exponential worst case with the number of points.

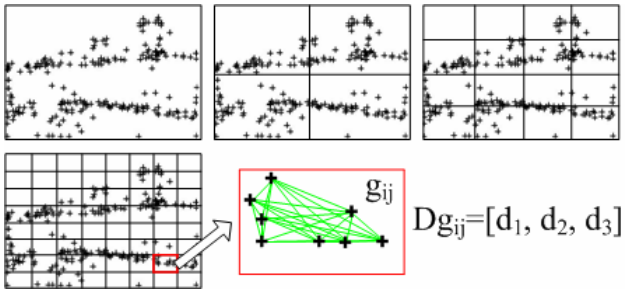


Figure 3: Modeling the distribution of local features. Grids with different resolution are imposed on the LIPs and a descriptor Dg_{ij} composed of three moments (d_1, d_2, d_3) are computed for each grid g_{ij}.

For speed reason, we adopt embedded EMD which provides a way to map the weighted point sets from the metric space into the normed space (Indyk and Thaper 2003) with low distortion. The basic idea of the EMD embedding is as follows: Let two point sets **P** and **Q** with equal cardinality *s*, each in $\Re^k$ and $\mathbf{V}=\mathbf{P}\cup\mathbf{Q}$. Imposing grids on the space $\Re^k$ of side length 2^i, $-1<i<\log(\triangle)$, where $\triangle$ is the diameter of **V**. Let G_i be grid of side 2^i, in order to embed a point set **P**, a vector $\mathbf{v}_i$ is constructed with one coordinate per cell, where each coordinate counts the number of points in the corresponding cell. Ultimately, by concatenating all $\mathbf{v}_i$ scaled by the side lengths, we can obtain the embedding of **P**:

Concepts	VK	LIP-D	VK+LIP-D	WT	CM
People Walking/ running	**0.150**	0.069	0.143	0.076	0.079
Explosion or fire	0.020	0.023	**0.031**	0.030	0.024
Map	0.127	0.244	0.281	0.270	**0.391**
US flag	**0.154**	0.020	0.082	0.045	0.072
Building exterior	0.127	0.067	0.131	0.067	**0.157**
Waterscape/waterfront	0.081	0.049	**0.147**	0.079	0.112
Mountain	0.107	0.085	**0.198**	0.074	0.100
Prisoner	0.001	0.000	0.001	**0.002**	0.001
Sports	0.100	0.104	0.206	0.134	**0.211**
Car	0.068	0.029	**0.107**	0.051	0.070
Mean Average Precision (MAP)	0.094	0.069	**0.133**	0.083	0.122

Table 1: Average precision of the 10 high-level concepts in TRECVID-2005 with different features. The best results are given in bold.

$$f(\mathbf{P}) = \left[\mathbf{v}_{-1}(\mathbf{P})/2, \mathbf{v}_0(\mathbf{P}), 2\mathbf{v}_1(\mathbf{P}) \ldots 2^i \mathbf{v}_i(\mathbf{P}) \ldots \right]. \quad (1)$$

In the embedded space, the normed distance between $f(\mathbf{P})$ and $f(\mathbf{Q})$ is an estimation of the exact EMD distance. The EMD embedding has a provable upper bound of distortion of $O(\log\triangle)$. Because the dimension of embedded vector is high, the locality sensitive hashing (LSH) technique is frequently used for nearest neighbor search (Grauman and Darrell 2005, Indyk and Thaper 2003).

In our approach, each LIP is indexed with its spatial location location (x, y) in the keyframe. To keep the length of feature vector in an acceptable level, we only impose grids with four side lengths, i.e., $1/8\triangle$, $1/4\triangle$, $1/2\triangle$ and $\triangle$ in this $2D$ space. Then, for each grid, the three moments of LIPs are computed to describe their distribution. The first moment counts the number of LIPs, while the second and third moments are the mean and variance of distances between all possible LIP pairs in the grid, as illustrated in Figure 3. Note that under the e-EMD setting, all grid points in the resolution i are grouped as a vector $v_i(\mathbf{P})$ weighted by 2^i in the subspace. In our case for semantic concept retrieval, instead of using LSH for fast searching, we adopt machine learning approach which is proved to have better performance than direct searching in a metric space. The SVM is expected to learn the decision boundary that discriminates the embedded vectors of a semantic concept from others in the one-against-all strategy.

