VIREO/DVMM at TRECVID 2009: High-Level Feature Extraction, Automatic Video Search, and Content-Based Copy Detection

Chong-Wah Ngo[†], Yu-Gang Jiang[†‡], Xiao-Yong Wei[†], Wanlei Zhao[†], Yang Liu[†], Jun Wang[‡], Shiai Zhu[†], Shih-Fu Chang[‡]

[†] *Video Retrieval Group (VIREO), City University of Hong Kong*
[‡] *Digital Video and Multimedia Lab (DVMM), Columbia University*
http://vireo.cs.cityu.edu.hk *http://www.ee.columbia.edu/dvmm*

Nov 25, 2009

Abstract

This paper presents overview and comparative analysis of our systems designed for 3 TRECVID 2009 tasks: high-level feature extraction, automatic search, and content-based copy detection.

High-Level Feature Extraction (HLFE):

Our main focus for the HLFE task is on the study of a new method named domain adaptive semantic diffusion (DASD) [1], which exploits *semantic context* (concept relationship) while also considers the *domain-shift-of-context* to improve concept detection accuracy. We apply our TRECVID 2008 HLFE system [2] to construct baseline detectors for the 20 evaluated concepts, where both local and global features are explored. Evaluation results show that our 2008 system is still able to produce strong performance (Run 5: MAP=0.156). Over the 20 strong baseline detectors, DASD consistently improves 17 concepts using a set of 300+ relatively much weaker detectors (from VIREO-374 [3]) as contexts (Run 1–4). Our 6 submitted runs are summarized below:

- A_vireo.dasd20scorelinear_1: DASD over a baseline using linear weighted fusion of local and global features. Concept affinity estimation method is the same to Run 3.
- A_vireo.dasd20fcs_2: DASD over Run 5; using ground-truth annotations and Flickr context to estimate concept affinity.
- A_vireo.dasd20score_3: DASD over Run 5; using ground-truth annotations and detection score to estimate concept affinity.
- A_vireo.dasd10_4: DASD over Run 5; using ground-truth annotations to estimate concept affinity (only applied for 10 concepts).
- A_vireo.localglobal_5: average fusion of local and global features.
- A_vireo.localalone_6: local feature alone - multiple detectors and spatial partitions.

Automatic Video Search:

For this task, in the past we have been focusing on concept-based video search [4, 5]. Given a textual query, various factors including semantic relatedness, co-occurrence, diversity, and detector robustness were jointly considered for better selection of the concept detectors. This year, in addition to textual queries, the visual query examples are also taken into account, and our main focus is on the combination of multiple search modalities. To this end we apply a concept-driven fusion scheme, which is able to dynamically discover the (near-)optimal modality weights for each query. Evaluation results confirm the effectiveness of our fusion approach, offering at least 10% improvement compared to the best uni-modality performance.

- F_A_N_CityUHK1: multi-modality fusion of concept-based search (a slight different setup based on Run 5), query-by-example (Run 9), and text baseline (Run 10).
- F_A_N_CityUHK2: multi-modality fusion of concept-based search (Run 5), query-by-example (Run 9), and text baseline (Run 10).
- F_A_N_CityUHK3: multi-modality fusion of concept-based search (Run 6), query-by-example (Run 9), and text baseline (Run 10).
- F_A_N_CityUHK4: multi-modality fusion of concept-based search (Run 7), query-by-example (Run 9), and text baseline (Run 10).
- F_A_N_CityUHK5: concept-based search; using both textual and visual example queries for concept selection.
- F_A_N_CityUHK6: concept-based search; using textual queries for concept selection based on semantic and context spaces [5].
- F_A_N_CityUHK7: concept-based search; using textual queries for domain adaptive concept selection based on Flickr context similarity.
- F_A_N_CityUHK8: concept-based search; using textual queries for concept selection based on Flickr context similarity.
- F_A_N_CityUHK9: query-by-visual-example.
- F_A_N_CityUHK10: text-based search.

Content-Based Video Copy Detection:

Our approach for copy detection is mainly based on our recent work on near-duplicate keyframe detection [6]. We consider only two features: bag-of-visual-words (BoW) based on SIFT and bag-of-audio-words (BoA) based on MFCC. To achieve fast and accurate BoW-based detection, indexing and various geometric verification techniques are employed. We submitted 2 video-only runs and 3 audio-video runs (see descriptions in Section 3).

1 High-Level Feature Extraction

In TRECVID 2009, we experiment our recently proposed algorithm, named domain adaptive semantic diffusion (DASD) [1], for context-based concept fusion. Starting from hundreds of

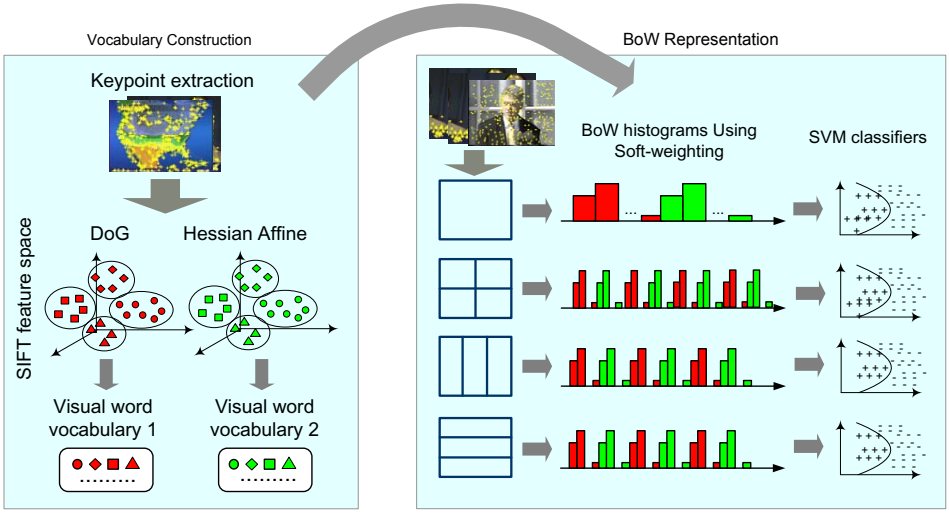

Figure 1: Our TRECVID-2009 local feature-based keyframe representation framework.

individually developed concept detectors, DASD exploits semantic context (concept relationship) to refine concept detection scores using graph diffusion technique. Particularly, it involves a semantic context adaptation process to cope with domain change between training and test data. We adopt our 2008 HLFE system as a baseline. In the end, we find that the well designed 2008 system which utilizes both local and global features still produces excellent performance with MAP=0.156, and the DASD algorithm is capable of consistently improving such a strong baseline for most of the evaluated concepts.

