Beyond Semantic Search: What You Observe May Not Be What You Think

Chong-Wah Ngo, Yu-Gang Jiang, Xiaoyong Wei
Wanlei Zhao, Feng Wang, Xiao Wu and Hung-Khoon Tan

VIREO Research Group
Department of Computer Science, City University of Hong Kong
{*cwngo,yjiang,xiaoyong,wzhao2,fwang,wuxiao,hktan*}*@cs.cityu.edu.hk*
http://vireo.cs.cityu.edu.hk

Abstract

This paper presents our approaches and results of the four TRECVID 2008 tasks we participated in: *high-level feature extraction*, *automatic video search*, *video copy detection*, and *rushes summarization*.

In high-level feature extraction, we jointly submitted our results with Columbia University. The four runs submitted through CityU aim to explore context-based concept fusion by modeling inter-concept relationship. The relationship is modeled not based on semantic reasoning, but by observing how concepts correlate to each other, either directly or indirectly, in LSCOM common annotation [1]. An observability space (OS) [2] is thus built on top of LSCOM [1] and VIREO-374 [3] for performing concept fusion. Since 19 of the 20 concepts evaluated this year appeared in VIREO-374, we apply OS to re-rank the results of both old models from VIREO-374 and new models from a joint baseline submission with Columbia.

- A_CityU-HK1: re-rank A_CU-run5 using OS – both positive and negative correlated concepts are used.
- A_CityU-HK2: re-rank A_CU-run5 using OS – only positive correlated concepts are used.
- A_CityU-HK3: re-rank old models from VIREO-374 using OS – both positive and negative correlated concepts are used.
- A_CityU-HK4: re-rank old models from VIREO-374 using OS – only positive correlated concepts are used.

In automatic search, we focus on concept-based video search. The search is beyond semantic reasoning, where we consider the fusion of detectors using concept semantics, co-occurrence, diversity, and detector robustness. Two runs are submitted based on the works in [2] and [4] respectively.

- F_A_2_CityUHK1_1: multi-modality fusion of concept-based search (Run-2), query example based search (Run-4 and Run-5), and text baseline (Run-6).

- F_A_2_CityUHK2_2: concept-based search by fusing semantics, observability, reliability and diversity of concept detectors [2].
- F_A_2_CityUHK3_3: concept-based search using semantics reasoning [4, 5].
- F_A_2_CityUHK4_4: query-by-example – using VIREO-374 detection scores as features.
- F_A_2_CityUHK5_5: query-by-example – using motion histograms as features.
- F_A_1_CityUHK6_6: text baseline.

In content-based video copy detection, we adopt a recently proposed near-duplicate video detection method [6, 7] based on the matching of local keypoint features. We submitted three runs:

- CityUHK_loose: we use cosine similarity of visual word histograms to generate candidate near-duplicate keyframe set. The set is further filtered by a recently proposed method called SR-PE [6].

- CityUHK_vkisect: same with CityUHK_loose except that we use histogram intersection instead of cosine similarity for candidate keyframe set generation.

- CityUHK_tight: similar to CityUHK_loose, but we add in few more heuristical constraints.

In BBC rushes summarization, we submitted one run using the same method with our last year's submission [8].

1 High-Level Feature Extraction (HLFE)

This year, we jointly submitted our HLFE results with Columbia University. Detailed descriptions of the joint submissions can be found in the notebook paper of Columbia [9]. For the four runs submitted by CityU, we aim to test context-based concept fusion based on a linear space (observability space) built from the observation derived from manual concept annotation.

1.1 Concept Fusion with Observability Space

The observability space (OS) is proposed to effectively model the co-occurrence relationship among concepts [2]. We refine the individual concept detectors by using simple and efficient linear weighted fusion of the target concepts with several peripherally related concepts, where both concept selection and fusion weights are determined by the OS.

Given a concept set V of n concepts, we first construct a $n \times n$ concept observability matrix $\mathbf{R}$ where each entry r_{ij} represents the co-occurrence relationship of a concept pair (C_i, C_j), measured by Pearson product-moment (PM) correlation:

$$r_{ij} = PM(C_i, C_j) = \frac{\sum_{k=1}^{|\mathcal{T}|}(O_{ik} - \mu_i)(O_{jk} - \mu_j)}{(|\mathcal{T}| - 1)\sigma_i\sigma_j} \quad (1)$$

where O_{ik} is the observability of concept C_i in shot k, and μ_i and σ_i are the sample mean and standard deviation, respectively, of observing C_i in a training set $\mathcal{T}$. We set O_{ik} to 1 if C_i presents in shot k, and 0 otherwise.

