Experimenting VIREO-374: Bag-of-Visual-Words and Visual-Based Ontology for Semantic Video Indexing and Search

Chong-Wah Ngo, Yu-Gang Jiang, Xiaoyong Wei
Feng Wang, Wanlei Zhao, Hung-Khoon Tan and Xiao Wu
Department of Computer Science
City University of Hong Kong
{*cwngo,yjiang,xiaoyong,fwang,wzhao2,hktan,wuxiao*}*@cs.cityu.edu.hk*
http://vireo.cs.cityu.edu.hk

Abstract

In this paper, we present our approaches and results of *high-level feature extraction* and *automatic video search* in TRECVID-2007.

In high-level feature extraction, our main focus is to explore the upper limit of *bag-of-visual-words* (BoW) approach based upon local appearance features. We study and evaluate several factors which could impact the performance of BoW. By considering these important factors, we show that a local feature only system already yields top performance ($MAP = 0.0935$). This conclusion is similar to our recent experiment of VIREO-374 on TRECVID-2006 dataset [1], except that the improvement, when incorporating with other features, is marginal. Description of our submitted runs:

- CityU-HK1: linear weighted fusion of 4 SVM classifiers using BoW, edge histogram, grid based color moment and wavelet texture.
- CityU-HK2: average fusion of 5 SVM classifiers using BoW, spatial layout of keypoints, edge histogram, grid based color moment and wavelet texture.
- CityU-HK3: average fusion of 4 SVM classifiers using BoW, edge histogram, grid based color moment and wavelet texture.
- CityU-HK4: Bag-of-visual-words (BoW).
- CityU-HK5: average fusion of 3 baseline classifiers using edge histogram, grid based color moment and wavelet texture.
- CityU-HK6: average fusion of 2 baseline classifiers using grid based color moment and wavelet texture.

In automatic search, we study the performance of *query-by-example* (QBE) and VIREO-374 ontology-based concept search. In QBE, the spatial properties of local keypoints and concept

detector confidence are utilized for retrieval. In concept-based search, a small set of VIREO-374 detectors are selected for query answering by measuring the similarity of query terms to semantic concepts in an Ontology-enriched Semantic Space. We submit six runs composing of concept-based, query-based, motion-based and text-based search.

- CityUHK-SCS: concept-based search in which one single concept is selected for each query.
- CityUHK-MCS: concept-based search in which top-3 concepts are selected.
- CityUHK-Concept: use 36-d concept detection confidence vectors of keyframes for QBE.
- CityUHK-ConceptRerank: use 36-d concept detection confidence vectors to rerank the result of text baseline.
- CityUHK-VKmotion-Rank: employ the motion histogram of visual keywords (VK) in video sequence to rerank the result of text baseline.
- CityUHK-Text: baseline run by ASR/MT transcripts.

1 High-Level Feature Extraction

This year, we mainly focus on exploring the upper limit of local features for concept detection. Our local feature approach is basically based on our previous work in [2]. We also implement three baseline features and examine the improvement of fusing the local features with the baseline visual features. For the selection of training samples, we only rely on this year's data and combine the two publicly available annotations from LIG [3] and MCG-ICT-CAS.

1.1 Baseline

We extract three baseline features, namely edge direction histogram (EH), grid-based color moments (CM) and wavelet texture (WT). In EH, we generate a 324-d edge direction histogram over 3 × 3 grid partitions based on the Sobel edge detector. In CM, we calculate the first 3 moments of 3 channels in *Lab* color space over 5 × 5 grids, and aggregate the features into a 225-d feature vector. For WT, we use 3 × 3 grids and each grid is represented by the variances in 9 Haar wavelet sub-bands to form a 81-d feature vector.

We combine the three baseline features as our baseline system. The combination is done in "late fusion" manner, i.e. the final decision is made by fusing of the outputs of separate SVM classifiers. We use "average fusion" to combine different baseline features.

1.2 Effective Local Features

The local patches (keypoints) are detected by DoG [4] and Hessian Affine [5] separately, and described using SIFT [4]. In our BoW approach, firstly we build a vocabulary of 500 visual words (clusters of SIFT descriptors) for each detector, denoted as V_1 for DoG and V_2 for Hessian Affine respectively. Then the soft-weighting scheme proposed in [2] is used to weight the significance of a word in the keyframes, resulting in two kinds of 500-d feature vectors F_1 and F_2. In SVM

Table 1: Summarization of 6 Runs for High-Level Feature Extraction.