1.1 Baseline Detectors Using Local and Global Features

Bag-of-visual-words (BoW) representation derived from local keypoint features has been playing a very important role in a successful concept detection system. For this, we slightly update our 2008 BoW representation framework (see Figure 2 in [2]), by removing a keypoint detector and adding in one more spatial partition. The new framework is show in Figure 1. Detector MSER is dropped since it did not help much in TRECVID 2008. As using multiple spatial resolutions tends to be helpful, we add in a 3×1 partition. At the end for each concept, there are four SVMs to be trained using BoW histograms. For more details about this BoW representation, please refer to [2, 7].

We extract two kinds of global features: grid-based color moments (CM) and grid-based wavelet texture (WT). For 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. Two SVMs are trained for each concept using the two global features respectively.

Given a test keyframe[1], the SVM classifiers are applied on the same set of features for

[1]For both HLFE and automatic search tasks, we extract 3 keyframes from each test shot.

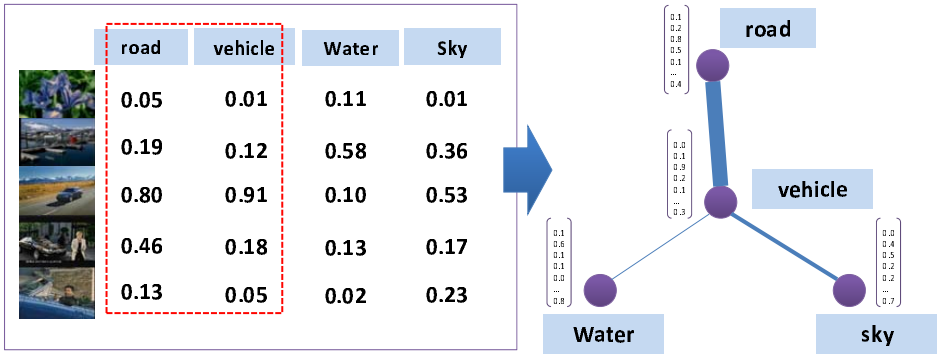

Figure 2: Illustration of DASD using four example concepts. Over a set of testing keyframes, detectors of frequently concurrent concepts tend to produce highly correlated prediction scores (left; *road* and *vehicle*). Therefore we model concept relationship in a graph structure where each node is a concept, and the edge weight (line width) indicates concept affinity (right). Prediction scores of the individual concept detectors are then refined w.r.t. the concept affinities using graph diffusion technique.

prediction. The raw outputs of the SVMs are converted into posterior probabilities (concept detection score). We then combine detection scores from the six SVMs in the "late fusion" manner, i.e. the final decision is made by fusing of the outputs multiple separate classifiers. In most of our experiments, "average fusion" is adopted to combine different classifiers.

1.2 Domain Adaptive Semantic Diffusion (DASD)

Most video concept detection systems assign single or multiple concept labels to a test sample (keyframe), where the assignment is often done independently without considering the inter-concept relationship. Due to the fact that concepts do not occur in isolation (e.g., *smoke* and *explosion*), more research attentions have been paid recently for improving detection accuracy by learning from semantic context (inter-concept relationship).

The learning of contextual knowledge, however, is often conducted in an offline manner based on training data, resulting in the classical problem of over-fitting. For large scale semantic concept detection which could involve simultaneous labeling of hundreds of concepts, the problem becomes worse when the unlabeled videos are from a domain different from that of the training data. For example, concept *weapon* always co-occurs with *desert* in news videos due to plenty of events about Iraq war. When such context relationship is captured by using news videos as training data, misleading detection results will be generated if it is applied to documentary videos where such relationship is seldom observed. This brings two challenges related to scalability for context-based learning: the need for adaptive learning and the demand for efficient detection.

DASD is designed to tackle these two challenges in a uniform fashion. As illustrated in Figure 2, one underlying assumption of DASD is that detectors of frequently concurrent concepts should produce highly correlated scores. We therefore construct an undirected and weighted graph, namely semantic graph, to model the concept affinities. The graph is then applied to

refine concept detection scores using a function level diffusion process. The aim is to recover the consistency of the detection scores w.r.t. the concept relationship. To handle the domain change problem, DASD further allows to simultaneously optimize the detection results and adapt the geometry of the semantic graph (concept affinity) according to the test data distribution. More formally, the cost function of DASD is defined as:

$$\mathcal{E}(g, W) = \frac{1}{2} \sum_{i,j=1}^{m} W_{ij} || \frac{g(c_i)}{\sqrt{d(c_i)}} - \frac{g(c_j)}{\sqrt{d(c_j)}} ||^2, \tag{1}$$

where $g(c_i)$ and $g(c_j)$ are the detection score vectors over a set of testing keyframes for concepts c_i and c_j; W_{ij} indicates the affinity between the two concepts; $d(c_i)$ and $d(c_j)$ are normalization factors; m is the total number of semantic concepts.

Apparently, this cost function evaluates the smoothness of g over the semantic graph. Therefore, reducing the function value of $\mathcal{E}$ makes the detection results g more consistent with the concept affinities captured by W. Specifically, we use gradient descent to reduce $\mathcal{E}$ by updating both g and W iteratively. The refinement of g is essentially a context-based concept fusion process, while the modification of W facilitates the domain adaptation of the semantic context.

The major advantage of DASD is twofold. First, it allows the online update of semantic context for addressing the problem of domain-shift. Second, with a linear complexity to the number of concepts and testing samples, it is scalable to large data sets where only a couple of minutes is required to complete DASD over hundreds of concepts for thousands of video shots. Interesting readers can refer to [1] for more details of DASD.