With $\mathbf{R}$, basis vectors $\mathbf{C}$ of OS can be estimated by solving following equation

$$\mathbf{C}^T\mathbf{C} = \mathbf{R}. \tag{2}$$

The above equation can be solved by performing spectral decomposition to $\mathbf{R}$. Note that the basis vectors in $\mathbf{C}$ are orthogonal. After generating the basis vectors, a concept C_i can be represented by a vector $\vec{C_i}$ in OS:

$$\begin{aligned} \mathbf{C}^T\vec{C_i} &= \vec{R_i} \\ \vec{C_i} &= (\mathbf{C}^T)^{-1}\vec{R_i}. \end{aligned} \tag{3}$$

where $\vec{R_i}$ is a vector obtained by computing the PM of C_i to all concepts in V. With this representation in OS, the observability score of two concepts is computed as

$$Observability(C_i, C_j) = \frac{\vec{C_i} \cdot \vec{C_j}}{|\vec{C_i}||\vec{C_j}|}. \tag{4}$$

An important property of OS is the offering of globally consistent space for observing the co-occurrence among concepts. The concepts are projected and represented as vectors in OS. The observability of two concepts is not simply based upon PM correlation, but also the observability of these two concepts with respect to the orthogonal bases computed based on the matrix $\mathbf{R}$. In other words, comparing observability of any two concepts is globally, instead of locally, measured in OS. In this way, the erroneous effect due to the missing annotation can be minimized. In fact, missing annotation commonly happens for instance in LSCOM annotation. A good example is that *snow* is not labeled together with *outdoor* in some sample shots by annotators. By employing a vector space as OS, the co-occurrence probability of *snow* and *outdoor* can still be discovered if *snow* is happened annotated together with *mountain*, and *mountain* is always to be labeled with *outdoor* in some sample shots. OS provides a novel view of concept correlation in a linear space, beyond what can be inferred from the semantic space that we employed in TRECVID 2007 [10]. For example, *car* appears together with *road*, *building* appears together with *window*.

Within OS, for each concept C_i, a set of positive correlated concept P is formed by selecting concepts with higher observability scores calculated using Eqn 4. Similarly, a negative correlated concept set N is generated containing the concepts with lower observability scores. Let S_{C_i} be a vector of prediction scores of concept C_i on a test data set, P and N is used to refine the S_{C_i} as follows:

$$\hat{S}_{C_i} = S_{C_i} + \sum_{C_j \in P} \lambda_{ij} S_{C_j} - \sum_{C_j \in N} \lambda_{ij} S_{C_j} \tag{5}$$

where λ_{ij} is the observability score of concept C_i and C_j; $\hat{S}_{C_i}$ is the prediction scores after concept fusion with OS.

Table 1: Components and performance of the submitted runs.

Run ID	*Baseline*	*Baseline Devel. Data*	*# Pos/Neg concepts selected by OS*	*# of concepts with improved AP*	*MAP*
CityU-HK1	A_CU-run5	TV'08 Devel	3/2	11	0.1548
CityU-HK2	A_CU-run5	TV'08 Devel	3/0	11	0.1561
Baseline-1	A_CU-run5	TV'08 Devel	-	-	0.1618
CityU-HK3	VIREO-374	TV'05 Devel	3/2	16	0.0598
CityU-HK4	VIREO-374	TV'05 Devel	3/0	16	0.0590
Baseline-2	VIREO-374	TV'05 Devel	-	-	0.0386

1.2 HLFE Results and Analysis

We conduct concept fusion experiments based on VIREO-374 detector set [3]. Except *two_people*, all the other 19 concepts are in the VIREO-374 detector set. The PM correlations between each pair of the 374 concepts are calculated from their labels on the broadcast news videos in TRECVID-2005 data set. Since we cannot accurately estimate the correlation of *two_people* with the other concepts, we only perform concept fusion for the other 19 concepts and in all the four submitted runs, we use the original prediction scores of *two_people* from A_CU-run5.

Table 1 summarizes the components of each submitted run. We used two baselines for concept fusion: 1) A_CU-run5 which is composed of 5 SVMs trained using various global features and local keypoint features [1]; 2) prediction scores using the old models from VIREO-374 which were trained on TRECVID 2005 development data. For each target concept, we empirically select three positive correlated concepts and two negative correlated concepts using OS. Based upon each baseline, we submitted two runs – respectively update the baseline with positive concepts only (Run-2 and Run-4) and both positive and negative concepts (Run-1 and Run-3).