4 Experiments

4.1 Data Set and Evaluation Criteria

We use TRECVID-2005 corpus to evaluate our proposed approach. The corpus contains more than 160 hours of broadcast videos, which were split in half chronologically. The first halves are used as test data while the second halves are used as development set for training. For the development set, TRECVID launched a collaborative annotation for several semantic concepts last year. In the experiment, we select a subset from the pool of manually annotated keyframes, together with some negative samples, for training of 10 concepts listed in Table 1. The 10 concepts are selected in TRECVID-2005 for feature evaluation.

We use the average precision over top-k retrieved shots for performance evaluation. Denote R as the number of true relevant shots in the corpus, and L as the ranked list of the retrieved shots. At any index j ($1<j<k$), let R_j be the number of relevant shots in the top j shots. Let I_j=1 if the j^{th} shot is relevant and 0 otherwise, the average precision is defined as

$$\mathrm{AP}(L) = \frac{1}{\min(R,k)} \sum_{j=1}^{k} \frac{R_j}{j} \times I_j. \quad (2)$$

We set k=2000, following the standard of high-level feature extraction in TRECVID evaluation.

4.2 Results and Discussion

We first evaluate the performance of visual keywords (VK) and LIP distribution (LIP-D) on 10 semantic concepts. Because VK and LIP-D are designed on two different attributes of LIPs, i.e., the description and spatial location respectively, we also use average fusion to combine both features. We compare the LIP-based features with grid based color moment (CM) and wavelet texture (WT). Both CM and WT have been shown as useful features in TRECVID-2005 corpus, which may partially because the time span of videos is about one month and there exist quite a lot of near-duplicate keyframes. In CM, three color moments (i.e., mean, standard deviation and skewness) are computed. Basically each keyframe is divided into 5 by 5 grids, and the color moments are computed for each grid in *Lab* color space. For WT, we use 3 by 3 grids and each grid is represented by the variances in 9 DB-4 wavelet sub-bands. While VK and LIP-D describe the LIPs (around corners and edges) that are robust to various transformations over different scale spaces, WT accounts for the statistical distribution of edge points in multi-resolution space.

Table 1 shows the performance comparison, where on average VK+LIP-D yields the best performance. VK is indeed useful for most of the concepts like *people walking/running*, *US flag*, *Building exterior*, *Car* and *Mountain*. This is due to the fact that these concepts are

Concepts	CM+WT	CM+WT+VK	CM+WT+LIP-D	CM+VK+LIP-D+WT
People Walking/running	0.137	0.198	0.167	**0.201**
Explosion or fire	0.035	0.040	0.041	**0.042**
Map	**0.398**	0.381	0.389	0.383
US flag	0.111	**0.148**	0.108	0.136
Building exterior	0.181	0.224	0.205	**0.225**
Waterscape/waterfront	0.172	0.215	0.210	**0.230**
Mountain	0.182	0.241	0.238	**0.288**
Prisoner	0.002	0.002	0.002	**0.002**
Sports	0.284	0.339	0.342	**0.367**
Car	0.140	0.181	0.153	**0.191**
Mean Average Precision	0.164	0.197	0.186	**0.207**

Table 2: Average precision of the 10 high-level concepts in TRECVID-2005 through average fusion of different features. The best results are given in bold.

mainly belong to objects or scenes, which can appear anywhere in the keyframes with different scales and viewpoints. VK, without any spatial information, is better for capturing these concepts. LIP-D is useful for the concepts like *map*, *sports*, and *mountain*, because these concepts exist with somewhat uniform background or contour pattern (e.g., *mountain* in Figure 1 that can be delineated with the location distribution of LIPs). On the contrary, LIP-D performs poorly for the concepts like *US flag*, *people walking/ running*, and *car*, which can arbitrarily appear anywhere in the keyframes and LIP-D cannot effectively capture their LIP distribution under the presence of background clutter. Overall, we can conclude that the two LIP-based features, VK and LIP-D, indeed complement each other. By the average fusion of VK and LIP-D, great improvement is achieved for most concepts. Moreover, in term of mean average precision (MAP), the LIP-based feature has better performance than both CM and WT, which are indeed quite effective features in TRECVID-2005. WT and LIP-based features, although redundant in certain sense, indeed complement each other as we will demonstrate in Table 2.