Run ID	*Component*	*Fusion Strategy*	*MAP*
Run 1	BoW+CM+WT+EH	Linear	0.0983
Run 2	BoW+LIP-D+CM+WT+EH	Average	0.0985
Run 3	BoW+CM+WT+EH	Average	0.0976
Run 4	**BoW**	-	0.0935
Run 5	CM+WT+EH	Average	0.0524
Run 6	CM+WT	Average	0.0510

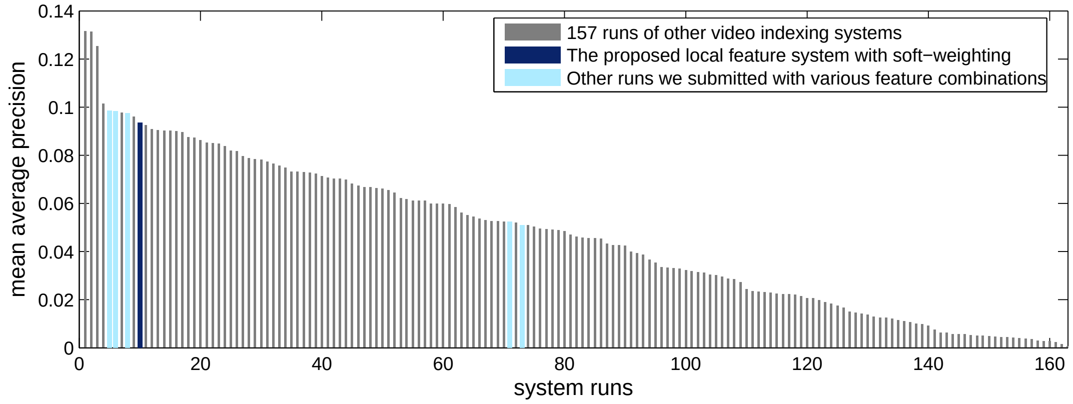

Figure 1: Overview of all video indexing runs submitted to TRECVID-2007, ranked according to mean average precision. Our BoW run is shown in dark blue, and our other official runs are shown in light blue.

classification, the two categories of keypoints are combined using kernel fusion. For a concept C_i, two kernel matrices K_1 and K_2 are generated for feature F_1 and F_2 using the popular χ^2 kernel. Finally a fused kernel K is obtained through the weighted combination of K_1 and K_2:

$$K = \alpha K_1 + (1 - \alpha) K_2. \tag{1}$$

In our experiments, α was empirically chosen as 0.5, while the optimized solution of α can be derived by the kernel alignment principle proposed in [6].

In addition to BoW, we also describe the location distribution of local keypoints under a multi-resolution grid representation [7], forming another feature called LIP-D.

1.3 Results and Analysis

Table 1 lists the summarization of our submitted runs. The evaluation results show that the local feature system, i.e. BoW, could offer a very competitive MAP as also shown in Figure 1. This indeed proves that the BoW approach, represented using the proposed soft-weighting scheme, is highly effective for concept detection.

The improvement of fusion local feature and color/texture features is about 5%. This is