1.2.1 Graph Construction Methods

In DASD, both the concept affinity W and the detection score g need to be initialized. The initial values of g can be directly set as the detection scores of individual classifiers. There can be many ways to compute the concept affinity. Here we evaluate three different knowledge sources:

Ground-truth annotation: One ideal way to estimate the concept relationship is to use a training set X_{trn} and its corresponding label matrix Y, where $y_{ij} = 1$ denotes the presence of concept c_i in the sample x_j, otherwise $y_{ij} = 0$. Base on this, the concept relationship can be computed using Pearson product moment correlation as

$$PM(c_i, c_j) = \frac{\sum_{k=1}^{|X_{trn}|} (y_{ik} - \mu_i)(y_{jk} - \mu_j)}{(|X_{trn}| - 1)\sigma_i \sigma_j}, \tag{2}$$

where μ_i and σ_i are the sample mean and standard deviation, respectively, of observing c_i in the training set X_{trn}.

The correlation calculated by the above equation can be either negative or positive. In this experiment, we only consider positive correlation and construct a semantic graph $\mathcal{G}$ as:

$$\mathcal{G} = (\mathcal{C}, E, W), \tag{3}$$

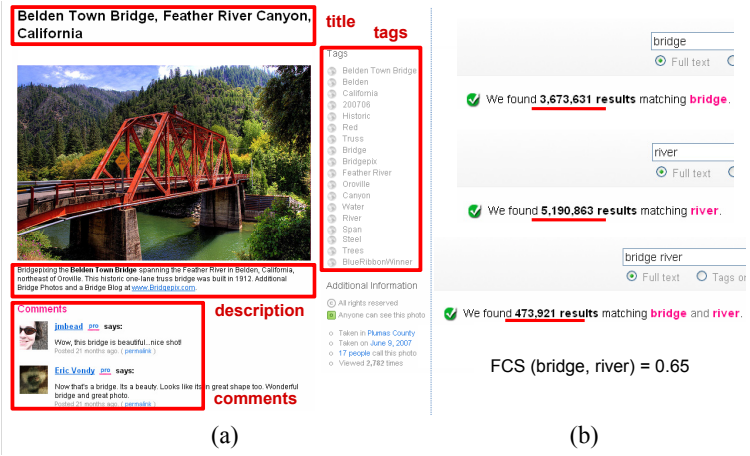

Figure 3: (a) Rich context information associated with a Flickr image. (b) The total number of images returned using keyword-based search in Flickr image context.

where $\mathcal{C}, E, W$ represent a node set (concepts), an edge set, and the edge weight matrix (concept affinities) respectively. An edge $e_{ij} \in E$ is established when $PM(c_i, c_j) > 0$ and $W_{ij} = PM(c_i, c_j)$.

Prediction score: An alternative source to estimate the concept affinities is the initial prediction scores from individual classifiers. Let $\mathcal{T}$ be the test data set and g_k^i be the baseline prediction scores of concept c_i in test shot k. Similar to Equation 2, the weight W_{ij} of the edge (c_i, c_j) can be calculated as $W_{ij} = \frac{\sum_{k=1}^{|\mathcal{T}|}(g_k^i - \mu_i)(g_k^j - \mu_j)}{(|\mathcal{T}|-1)\sigma_i\sigma_j}$, where μ_i and σ_i are the mean and standard deviation of the prediction scores of c_i, respectively, in $\mathcal{T}$.

Compared to ground-truth annotation, using prediction score is more economic (less accurate, though) since the former requires a fully labeled training set. Manual labels are difficult to obtain in practice, especially when the number of concepts is in the order of thousands.

Flickr context: The growing practice of online photo sharing has resulted in a huge amount of consumer photos accessible online. In addition to the abundant photo content, another attractive aspect of such photo sharing activity is the context information generated by users to depict the photos. As shown in Figure 3 (a), the rich context information includes title, tags, description and comments. Here we make use of such context information for concept affinity estimation.

Given two concepts, we compute their relatedness based on the number of Flickr images associated with the concept names. With the number of hits returned by Flickr, we apply NGD derived from Kolmogorov complexity theory to estimate concept distance [8]:

$$\mathrm{NGD}(x, y) = \frac{\max\{\log h(x), \log h(y)\} - \log h(x, y)}{\log N - \min\{\log h(x), \log h(y)\}}, \tag{4}$$

where $h(x)$ denotes the number of images associated with concept x in their context, and $h(x, y)$ denotes the number of images associated with both concepts x and y; N is the total number of images on Flickr, which is roughly estimated as 3.5 billion by the time we did the experiments. The NGD is then converted to Flickr context similarity (FCS) using a Gaussian kernel, defined

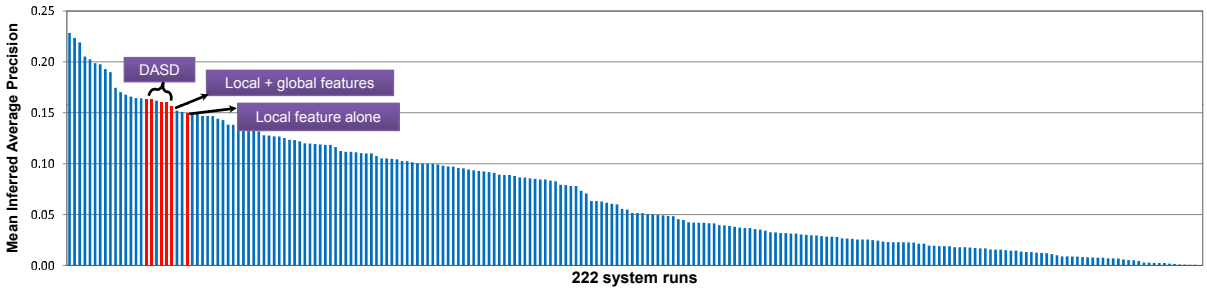

Figure 4: Mean average precision of all 222 HLFE runs submitted to TRECVID-2009. Our submissions are marked in red.

as

$$\mathrm{FCS}(x, y) = e^{-\mathrm{NGD}(x,y)/\rho}, \tag{5}$$

where the parameter ρ is empirically set as the average pairwise NGD among a randomly pooled set of words. Similar way of setting ρ has been shown to be effective for kernel based classification tasks [7]. An example of calculating FCS is shown in Figure 3 (b).

FCS can be directly used as concept affinity for DASD. Similar to detection score, it also does not require a fully labeled training set. Since Flickr contexts are basically composed of descriptions to photos, the word co-occurrence estimated using FCS somewhat reflects visual co-occurrence of both terms in images. This is the major advantage of FCS compared to other word similarity measures (e.g., those using WordNet ontology). More descriptions/evaluations of FCS can be found in [9].