Figure 1 shows the mean average precision (MAP) of all type-A submissions. It is interesting to see that our Run-3 and Run-4 lie in the upper half of the 161 submissions, though no training data from TRECVID-2008 was used. The OS-based concept fusion method performs very well on baseline-2 by improving 53% using positive concepts only and 55% using both positive and negative concepts (cf. Table 1). However, no improvement is observed from the other two runs (Run-1 and Run-2) generated upon baseline-1. We believe this is due to the problem of normalizing the detection scores from new model (A_CU-runs) and old model (VIREO-374). Both models provide different distribution of detection scores, which results in difficulty in combining the detectors even if the weights of detectors can be properly predicted with OS. We will study this issue which is related to the fusion of multiple detector sets, from which each set is trained using different models, data and features. Per-concept performance of our submitted runs is shown in Figure 2.

[1] For the ground-truth labels, we used both the collaborative annotation and the MCG-ICT-CAS annotation [11]

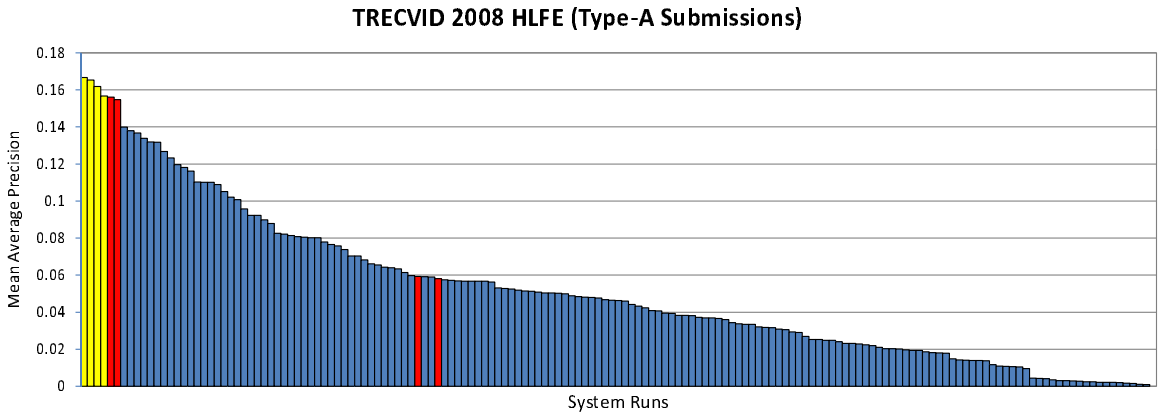

Figure 1: Mean average precision of all 161 type-A runs submitted to TRECVID-2008. The joint submissions with Columbia are shown in Yellow, and the four OS-based concept fusion runs submitted from CityU are shown in Red. Note that for the two runs with MAP around median, no training data from TRECVID 2008 were used.

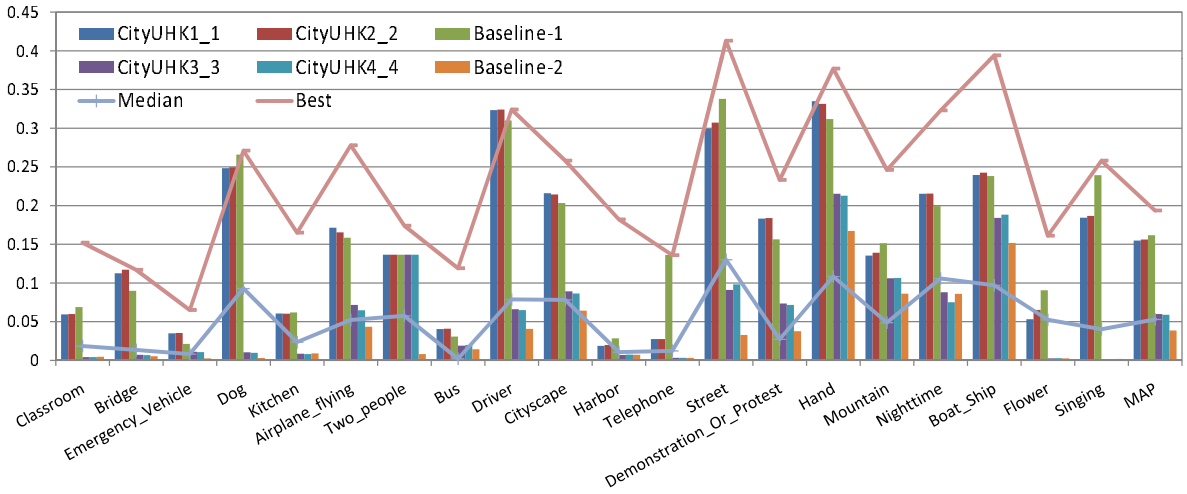

Figure 2: Per-concept performance of our submitted runs and two baselines. Median and best performances of all submitted runs are shown using lines.

Table 2: Positive and negative concepts suggested by OS.