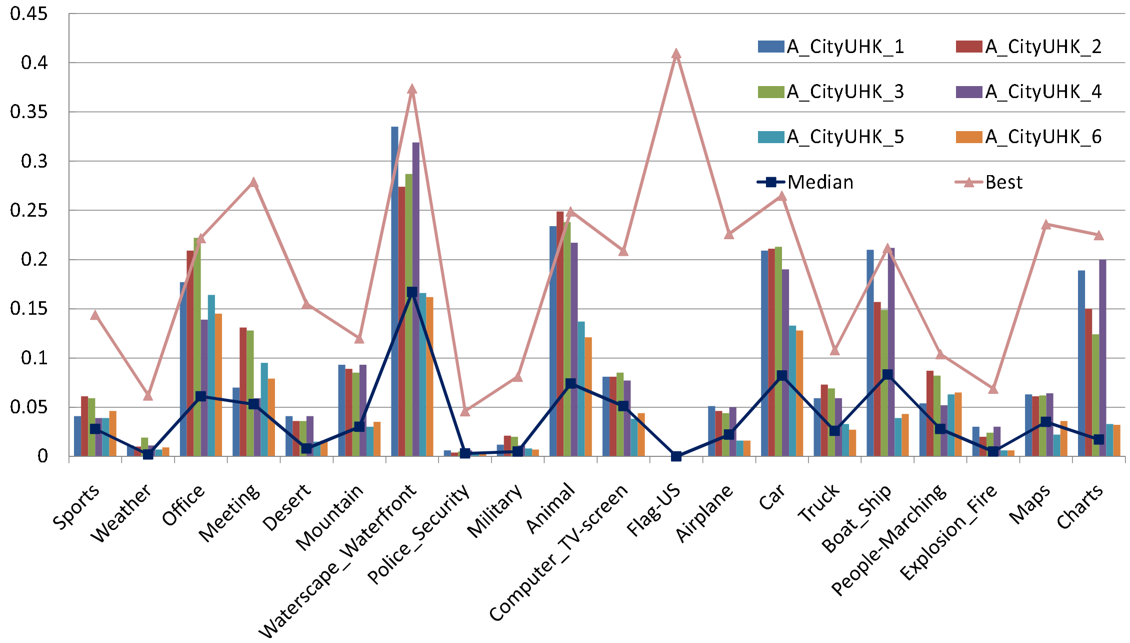

Figure 2: Performance of our submitted runs for each concept *vs.* median and best performance of all submitted runs.

much lower than our results on last year's data set (around 50%) [1]. This may be partially due to the fact that this year's data contain no commercials and less duplicates, resulting in worse performance of the global color/texture features.

A detailed per-concept analysis with the best and median of all submitted runs is shown in Figure 2. Our BoW (run 4) performs better/comparable than the baseline features for most concepts except "sports", "office", "meeting", and "people_marching". This may be because these four concepts generally appear in the whole keyframes (in contrast to the object concepts which can appear anywhere), which can be well captured by the global baseline features. Furthermore, among all the 163 submitted runs, we perform best for 3 concepts "office", "animal", and "boat_ship".

2 Automatic Video Search

2.1 Text Baseline

In text search, we only employ the English output of ASR/MT [8]. Instead of traditional method like TF-IDF, our text search model adopts Okpai [9] to index the transcripts. Based on our previous experiments, synonyms usually hurt the performance, so our text search model only uses the original noun query items as query input to avoid the negative affect of irrelevant words. The application interface provided by Lemur [10] was used.

2.2 Concept-based Video Search in Semantic Space

Concept-based video retrieval has recently attracted a new spurt of research attention, attributed to its potential in bridging semantic gap. In this experiment, for each query, we focus on the selection of top-k related semantic visual concepts with an Orthogonal Ontology-enriched Semantic Space (OS^2) to disambiguate query-concept relationship.

2.2.1 Concept Vocabulary

We use the VIREO-374 [1, 2] shared with the TRECVID community as the basic concept vocabulary for selection. The VIREO-374 uses three features, i.e. local feature, CM, and WT, to train a set of 374 semantic concepts using LSCOM annotation [11]. The VIREO-374 performs very good on TRECVID-2006 test set, yielding a MAP of 0.154 on the 20 semantic concepts selected during TRECVID-2006 HLFE evaluation. As all the LSCOM annotations are on the old news videos, we replace the 36 LSCOM-lite concepts using our this year's submission (Run 1 in Section 1). Although this year's data is quite different from news videos as in previous years, we expect that a huge amount of detectors trained on news videos are still useful for concept based search. In the experiments, we removed those concepts without definitions in WordNet [12], resulting in a smaller set of 242 concepts.

2.2.2 Concept Selection Using OS^2

In [13], an OSS (ontology-enriched semantic space) was built to ensure globally consistent comparison of semantic similarities. OSS is a space learnt by modeling the ontological interrelatedness of concepts. The concepts are clustered according to their closeness. One concept from each cluster is then selected to form the bases of the semantic space. Basically, OSS can be view as the "vantages" of the original ontology space, where the vantages are the bases formed by real concepts which are representative but not strictly orthogonal. Since the basis axes are not strictly orthogonal to each other, the semantic similarity will somewhat be biased. We improve OSS by constructing an Orthogonal Ontology-enriched Semantic Space (OS^2), as an extension of OSS. In contrast to OSS, OS^2 performs spectral decomposition to transform the semantic space into a novel space with orthogonal bases. The bases in OS^2 are not formed by the real concepts. They are, however, more powerful in terms of expressive and generalization abilities for having the orthogonal property which optimally covers the semantic space. Thus a more consistent way of comparing concept similarity scores can be guaranteed in OS^2. The OS^2 also facilities the concept selection within a computable linear space. With OS^2, we are able to represent each semantic concept or query item as a vector. Concept selection could be easily done by selecting top-k nearest neighbors of a query item. For multiple concept selection, we use linear fusion to fuse multiple detectors, weighted by concept-query similarities.