1.3 HLFE Results and Analysis

As described in the abstract, we have six submissions including two baseline runs (Run 5&6) and four DASD runs (Run 1–4). Figure 4 shows MAP performance of all the official submissions this year. Our runs are marked in red color. As can be seen from the figure, our 2008 system which judiciously uses local and global features still produces impressive performance, with a local feature alone run of MAP at 0.150, and a local+global feature run of MAP at 0.156. The light-weight modification of the system does not show clear performance improvement. From our internal evaluation, the MAP performance merely drops to 0.149 without the newly added 3×1 partition. Thus we conclude that although using multiple spatial partitions tends to be helpful, adding too many layers will deteriorate the feature mismatch problem, especially for objects spanning multiple cells.

The major difference of the four DASD runs is semantic graph construction method, i.e., the way to compute concept affinities. We use VIREO-374[2] detectors [3] as contextual knowledge, and construct a graph of 394 nodes by directly adding the 20 evaluate concepts into the semantic

[2]Download site: http://vireo.cs.cityu.edu.hk/research/vireo374/. Features and detection scores of VIREO-374 on TRECVID 2009 data collection have been released. Detection scores of this year's 20 concepts using our new models (for Run 5) are also available at http://www.ee.columbia.edu/ln/dvmm/CU-VIREO374/.

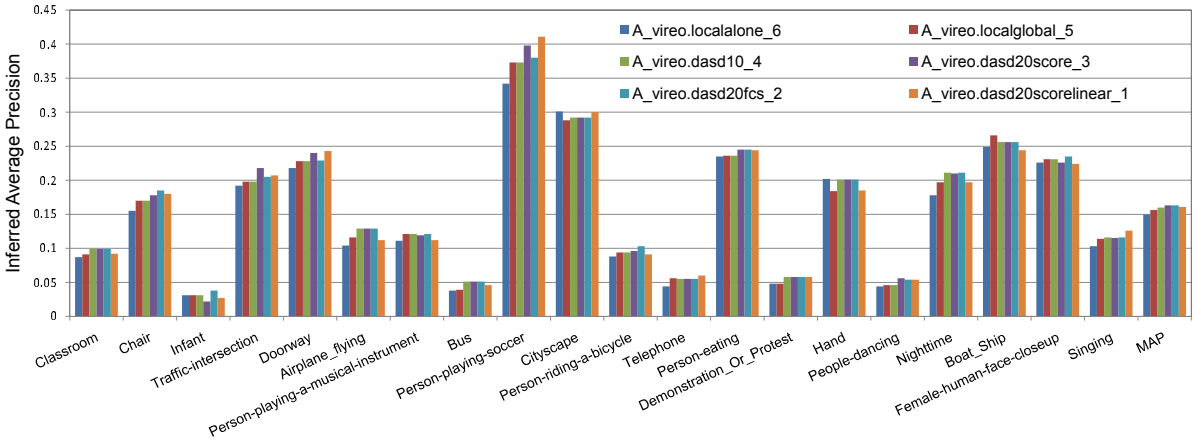

Figure 5: Per-concept performance of our submitted runs.

space. Note that although about half of the 20 evaluated concepts exist in the 374 concept set, we treat them as different nodes in the graph since the training set and feature representations of VIREO-374 are quite different from our new basline[3]. Throughout the experiments, we use ground-truth annotations (on TRECVID 2005 development set) to compute concept affinity for the new concepts with their names existing in VIREO-374 (10 out of 20). Based on this we submitted Run 4 where DASD is only applied to the 10 concepts (MAP gain for the 10 concepts: 5%). For the remaining 10 *out-of-vocabulary* concepts, we compute their affinities to VIREO-374 concepts using either detection score (Run 3) or Flickr context (Run 2). For Run 2–4, we use Run 5 as baseline, while for Run 1, we adopt the same semantic graph as Run 3, but use a slightly different (and worse) baseline from (than) Run 5.

As shown in Figure 4, all the DASD runs steadily outperform the baseline, with MAP of 0.160–0.163. The overall improvement over the baseline is about 4.5%. This is lower than that from our previous experiments on TRECVID 2005–2007 [1], where the improvements range from 11.9% to 17.5%. This is due to the fact that for the previous TRECVID collections we used similar baselines for all the concepts, while for TRECVID 2009, the baseline detectors of the 20 target concepts are much *stronger* than that of VIREO-374. Using *weaker* detectors as contexts indeed largely limits the performance gain of DASD in this experiment.

Figure 5 gives the per-concept performance of our six submissions. Similar to our observations in [1], we see that DASD consistently outperforms the baseline for most of the concepts (e.g., Run 2 improves 16/20 concepts over Run 5). Significant improvements are obtained for many concepts (e.g., 30% for *Bus* and 10% for *person-riding-bicycle*). Very few concepts surfer from minor performance degradation (e.g., *Boat_Ship*).

To clearly compare the effectiveness of FCS and detection score for graph construction (concept affinity estimation), we show per-concept performances of Runs 2, 3, and 5 for the 10 out-of-vocabulary concepts in Figure 6. Both methods do not require all the concepts being

[3]VIREO-374 detectors were trained on broadcast news videos (TRECVID 2005 development set) using BoW (500-d; no spatial partition), CM, and WT features.

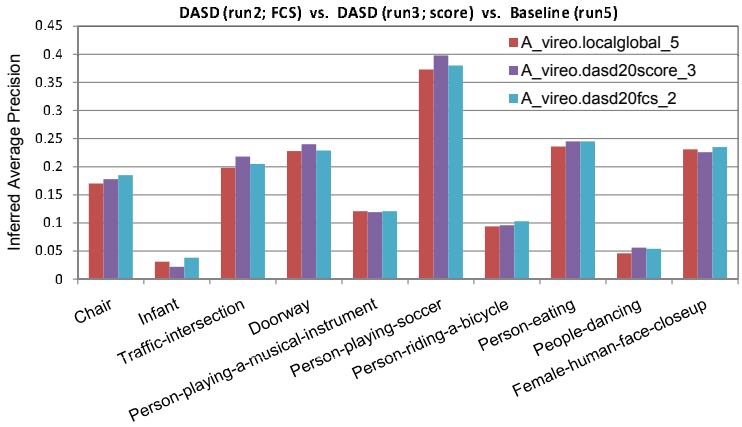

Figure 6: Comparison of different graph construction methods (using FCS and detection score respectively) over the 10 out-of-vocabulary concepts.

fully labeled over a large training set (in contrast to computing affinity using ground-truth annotation). We see that there is no clear winner of the two, with equal overall MAP improvement at 4%. Thus we conclude that detection score may be used for out-of-vocabulary concepts if the test set is large enough to compute score correlation (cf. Section 1.2.1). Otherwise, FCS is a good choice.