Target	Positive Concepts	Negative Concepts
Classroom	Boy, Child, Asian People	Outdoor, Daytime Outdoor
Bridge	Road Overpass, Highway, River Bank	Person, Face
Emergency_Vehicle	Ground Vehicles, Car, Vehicle	Face, Civilian Person
Dog	Animal, Animal_Pens_And_Cages, Commercial AD	Adult, Person
Kitchen	Food, Dining Room, Room	Outdoor, Daytime Outdoor
Airplane_Flying	Airplane, Helicopter Hovering, Bicycle	Person, Adult
Two People	—	—
Bus	Ground Vehicles, Vehicle, Windows	Single Person, Person
Driver	Car, Vehicle, Windows	Ties, News Studio
Cityscape	Office Building, Urban Scenes, Tele. Tower	Civilian Person, Person
Harbor	Boat_Ship, Waterways, Lakes	Face, Person
Telephone	Cell Phones, Landlines, Cordless	News Studio, Studio
Street	Road, Car, Ground Vehicles	Face, Adult
Demon._Or_Protest	Protesters, People-Marching, People_Marching	Individual, Commercial AD
Hand	Body Parts, Attached Body Parts, Pipes	Face, Adult
Mountain	Hill, Landscape, Sky	Face, Adult
Nighttime	Moonlight, Outdoor, Sidewalks	Daytime Outdoor, Sitting
Boat_Ship	Waterways, Ship, Waterscape_Waterfront	Face, Adult
Flower	Furniture, Meeting, Interview On Location	Outdoor, Commercial AD
Singing	Dancing, Celebrity Entertainment, Entertainment	Outdoor, Male Person

Table 2 shows the positive and negative concepts selected by OS. As shown in the table, a good example is that a positive concept set {*Airplane, Helicopter_Hovering, Bicycle*} is selected for concept *Airplane_Flying*, with which the performance is improved from 0.0434 (Baseline-2) to 0.0646 (Run 4). When a negative concept set {*Person, Adult*} is in use, the performance is further boosted to 0.0717 (Run-3). However, this is not always the case for the other concepts. For example, *furniture* and *meeting* are selected for concept *flower*, although intuitively they are not really related to *flower*. This may be due to the fact that we only have 374 concepts in OS and thus sometimes irrelevant concepts may be selected. Instead of using fixed top-k concepts in positive set P, using an adapted number could be a possible way to improve our current method. On the other hand, the reliability of the concept detectors could be another important factor that should be considered in our future work. For example, *Food* is a reasonable selection for *Kitchen*, but the detector *Food* is highly unreliable and using it for concept fusion will hurt the performance.

2 Automatic Video Search

We experiment concept-based, query-by-example based and text-based video search. Concept-based and text-based search use text queries, while query-by-example uses image and video examples as queries. We submit six runs: two runs based on concept search, two runs based on

query-by-example, one run as text baseline, and one run basically fuses the results from concept, query-by-example and text searches.

2.1 Concept-Based Video Search

The submitted runs (run 1-3) are based on our recent works in [2]. We consider four aspects of detectors: semantics, reliability, observability and diversity for concept-based search. Two spaces are built for this purpose: semantic space (SS) and observability space (OS) mentioned in Section 1.1. SS and OS, respectively, model the semantic and co-occurrence relationships of concepts. Inside both spaces, each detector is treated as a vector. A multi-level fusion strategy is employed to novelly combine detectors, allowing the enhancement of detector reliability while enabling the observability, semantics and diversity of concepts being utilized for query answering. For details, please refer to [2].

2.1.1 Concept Selection

Concept selection is performed in three major stages. First, each query term is mapped to SS as a vector. By cosine similarity, one detector is picked and assigned to each query term. We call these detectors "Anchor concepts", denoted as $\mathcal{A}$. Second, the detectors in $\mathcal{A}$ are mapped to OS. From OS, another set of detectors which reside in the subspaces formed by anchor concepts are mined. We call these detectors "Bridge concepts", denoted as $\mathcal{B}$. Third, for each bridge concept, we further select the set of positively correlated detectors (denoted as $\mathcal{P}^+$) and negatively correlated detectors (denoted as $\mathcal{N}^-$). Together, the four sets of detectors ($\mathcal{A}$, $\mathcal{B}$, $\mathcal{P}^+$, $\mathcal{N}^-$) form the concept pool to answer queries. Note that the number of detectors ultimately selected is different from query to query. This is very different from our work in TRECVID 2007 [10], where each query is assigned one or three most similar detectors based on SS.

2.1.2 Multi-Level Detector Fusion

We experiment multi-level fusion of detectors, based on detector reliability, observability, semantics and diversity. At the lowest level, the detectors in $\mathcal{P}^+$ and $\mathcal{N}^-$ are fused using Equation-5 to enhance the reliability of detectors in B. The enhanced detectors are further linearly fused to support the detectors in A. The fusion weights are derived directly from OS. Finally, considering the semantics and diversity of concepts, the detectors in A are fused again to determine the final scores of shots. Interested readers please refer to [2] for the details of fusion algorithm.