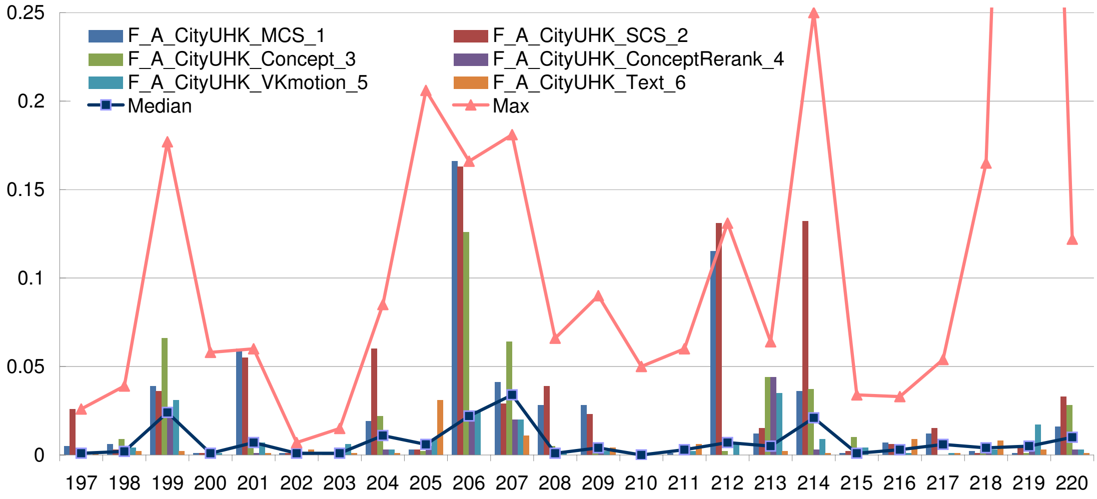

Figure 3: Performance of our automatic search runs *vs.* median and best performance of all submitted runs.

2.3 Query-by-Example (QBE)

Our QBE search CityUHK-Concept is based on concept detection. For each image in the query, we employ the 36 concept detectors in Section 1 to compute a vector $V_{concept} = \{c_1, c_2, \cdots, c_{36}\}$, where $c_i (i = 1, 2, \cdots, 36)$ is the confidence value that the image contains concept i. The Euclidean distance between concept vectors of query images and keyframes in the test set are calculated for retrieval.

2.4 Reranking

This year we develop two runs to explore the usefulness of reranking for improving search results. In one run CityUHK_ConceptRerank_4, we employ the concept vector used in Section 2.3 to rerank the result of text baseline. In the other run CityUHK_VKmotion_5, instead of selected keyframes, we employ motion information in an entire shot for search. Between neighboring frames in each shot, we track and compute a motion histogram of visual words which are constructed in Section 1. This motion histogram is then employed to measure the distance between the query shots and the test videos to rerank the result of text baseline.

2.5 Results

We submitted 6 runs of automatic search task this year. The results are shown in Figure 3, in comparison with the max and median APs for each topic.

The first two runs are concept-based search, namely F_A_CityUHK_SCS_2 (SCS) which selects a single concept and F_A_CityUHK_MCS_1 (MCS) which selects top-3 concepts. Since our focus is mainly on semantic-based concept selection, we did not fuse the result with other features such as text baseline. We only use text queries, ignoring image and clip queries. The results of these two runs are listed in Table 2, together with the top-3 concepts selected from VIREO-374

Table 2: Comparison of two concept selection strategies in OS^2 for automatic video search. The best result is given in bold.

Topic	**Selected Detectors**			**Average Precision**	
	Top-1	Top-2	Top-3	SCS	MCS
197	People	Walking	Running	**0.026**	0.005
198	Observation Tower	Office Building	Hotel	0.003	**0.006**
199	Walking	Person	Bicycle	0.036	**0.039**
200	Hand	Computer	Cable	0.001	0.001
201	Canal	River	Riverbank	0.055	**0.060**
202	Person	Telephone	Talking	0.001	0.001
203	Street	Supermarket	Shopping Mall	0.001	0.001
204	Street	Parade	Alley	**0.060**	0.019
205	Bus	Vehicle	Bicycle	0.003	0.003
206	Mountain	Hill	Valley	0.163	**0.166**
207	Building	Office Building	Water	0.029	**0.041**
208	Nighttime	Street	Alley	**0.039**	0.028
209	People	Table	Furniture	0.023	**0.028**
210	Walking	Dog	Running	0.000	0.000
211	Horse	Dog	Bird	0.000	**0.001**
212	Boat	Canoe	Rowboat	**0.131**	0.115
213	Talking	Interview	People	**0.015**	0.012
214	Crowd	People	Group	**0.132**	0.036
215	Conference Room	Bathroom	Classroom	**0.002**	0.001
216	Bridge	Stadium	Observation Tower	0.006	**0.007**
217	Road	Vehicle	Highway	**0.015**	0.012
218	People	Group	Weapon	0.001	**0.002**
219	Construction Worker	Judge	Officer	**0.004**	0.001
220	Building	Street	People	**0.033**	0.016
MAP				*0.0325*	*0.0250*