The DASD algorithm is highly efficient. The complexity is $O(mn)$, where m is the number of concepts and n is the number of testing samples (keyframes). Specifically, performing DASD for one testing sample only requires 2 milliseconds on a regular PC. More detailed empirical evaluations of DASD can be found in [1].

2 Automatic Video Search

For the automatic search task, our main focus is on the combination of three modalities: text-based search, concept-based search, and query-by-example (QBE). We submitted four multi-modality search runs (Run 1–4) and six uni-modality search runs (four of them are concept-based search runs with different concept selection schemes). In the following we first describe each of the three modalities. After that, we introduce our multi-modality fusion method and analyze search results.

2.1 Text-based Search

In text search, we only employ the English output of ASR/MT [10]. Instead of traditional method like TF-IDF, our text search model adopts Okpai [11] 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 [12] is used.

2.2 Concept-based Search

We use VIREO-374 [3] detectors for concept-based search, and submit four runs (Run 5–8) based on different concept selection schemes, which are described as follows.

2.2.1 Semantic Space and Context Space (SSCS) Reasoning

For this, we directly apply our previous work in [5]. Only textual queries are considered for concept selection. Two spaces, semantic space (SS) and context space (CS) are constructed using WordNet ontology and manual annotations (on TRECVID 2005 development data) respectively. SS and CS are then jointly used to model the semantic and contextual relationship between concepts and query terms. Concepts that are semantically and/or contextually relevant to the textual query terms are selected. Finally, a multi-level fusion strategy is further employed to combine the selected concepts. Note that the aim of this multi-level fusion is for the combination of multiple concepts, which is different from the multi-modality fusion introduced later on. For more details about this component, please refer to [5].

2.2.2 Flickr Context Similarity (FCS)

In addition to semantic and context reasoning, we also test FCS described in Section 1.2.1 for concept selection. The relevancy of a concept C to a query term q is measured by FCS(C, q). For each query term, three concepts are empirically selected. All the selected concepts are finally linearly fused to response the query, where the fusion weight for each concept is determined by FCS.

2.2.3 Signed Fisher Ratio (SFR)

Different from the text-based schemes introduced above, in this section we perform concept selection by investigating visual query examples. The intuition is, by surveying the presence (absence) of concepts in positive (negative) examples, concepts are picked accordingly based on their prevalence and discriminativeness. There are several assumptions made here. First, the presence of a concept in visual examples is not known a priori in practice. We thus adopt concept detector to predict the relevance of a concept to the visual query examples. Second, query examples are treated as positive training samples, and in general the number of examples is very small. Third, there is no negative sample comes along with a query. We draw samples randomly from the training set as pseudo-negative examples, assuming that majority of the training samples are not relevant to the query.

Denote $\{u_+, \sigma_+\}$ and $\{u_-, \sigma_-\}$ as the mean and standard deviation of prediction scores from a concept C on the positive and negative training examples respectively. We propose a signed Fisher ratio (SFR) to measure the relevance of C to query as

$$V_{rel}(C) = sign(u_+ - u_-) \cdot \frac{(u_+ - u_-)^2}{\sigma_+^2 + \sigma_-^2}, \tag{6}$$

where the *sign* function contrasts the prevalence of C in positive and negative samples. Positive value will be assigned, indicating the usefulness of C to visual query, if C receives higher prediction scores on average in positive than in negative samples. This function basically groups the available set of concepts into positive and negative concepts with respect to the query. The discriminativeness of a concept is further determined by the second part of the equation, which is the original formula of Fisher ratio, for measuring class separability. By SFR, all concepts are eventually ranked based on their relevance. We consider the top-k most relevant concepts where the value of k is empirically set equal to the number of concepts selected by semantic and context reasoning.

2.3 Query-by-Example (QBE)

The visual query examples include images and/or short video clips. We adopt the supervised learning approach (Multi-bag SVMs) in [13] for QBE, by training ten SVMs for each query. Visual examples are used as the positive training samples for all the SVMs. Ten sets of pseudo-negative examples are randomly sampled from the dataset and used separately for each SVM. The visual features for learning SVMs are concept scores produced by VIREO-374 concept detectors. In other words, each sample is represented by a vector of 374 elements, where each element is a probability indicating the confidence of detecting the corresponding concept in the sample. With the outputs from all the ten SVMs trained for each query, finally we adopt average fusion to combine the results for ranking video shots.

2.4 Concept-Driven Multi-Modality Fusion

In this section we introduce a concept-driven approach for multi-modality fusion (text-based search, concept-based search, and QBE). Since the importance of the each search modality for a user query cannot be directly obtained in advance, we explore its selected concepts in concept-based search to implicitly predict fusion weights.

To this end, we treat each semantic *concept* as a *simulated query* by using *concept name* as *textual query* and *ten randomly chosen positive samples* as *visual query*. Based on both *textual and visual queries* we perform search for this *concept (simulated query)* against a training dataset using the three modalities (text-based, concept-bsed, and QBE). Then the uni-modal search performance for the concept/simulated-query can be easily computed using manual labels of the concept. With these simulated search evaluations, given a real user query, we estimate its multi-modality fusion weights on-the-fly, by jointly considering query-concept relatedness and the simulated search performance of all selected concept.

2.5 Search Results and Analysis

Table 1 summarizes our submitted runs. With such a very small number of positive samples per-query, the multi-bag SVMs work very well with MAP at 0.029 (Run 9), but the text-based search (Run 10) performance is surprisingly bad this year. Among all the concept-based

Table 1: Description and MAP of our submitted runs.