2.2 Query-by-Example (QBE) Based on VIREO-374 Concept Scores

In addition to selecting appropriate detectors, we also experiment a very commonly used strategy for concept-based video search. In this strategy, each shot is represented as a vector of 374 dimensions. Each component in the vector represents the detection score of a detector, indicating the likelihood of finding a particular concept in the shot. In the submitted run (Run 4), the detection scores is derived based on VIREO-374 detector set [3]. Cosine similarity is used to

measured the similarity between shots and queries. We used both image examples and video examples as query in this run. For video examples, we uniformly extracted one keyframe per second.

2.3 QBE Based on Motion Histogram

Since some queries are motion or event driven, the submitted run (Run 5) aims to test the feasibility of using motion for video search. The feature we consider is grid-based directional motion histogram. The histogram is formed based on the motion vectors extracted directly from MPEG compressed domain. Basically each frame is divided into 5×5 grid, and a 4-directional motion histogram is computed for each grid. By accumulating the motion histograms over frames, each shot is represented as a feature vector of 100 dimensions. Cosine similarity is then employed for measuring the similarities between shots and queries. In this run, we only use video examples as query.

2.4 Query-Class Dependent Fusion

The submitted run (run 1) aims to fully leverage the available modalities (text, image and video examples, concept detectors, motion features) for query answering. This run basically fuse the previous four submitted runs based on concepts, motion and text. The underlying question is how to effectively fuse different modalities. For instance, motion features should be assigned higher weight when answering event-driven queries. We adopt query-class dependent fusion based on our prior work in [4], where the fusion weights are determined based on fuzzy synthetic evaluation (More details in [4]).

2.5 Text Baseline

In text search, we only employ the English output of ASR/MT [12]. Instead of traditional method like TF-IDF, our text search model adopts Okpai [13] 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 [14] is used.

2.6 Search Results and Analysis

We submitted six automatic search runs. The component and MAP of each submitted run is shown in Table 3. Among the six runs, concept based search using our work in [2] performs best with $MAP = 0.0424$ (Run-2). Considering the fact that only text queries are used in this run, it is pretty interesting to see that this run already exhibits strong performance compared to other submissions (see Figure 3). This confirms the effectiveness of our multi-level detector fusion framework for concept-based video search, which cleverly utilizes the semantics, reliability, observability, and diversity of the VIREO-374 detectors (36 detectors were replaced with models trained on TV'07 development data). Compared to another concept-based search run (Run-3)

Table 3: Component and MAP of each submitted automatic search run.

Run ID	Description	*MAP*
Run 1	Query-dependent fusion	*0.0406*
Run 2	Concept-based search (beyond semantics) [2]	*0.0424*
Run 3	Concept-based search (semantics only) [4, 5]	*0.0388*
Run 4	QBE: VIREO-374 concept scores	*0.0099*
Run 5	QBE: Motion histogram	*0.0039*
Run 6	Text only	*0.0081*

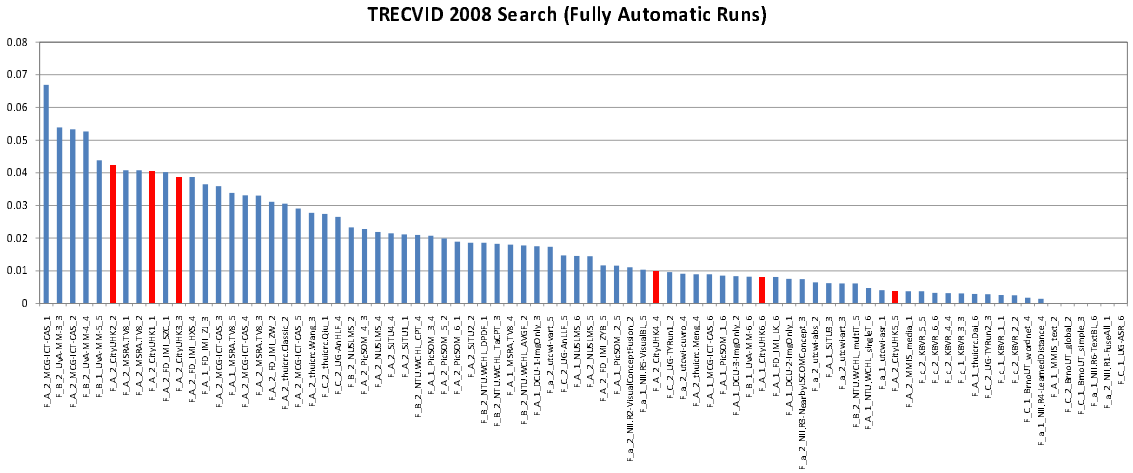

Figure 3: Mean average precision of all fully automatic runs submitted to TRECVID-2008.