Our next experiment aims to evaluate the degree of improvement when fusing different features. Table 2 shows the performance of various feature combination, with CM+WT as baselines. With different kinds of feature combination, the average improvement over baseline ranges from 13% to 26%. The result indicates that when incorporate VK or LIP-D upon the baseline, both of them could have better performance. In particular, obvious improvement is noticed for the concepts *waterscape /waterfront*, *mountain* and *sports*. Overall, the best MAP for almost all the semantic concepts is attained when fusing all the four features together, see Figure 5 for *car* results. The only two exceptions are *map* and *US flag*. For *map*, most keyframes are indeed from weather news, for which color alone is enough to achieve high precision. The *US flag*, as mentioned before, can appear anywhere in keyframes with different shapes and scales. So, global features can not capture useful information and are easily affect by background clutter.

To have a better view of improvement when VK and LIP-D are incorporated upon the baseline, Figure 4 compares the performance of all four features against the median and best results in the evaluation of TRECVID-2005. By only using four features, seven out of ten concepts are higher than median while the other three are around median. The results are still far from the best mainly due to the fact that we just focus on the feature level point of view, and did not pay particular attention on the aspects of machine learning and multi-modality fusion. Even so, the result indeed indicates the potential of delivering state-of-the-art performance, by focusing only on the expressive power of feature-level information with simple fusion strategy.

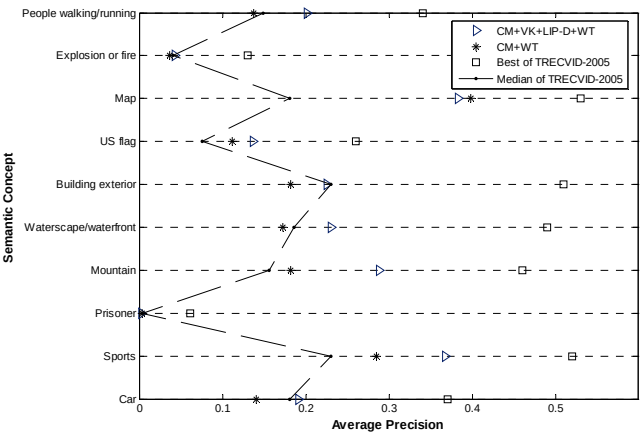


Figure 4: Comparison of our results with median and best of TRECVID-2005

4.3 Run Time

The generating process of our visual keywords takes 4 hours. Then, based on the visual dictionary, we use KD-tree to speed up the vector quantization for all the keyframes. The entire vector quantization process, including all training and testing data sets, takes 30-40 hours. For each concept, training a SVM model takes several seconds for CM, WT and LIP-D, and from several seconds to few minutes for VK, depends on the number of training keyframes. The predicting procedure of a semantic concept on the 80 hours testing data set takes 1-4 minutes for CM, WT and LIP-D, and 5-20 minutes for VK.

All of the computational costs are evaluated on a P4 3G machine with 512MB RAM.

5 Conclusion

We have presented two LIP-based features for semantic concept retrieval. One is visual keywords which are obtained by clustering of PCA-SIFT descriptors. The other feature is constructed based on the spatial distribution of the LIPs with EMD embedding. The experimental results on TRECVID-2005 data set show that local features are useful for most of the semantic concepts retrieval. In particular, the improvement of MAP is more significant for scenes (e.g., *mountain* and *water*) than that of objects (e.g., *car*). We conclude that the LIP-D is a good feature for most concepts except those object concepts with large variance in appearance and can appear anywhere in a keyframe. VK, on the other hand, is quite useful for those object concepts since it do not count in spatial information. Overall, by fusing the proposed feature descriptors with baseline features, the retrieval performance is boosted. We believe that the current result can be further improved if more advanced techniques in fusion and learning can be jointly considered.