Run ID	Description	MAP
Run 1	Multi-modality Fusion (same to Run2 but with a different SFR setup)	0.044
Run 2	Multi-modality Fusion (Run 5+Run 9+Run 10)	0.035
Run 3	Multi-modality Fusion (Run 6+Run 9+Run 10)	0.046
Run 4	Multi-modality Fusion (Run 7+Run 9+Run 10)	0.040
Run 5	Concept-based Search (SSCS+SFR)	0.030
Run 6	Concept-based Search (SSCS)	0.033
Run 7	Concept-based Search (FCS+adaptation)	0.036
Run 8	Concept-based Search (FCS)	0.049
Run 9	QBE	0.029
Run 10	Text-based Search	0.002

search runs, FCS which only uses textual queries generates an impressive MAP at 0.049 (Run 8). Compared with the other two concept-based runs which select concepts using SSCS (Run 6) and its combination with SFR[4] (Run 5), FCS achieves MAP gains of 48.4% and 63.3% respectively. This indeed signified the advantage of FCS for concept selection – it is able to reflect visual co-occurrence of words (query terms/concepts). A detailed evaluation of FCS can be found in [9]. Additionally, in Run 7, we further applied a new adaptation technique to refine the selected concept set of FCS according to test data characteristics [9]. However, we performed this adaptive search on shot-level after consolidating concept detection scores from keyframe-level[5], which seems to degrade the performance significantly.

On the other hand, from comparing Run 5 with Run 6, it turns out that collecting additional concepts from SFR does not improve the performance. We investigated the results and found that some concepts are selected twice by SSCS and SFR. As a result our simple polling strategy largely over-weighs the importance of these concepts.

For the four multi-modality search runs, we see consistent MAP improvements over corresponding uni-modality performance (e.g., 46.7% of Run 1 over Run 5, 16.7% of Run 2 over Run 5, 39.3% of Run 3 over Run 6, and 13.9% of Run 4 over Run 7). This confirms the effectiveness of our concept-driven multi-modality fusion approach.

Figure 7 further gives per-query results of our submitted runs, together with the mean and best performance of all submissions. We achieved the best performance for Query-273 (*for a closeup of a hand, writing, drawing, coloring, or painting*), and above-mean performance for most of the remaining queries. It is also worth noting that although concept-based search produced the best performance among our submissions (Run 8), it is largely limited by the number/quality of available concept detectors. For example, a concept *cable* is selected for Query-283 "*find a person playing a piano*" because there is no better choice. In the future we plan to put more efforts in building large-scale reliable concept detectors.

[4]For this run we simply pool all the selected concepts by SSCS and SFR together for concept-based search.

[5]All the other runs are conducted on keyframe-level.

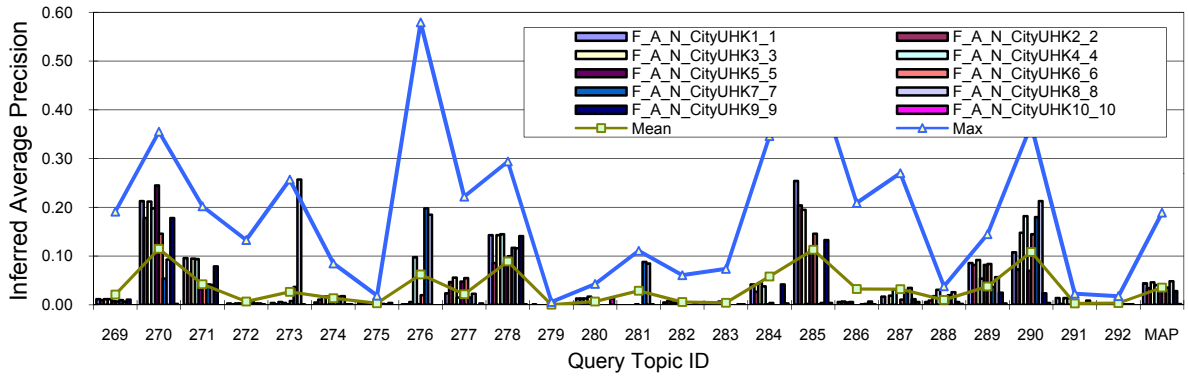

Figure 7: Per-Query performance of our submitted runs vs. the mean and best performance of all submissions of fully automatic search.

3 Video Copy Detection

3.1 Video-Only Copy Detection

Our submitted video-only runs for CBCD are mainly based on our recent works on near-duplicate video search. Figure 8 shows our framework, which consists two major parts: offline indexing and online retrieval. We only consider one feature - bag-of-visual-words (BoW) generated from local keypoints. For offline indexing, one video frame is sampled every 2.5 seconds (567,056 frames in the test/reference set). Keypoint features are computed using SIFT [14]. A vocabulary of 20,000 visual words is generated for BoW representation. We adopt inverted file to index the frames extracted from the reference set.

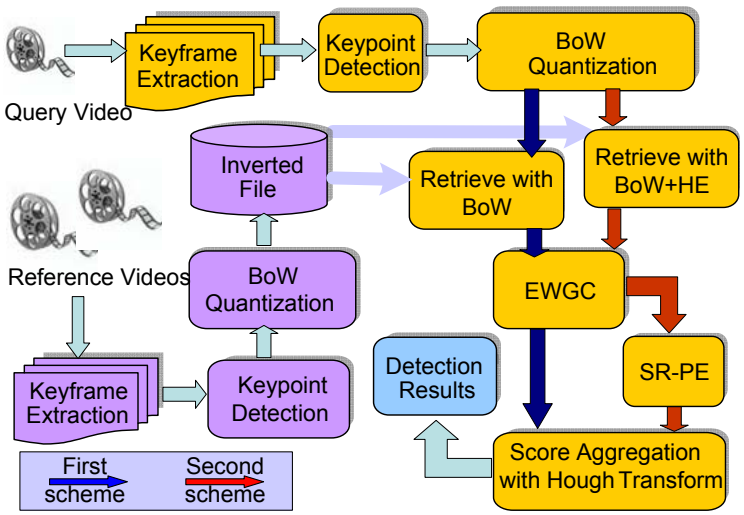

Figure 8: Video-only copy detection framework.

During online retrieval, we extract one frame per 1.25 seconds from queries. The frames are represented with BoW, and mapped into the inverted file structure for efficient search.

To alleviate the information loss in BoW quantization, we adopt an enhanced weak geometric consistency verification strategy (EWGC). EWGC is a modified version of WGC proposed in [15]. To consolidate matching results from frame level to video level, we employ 2D Hough transform (HT) to aggregate scores from the matched frames and localize copy segments.