which only used semantic reasoning, our new framework can improve 9.3% in term of MAP. For the query-by-example submissions, our Run-4 based on VIREO374 detection scores performs better than Run-5 using motion histogram. This may be because in Run-4 both query images and video clips were used, while in Run-5 motion feature can only be extracted from the query video clips. The query-class dependent fusion (Run-1) is not able to improve the MAP. This is because the reliability of each feature component is not considered in the fusion step. For example, in topic-239 "*Find shots of one or more people standing, walking, or playing with one or more children*", the query-dependent fusion assigned the highest weight for the motion based QBE since this query is related to event. However, the motion-based QBE is not reliable for this query (cf. Figure 4). Thus, knowing how to measure the reliability of each component could be a possible direction to improve query-dependent fusion.

The detailed per-query AP is shown in Figure 4. Among all the automatic search submissions, we achieve the highest AP for 8 out of 48 queries. For instance, our concept-based search run-2 get the highest AP for topic-254 "*Find shots of a person talking behind a microphone*". For this query, when the semantic reasoning is in use (run-3), three concepts are selected: *person*, *talking* and *microphone*. In addition to these three concepts, our method in Run-2 is able to select two more positive concepts *individual* and *face*, two negative concepts *sky* and *weapon*, and one bridge concept *speaker*. This again confirms the effectiveness of our approach in [2].

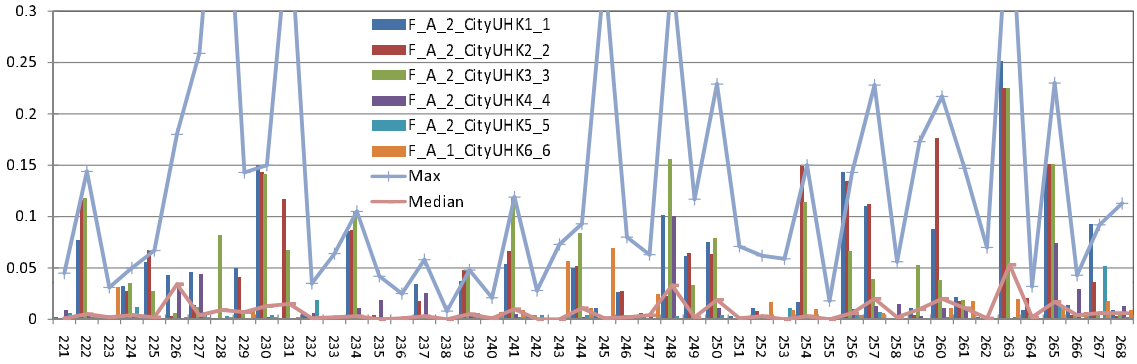

Figure 4: Per-query performance of our submitted runs vs. median and best performance of all fully automatic runs.

3 Video Copy Detection

Considering that video copy can be treated as a special case of near-duplicate (ND) video, we extend our prior work in [6] for video copy detection. Figure 5 shows the framework for this task. The framework is composed of two parts: off-line quantization and online detection.

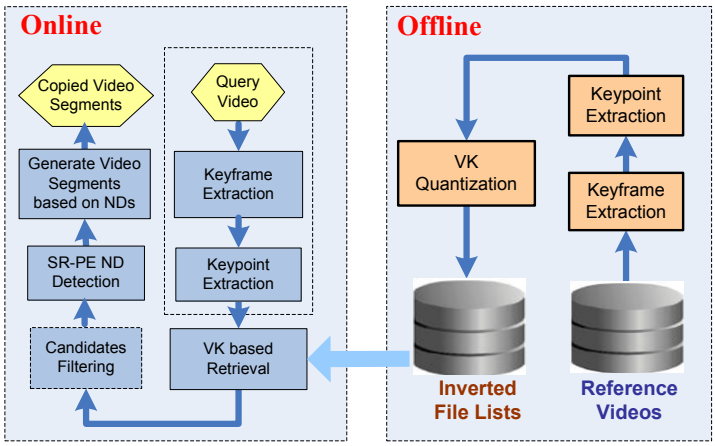

Figure 5: Video Copy Detection Framework.

OFFLINE At the offline stage, keyframes are first extracted from each reference video. Keypoints are detected by DoG (Difference of Gaussian) [15] and described by PSIFT [6], which is a compact representation of SIFT [15]. With the keypoints, a visual vocabulary of $10,000$ visual words is constructed by k-means. In the off-line quantization stage, keypoints of each keyframe are quantized based on the visual vocabulary. In the end, each keyframe is represented as a visual word histogram. To facilitate efficient online retrieval, the visual word histograms are stored using inverted lists.