(for SCS, the first concept is selected). Surprisingly, the run with single best concept selection performs better with a *MAP* of 0.0325. We analyzed the results and found that the major reason is due to the absence of good strategy to determine the number of concepts to be selected. Fixing the number of concepts to 3 actually hurts the performance by introducing irrelevant concepts. There are about eight queries having this problem (e.g., topic 215: *a classroom scene with one or more students*). The reliability of concept detectors, to certain extent, has seriously affected the search performance of a few queries (e.g., topic 200: *hands at a keyboard typing or using a mouse*). Meanwhile, some topics also need linguistic analysis to capture the query semantics (e.g., in topic 207: *waterfront with water and buildings*, for which our method will assign equal weights to *waterfront* and *building*). Another issue is that VIREO-374, currently containing only 374 concepts, is not enough to cover and interpret all the queries terms. There are five topics we did not find any appropriate detectors to support. For instance, in topic 211: *find shots with sheep or goats*, both "sheep" and "goat" are not included in VIREO-374. With the two runs of semantic-based concept selection, we get the highest AP for four query topics.

Another two runs for reranking improve the result of text baseline from 0.004 to 0.006 (CityUHK-Concept-Rank) and 0.008 (CityUHK-VKmotion-Rank) respectively. However, considering the poor performance of the baseline, we cannot draw a conclusion whether or by how much the reranking can improve the search result. More experiments are needed to test the usefulness of reranking by applying it to other runs such as SCS and MCS.

3 Conclusions

This year, our main focus is on the exploration of local features and bag-of-visual-words (BoW) for semantic indexing and automatic search. Our developed VIREO-374 are experimented for these purposes. Similar to our conclusion in [1, 2], we show that BoW, with careful representation choices, can be very competitive for high-level feature extraction.

In automatic search, we experiment VIREO-374 and orthogonal ontology-enriched semantic space OS^2 for concept-based video search. By using only text-query, the search methodology outperforms our other runs on keyframe-based query-by-visual-examples and sequence-based motion features. In view that there are about 50% of search queries involving events, we do not want to read too much into the *MAP* performance. In general, using keyframe-based detectors to answer event-based queries is still difficult. We will put more efforts in this aspect in future.

Acknowledgment

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 118906 and CityU 118905).

References

[1] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "VIREO-374: LSCOM semantic concept detectors using local keypoint features," in *http://vireo.cs.cityu.edu.hk/research/vireo374/*.

[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] S. Ayache and G. Quenot, "Evaluation of active learning strategies for video indexing," in *Proc. of fifth international workshop on content-based multimedia indexing*, 2007.

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

[5] K. Mikoljczyk and C. Schmid, "Scale and affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, pp. 63–86, 2004.

[6] G. Lanckriet, N. Cristinanini, P. Bartlett, L. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semi-definite programming," in *ICML*, 2002, pp. 323–330.

[7] Y.-G. Jiang, X.-Y. Wei, C.-W. Ngo, H.-K. Tan, W. Zhao, and X. Wu, "Modeling local interest points for semantic detection and video search at TRECVID 2006," in *TRECVID online proceedings*, 2006.

[8] M. Huijbregts, R. Ordelman, and F. de Jong, "Annotation of heterogeneous multimedia content using automatic speech recognition," in *Proceedings of the second international conference on Semantics And digital Media Technologies (SAMT)*, ser. Lecture Notes in Computer Science. Berlin: Springer Verlag, December 2007.

[9] S. E. Robertson and S. Walker, "Okapi/keenbow at trec-8," in *Text REtrieval Conference*, 2000, pp. 151–163.

[10] Lemur, "The lemur toolkit for language modeling and information retrieval," in *http://www.lemurproject.org/*.

[11] "LSCOM lexicon definitions and annotations," in *DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report #217-2006-3*, 2006.

[12] C. Fellbaum, *WordNet: an electronic lexical database.* The MIT Press, 1998.

[13] X.-Y. Wei and C.-W. Ngo, "Ontology-enriched semantic space for video search," in *ACM Multimedia*, 2007.