We also perform and test an internal run by including hamming embedding (HE) and pattern entropy checking (SR-PE [6]), before and after the EWGC, respectively. HE is adopted as a pre-filtering step, to reduce the number of candidate frame pairs for EWGC and thus improve the overall detection efficiency. The additional process of SR-PE further prunes false alarms which cannot be filtered by EWGC. Our video-only runs are listed as follows. Table 2 summarizes the detailed settings of these runs.

Table 2: Configurations of different video-only runs.

Run ID	Retrieval Scheme	HE	EWGC	SR-PE	Weighting Scheme
Vireo.v.BALANCED.tgc	First Scheme		√		$\text{Sim}^1 \times \text{Score}$
Vireo.v.BALANCED.norm	First Scheme		√		$\text{Sim}^2 \times \text{Score}$
Vireo.v.NOFA.srpe	Second Scheme	√	√	√	$\text{Sim}^1 \times \text{N}_c$

Sim^1: cosine distance; Sim^2: ratio of matched visual words between two frames; Score: EWGC score; N_c: # of correct visual word matches estimated by SR-PE

Vireo.v.BALANCED.tgc: This run employs EWGC and 2D HT, shown as "First Scheme" as indicated in Figure 8. For each query, we return top-2 retrieved videos for submission.

Vireo.v.BALANCED.norm: Similar to the previous run but with a different similarity measure (see Table 2).

Vireo.v.NOFA.srpe: This run employs HE, EWGC, SR-PE and 2D HT, shown as "Second Scheme" in Figure 8. Only top-1 retrieved video is returned. Note that this run is not officially submitted.

Table 3 shows the performance of three runs, measured in terms of NDCR and F1, and compared against the best and median results. Among the two official runs we submitted, Vireo.v.BALANCED.tgc shows better performance in terms of NDCR. This implies that cosine similarity is a better choice over the ratio of matched visual words. The updated run, Vireo.v.NOFA.srpe, shows much better performance. In terms of NDCR, it is even better than the best reported results for T3, T4, T5 and T6. For T2 (Pic-in-Pic), T8 and T10 (with mirror), our approach shows relatively poor performance (compared to the best results) due to the limitation of local features which cannot deal with these types of transformations well.

Because our framework is basically targeted for near-duplicate detection, the performance is not so good in terms of F1 measure, especially for reference videos containing near-duplicate scenes to the true video copy segment. A typical example is shown in Figure 9, where many frames in the ground-truth video are visually similar to the query video. In this case, the exact copy segment cannot be correctly located by our approach, even though this video can always be correctly identified.

Table 3: Performance of different video-only runs.

	Opt.NDCR					Opt.F1				
	tgc	norm	srpe	median	best	tgc	norm	srpe	median	best
T2	0.985	1.204	0.940	1.005	**0.938**	0.920	0.95	0.637	0.165	**0.982**
T3	0.898	0.970	**0.022**	0.998	0.216	0.780	0.783	0.612	0.801	**0.995**
T4	0.993	0.993	**0.478**	1.000	0.493	0.720	0.720	0.613	0.725	**0.938**
T5	0.993	0.993	**0.060**	1.001	0.142	0.720	0.720	0.624	0.775	**0.995**
T6	1.000	1.032	**0.015**	1.002	0.408	0.000	0.804	0.597	0.800	**0.952**
T8	0.985	0.993	0.851	1.002	**0.420**	0.869	0.944	0.488	0.820	**0.993**
T10	1.032	1.102	0.940	1.005	**0.723**	0.875	0.944	0.468	0.748	**0.990**

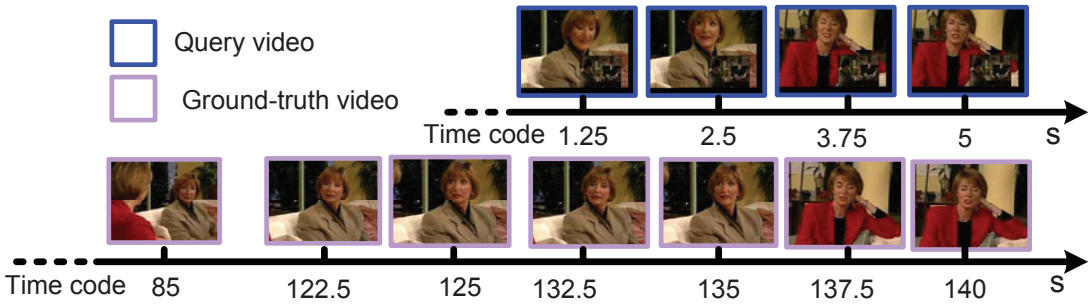

Figure 9: A typical example on which our approach shows low localization accuracy. For a reference video which has self-duplicate or similar segments throughout the sequence, our approach is not able to accurately localize the copy.

For query processing and retrieval time, the two official runs take approximately 250 seconds to complete the search for a query video of average length (178 seconds). For the new run NOFA.srpe, due to the use of HE as a pre-filtering step, the search time is significantly reduced to 134 seconds.

3.2 Audio-Video Copy Detection

Following the similar idea of BoW, we propose the use bag-of-audio-words (BoA) to perform audio-based copy detection. For both query and reference video sequences, we extract audio signals and converted them to mono PCM (16 bits) with $44.1KHz$ sampling rate. After that, the audio signals are divided into segments of $40ms$-length with 50% overlap between consecutive ones. Each segment is then represented by a 39-dimensional Mel-Frequency Cepstrum Coefficients (MFCCs). Finally, every 50 continuous audio segments (1 second audio signal) are quantized using a pre-constructed audio word vocabulary ($15,000$ words), and represented by a BoA histogram. The BoA histogram is used as audio feature of a video frame closet to the center of the 1 second time slot. With the BoA features, similar to our video-only approach, Hough transform is used to aggregate frame-level similarity scores. For each audio query, the top one video is returned.

For joint audio-video detection, we use late fusion to fuse our audio-only result with several

Table 4: Configuration of the video-only component in our audio-video runs.