ONLINE Similar to the offline stage, keyframes and keypoints are first extracted from each query video segment. Each keyframe is described using a visual word histograms. Then fast retrieval based on the inverted lists is performed to retrieve top-k candidate keyframes for each keyframe in the reference video database. After that a recently proposed pattern learning

method, SR-PE [1], is employed to accurately detect ND pairs within the candidate set. To reduce the candidate pairs for ND detection, a candidate filtering step is adopted to prune the unnecessary comparison. The filtering is based on the fact that keyframes are densely extracted from query videos and the neighboring keyframes should be visual similar. Thus we remove candidate keyframes that only appear in the candidate set of one query keyframe within a temporal window. Finally, the matched ND pairs are organized into segments. For each detected copied segment, a confidence score is given based on the frame level similarity.

3.1 CBCD Results and Analysis

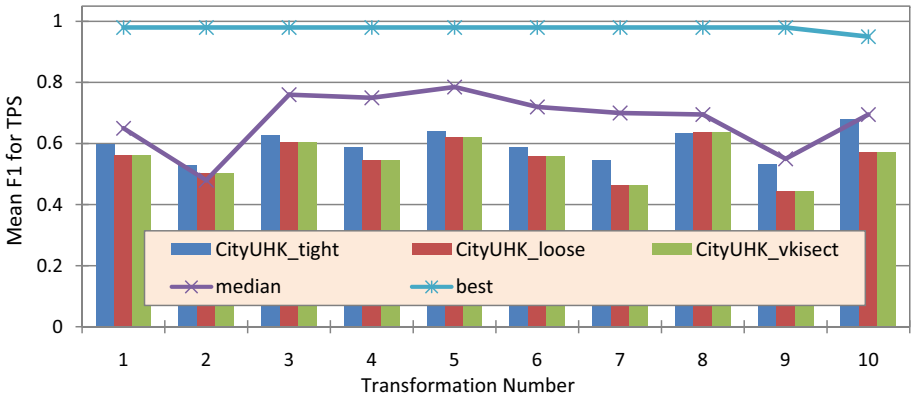

Figure 6: Performance of the three submitted CBCD runs.

We submitted three CBCD runs:

- CityUHK_loose: we use cosine similarity of visual word histograms to generate candidate near-duplicate keyframe set. The set is further filtered by a recently proposed method called SR-PE [6].

- CityUHK_vkisect: same with CityUHK_loose except that we use histogram intersection instead of cosine similarity for candidate keyframe set generation.

- CityUHK_tight: similar to CityUHK_loose, but we add in few more heuristic constraints.

Figure 6 shows the F1 performances with respect to each of the 10 transformations. The performance of our submitted runs is lower than median. Among the three runs, "tight" performs better than the other two runs for 9 out of 10 transformations, which indicates that the filtering step could eliminate some false alarms from visual word based retrieval. The performance of "vkisect" run is almost the same to that of the "loose" run, showing that histogram intersection and Cosine similarity are similar.

There are several possible reasons that our approach does not perform very well. Firstly, since our approach is based on local keypoints, it can be easily affected by transformations such as resolution changes. This is due to the fact that keypoints are usually detected around corners and thus may not be found when a frame is highly blur. Secondly, our visual word based retrieval

step is not suitable for the "picture in picture" transformation, since we quantize all keypoints from a frame to form a feature vector and the huge portion of background irrelevant keypoints will dramatically affect the retrieval performance. Finally, we only extracted 170,000 keyframes from the 200 hours reference video data set (1 frame from every four seconds), which may be too small for this task and could be a bottleneck where some copied segments are simply missed by this uniform and coarse keyframe sampling scheme.

4 BBC Rushes Summarization

4.1 Approach

We submitted 1 run for the task of BBC Rushes Summarization, where we employed the same algorithm as last year. The details can be found in [8]. Since the summary duration is limited to 2% instead of 4% of the original video as in 2007, we adapt our algorithm by simply sampling 1 frame from every 2 frames.

4.2 Rush Summarization Results

Table 4 shows our results for the 8 criteria used in the summary evaluation. For the 4 criteria DU, XD, TT, and VT which measures the summary duration and play time, the results are similar to last year and slightly better than the average of all submissions. For IN (fraction of inclusions found in the summary), we also get a score (0.65) similar to 2007 (0.64), which is significantly higher than the average (0.44). However, the baseline (0.83) is ranked 1st for this criterion. This is probably because the baseline is generated by sampling, which covers almost all the content in the original video. Different from TRECVID 2007, two criteria (JU and RE) are used to measure the redundancy of the summary in this year's evaluation. For JU (summary contains lots of junk), we get a relatively good result. However, for RE (summary contains lot of duplicate video), our results are just close to the average. So, our approach for junk shot removal is effective. At the same time, some repetitive video clips are not identified. For TE (summary has a pleasant tempo/rhythm), our results are not good as last year, where we were ranked 1st for a similar criterion (EA). One important reason is that we sped up the play of the video in order to meet the requirement on the summary duration. As a result, the average length of each selected video clip is shortened to only 0.5 second and this reduces the enjoyability of the video summary. Among these criteria, there is a tradeoff between IN and others. We can achieve good results for IN and JU while most of other criterions are still better or similar to the average.