Acknowledgement

The work described in this paper was fully supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 118905 and CityU 118906).

References

Amir, A. et al. (2005): IBM research trecvid-2005 video retrieval system. *Proc. of TRECVID workshop*, Gaithersburg, MD, USA, 133-149, NIST.

Chang, S. F. et al. (2005): Columbia university trecvid-2005 video search and high-level feature extraction. *Proc. of TRECVID Workshop*, Caithersburg, MD, USA, 41-48, NIST.

Grauman, K. and Darrell, T. (2005): Efficient image matching with distribution of local invariant features. *Proc. of International Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 627-634, IEEE Computer Society.

Indyk, P. and Thaper, N. (2003): Fast image retrieval via embeddings. *Proc. of 3rd International Workshop on Statistical and Computational Theories of Vision*, Nice, France.

Ke, Y. and Sukthankar, R. (2004): PCA-SIFT: A more distinctive representation for local image descriptors. *Proc. of International Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, 506-513, IEEE Computer Society.

Ke, Y., Sukthankar, R. and Huston, L. (2004): Efficient near-duplicate detection and sub-image retrieval. *Proc. of ACM Multimedia Conference*, New York, USA, 869-876, ACM Press.

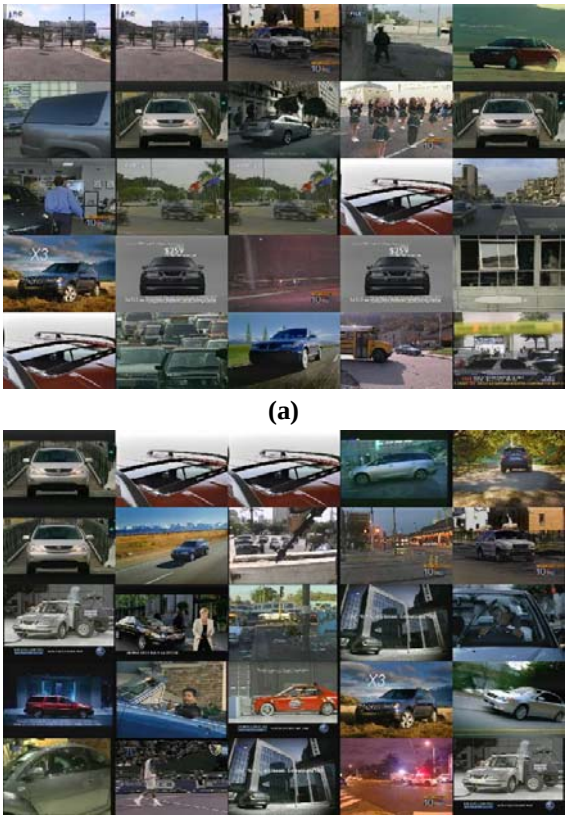
Lowe, D. (2004): Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*. **60**(2):91:110.

Sivic, J. and Zisserman, A. (2003): Video google: A text retrieval approach to object matching in videos. *Proc. of International Conference on Computer Vision*, Nice, France, 1470-1477, IEEE Computer Society.

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

Zhang, D. Q. and Chang, S. F. (2004): Detecting image near-duplicate by stochastic attributed relational graph matching with learning. *Proc. of ACM Multimedia Conference*, New York, USA, 877-884, ACM Press.

Zhao, W. L., Jiang, Y. G. and Ngo, C. W. (2006): Keyframe retrieval by keypoints: Can point-to-point matching help? *Proc. of International Conference on Image and Video Retrieval*, Tempe, AZ, USA, 72-81, Springer.



(a)

(b)

Figure 5: Top 25 results for *car*. Ordered left to right and top to bottom. (a) Baseline (20 are correct); (b) Average fusion of four features (24 are correct).