Run ID	Retrieval Scheme	HE	EWGC	SR-PE	Weighting Scheme
Vireo.v.NOFA.he	Second Scheme	√	√	√	$\text{Sim}^1 \times \text{N}_c$
Vireo.v.BALANCED.ewgc	Second Scheme	√	√	√	$\text{Sim}^1 \times \text{Score}$
Vireo.v.BALANCED.retri	Second Scheme	√	√		$\text{Sim}^1 \times \text{Score}$
Vireo.v.NOFA.srpe	Second Scheme	√	√	√	$\text{Sim}^1 \times \text{N}_c$

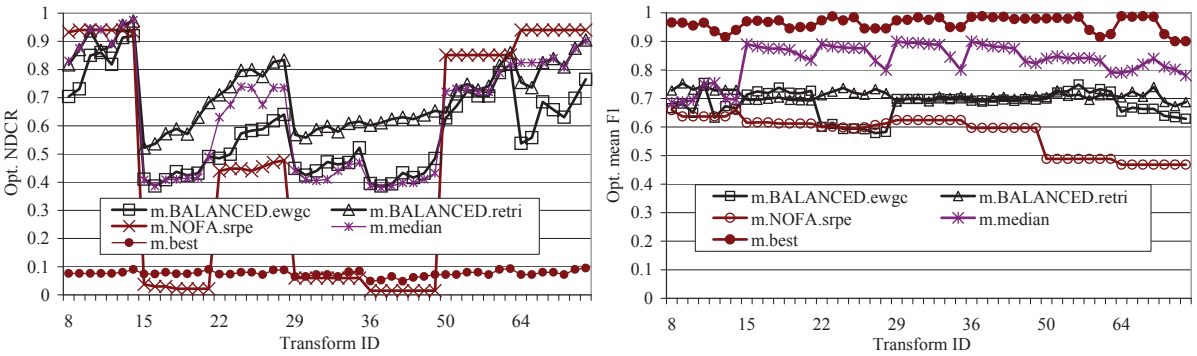

Figure 10: Audio-video copy detection performance.

video-only runs under different settings. We perform four runs as described below. Detailed configurations of the video-only part in our audio-video runs are summarized in Table 4.

Vireo.m.NOFA.he: This run fuses audio-only with Vireo.v.NOFA.he. There are programming mistakes in this run. Results will not be discussed.

Vireo.m.BALANCED.ewgc: This run fuses audio-only with Vireo.v.BALANCED.ewgc.

Vireo.m.BALANCED.retri: This run fuses audio-only with Vireo.v.BALANCED.retri.

Vireo.m.NOFA.srpe: This run serves as an update of Vireo.m.NOFA.he. It fuses audio-only with Vireo.v.NOFA.srpe. Note that this run was not submitted.

We linearly fuse video and audio results. Because scores from audio-only are much smaller than those of video-only, we use weights of 10:1 to combine both modalities. Figure 10 shows the performance of the three runs, measured in terms of NDCR and F1, and compared against the best and median results[6]. For the two official runs (m.BALANCED.ewgc and m.BALANCED.retri), performances are around median in terms of NDCR, but are relatively worse in terms of F1 measure. This indicates that the audio-only part contributes little to reducing the localization errors in the video-only runs. The performance of the updated run (m.NOFA.srpe) is above median in terms of NDCR. It is worth noting that, for 21/49 transformations, the NDCR scores of this run are below 0.05. From our investigation, compared to video-only, the fusion is able to improve 10% in terms of recall while degrade 7% for precision. On one hand, audio-only results are complementary to video-only results, especially when video-only cannot handle the Pic-in-Pic and flip transformations well. On the other hand, however, since MFCC feature is not discriminative enough, a number of false-alarms are also brought into the audio-video results.

[6]The best and median results are roughly estimated by ourselves from the figures officially released.

For computational time, the video-only part used for all audio-video runs finishes one query in 134 seconds on average. Because of the simplicity of our audio-only system, audio-only copy detection is much more efficient – taking just 15 seconds on average for a query. In total, the average time cost for a query is around 150 seconds.

4 Summary

This year, we applied our TRECVID-2008 HLFE system [2] as baseline. As expected, we again demonstrated strong performance of this system which largely relies on local keypoint features (MAP=0.156). By considering context-based concept fusion using DASD, we obtained some further improvements. Similar to our observations in [1], we have seen consistent improvement for most of the evaluated concepts from using DASD. For automatic search task, we demonstrated the power of using Flickr context for query-concept similarity estimation. Compared to other ontology-based word relatedness measures, it has the advantage of reflecting word co-occurrence in images rather than text corpus. We also presented a concept-driven multi-modality fusion method, which offers significant performance gains of 14-47% over uni-modality search. We believe that using multiple search modalities with query-adaptive fusion scheme is the right direction towards a successful image/video search system.

For copy detection, video elements are represented using both bag-of-visual-words and bag-of-audio-words. From our evaluation, video feature plays a more important role in copy detection. Our additional video-only runs achieve the best results for 4 out of 7 transformations in terms of NDCR. However, for F1 measure, since our framework is mainly developed for near-duplicate detection, the performance is not satisfactory due to the excessive number of false alarms caused by incorrect copy localization.

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 119508 and CityU 119709).

References

[1] Y.-G. Jiang, J. Wang, S.-F. Chang, and C.-W. Ngo, "Domain adaptive semantic diffusion for large scale context-based video annotation," in *ICCV*, 2009.

[2] 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.

[3] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. G. Hauptmann, "Representations of keypoint-based semantic concept detection: A comprehensive study," *IEEE Trans. on Multimedia, in press*, 2010.

[4] C.-W. Ngo, Y.-G. Jiang, X. Wei, W. Zhao, F. Wang, X. Wu, and H.-K. Tan, "Beyond semantic search: What you observe may not be what you think," in *NIST TRECVID workshop*, 2008.

[5] X.-Y. Wei and C.-W. Ngo, "Fusing semantics, observability, reliability and diversity of concept detectors for video search," in *ACM 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*, vol. 18, pp. 412–423, 2009.

[7] 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.

[8] R. L. Cilibrasi and P. M. Vitanyi, "The google similarity distance," *IEEE Trans. on Knowledge and Data Engineering*, vol. 19, pp. 370–383, 2007.

[9] Y.-G. Jiang, C.-W. Ngo, and S.-F. Chang, "Semantic context transfer across heterogeneous sources for domain adaptive video search," in *ACM Multimedia*, 2009.

[10] 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.

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

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

[13] A. P. Natsev, M. R. Naphade, and J. Tesic, "Learning the semantics of multimedia queries and concepts from a small number of examples," in *ACM Multimedia*, 2005.

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

[15] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *ECCV*, 2008.