5 Conclusions

This year, our major aim is to study the appropriate utilization of bag-of-words (BoW) and bag-of-concepts (BoC) for search related tasks. For BoW, similar to last few years [10, 16],

Table 4: Results of BBC Summarization Task

Criterion	**DU**	**XD**	**TT**	**VT**	**IN**	**JU**	**RE**	**TE**
Mean score of 43 runs	27.11	4.60	41.20	29.36	0.44	3.16	3.27	2.73
Baseline score	31.31	0.40	59.59	31.36	0.83	2.66	2.02	1.44
Our score	23.41	8.30	39.52	24.96	0.65	3.58	3.24	2.74
Baseline rank	35	35	43	25	1	41	43	43
Our rank	10	10	15	9	5	3	25	24

we again demonstrate the strong performance of BoW (MAP=0.157; in joint submission Run-6 with Columbia). Even when we only use the TRECVID 2005 as training data (except that for *two_people*, 2008 training data is used), the BoW still performs better than the median performance. By further considering context-based fusion using observability space, we demonstrate some improvements (though this part is yet to be further investigated). For BoC, we demonstrate the power of beyond semantic search, where the four properties of bag-of-concepts (semantics, observability, diversity, reliability) are jointly considered for fusing detectors. An interesting fact is that when considering observability space for selecting concepts, either for high-level feature extraction or automatic search, some selected concepts often surprise us, where the fact of "what we observe may not be what we think" is always true. This indeed in line with several recent studies where different aspects of concepts (e.g., statistical properties, detector robustness) play important roles for concept-based video search. One aspect worth to pinpoint is that: using only text-based queries and relying only on a small pool of detectors (VIREO-374), relatively high MAP is achieved compared to other runs submitted this year. Of course, the MAP (of less than 0.05) is still hard for us to make any concrete conclusion on our current findings.

For copy detection, we learn the lesson that copy detection could be very different from near-duplicate (ND) detection. The fact that ND imposing "less restrict" detection conditions can easily result in false alarms. Other clues such as keypoint/motion trajectories could be better features for copy detection, though these clues may not be directly applicable to ND detection. For rush summarization, a simple re-run of our TRECVID 2007 system still shows that our performance is reasonable. While we obtain relatively high performance in IN and JU, it seems difficult to equally balance other criteria especially when the compression ratio is high (2% of the original video).

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] "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.

[2] X.-Y. Wei and C.-W. Ngo, "Fusing semantics, observability, reliability and diversity of concept detectors for video search," in *ACM Multimedia*, 2008.

[3] 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/*.

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

[5] X.-Y. Wei, C.-W. Ngo, and Y.-G. Jiang, "Selection of concept detectors for video search by ontology-enriched semantic spaces," *IEEE Trans. on Multimedia*, 2008.

[6] W.-L. Zhao and C.-W. Ngo, "Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection," *IEEE Trans. on Image Processing*, 2008.

[7] C.-W. Ngo, W.-L. Zhao, and et al., "Near-duplicate keyframe detectoin based on interest point matching," in *http://vireo.cs.cityu.edu.hk/research/NDK/ndk.html/*.

[8] F. Wang and C. W. Ngo, "Rushes video summarization by object and event understanding," in *TRECVID BBC Rushes Summarization Workshop at ACM Multimedia*, 2007.

[9] S.-F. Chang, J. He, Y.-G. Jiang, E. El Khoury, C.-W. Ngo, A. Yanagawa, and E. Zavesky, "Columbia university/vireo-cityu/irit trecvid-2008 high-level feature extraction and interactive video search," in *TRECVID workshop*, 2008.

[10] C.-W. Ngo, Y.-G. Jiang, X.-Y. Wei, F. Wang, W. Zhao, H.-K. Tan, and X. Wu, "Experimenting VIREO-374: Bag-of-visual-words and visual-based ontology for semantic video indexing and search," in *TRECVID workshop*, 2007.

[11] S. Tang, J. Li, and et al., "Trecvid 2008 high-level feature extraction by mcg-ict-cas," in *TRECVID workshop*, 2008.

[12] 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. LNCS. Berlin: Springer Verlag, December 2007.

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

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

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

[16] 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.