VIREO at TRECVID 2010: Semantic Indexing, Known-Item Search, and Content-Based Copy Detection

Chong-Wah Ngo[†], Shi-Ai Zhu[†], Hung-Khoon Tan[†], Wan-Lei Zhao[†], Xiao-Yong Wei[†‡]

[†] *Video Retrieval Group (VIREO), City University of Hong Kong*

[‡] *School of Computer Science, Sichuan University, China*

http://vireo.cs.cityu.edu.hk

Oct 19, 2010

Abstract

This paper presents our approaches and the comparative analysis of our results for the three TRECVID 2010 tasks that we participated in: *semantic indexing*, *known-item search* and *content-based copy detection*.

Semantic Indexing (SIN):

Our main focus for the SIN task is on the study of the following two issues: 1) the effectiveness of concept detectors for indexing web video dataset, and 2) how to leverage the ontology relationships to reinforce concept detection. Our baseline detectors are similar to those of our TRECVID 2009 system, where both local and global features are employed to train the SVM model for each concept. Based upon the baseline detectors, we propose two approaches to refine the detection scores. The first integrates the ontology information into the random walk framework. The second seeks the agreement among the set of ranked lists generated by semantically related concepts. Our four submitted runs are summarized below:

- F_A_VIREO.randomwalk_1: perform a random walk over the baseline result using local feature alone. Flickr distance and ontology relationship are used to build the context.
- F_A_VIREO.agreement_2: re-rank the videos by seeking the agreement among semantically related concepts on the relevant videos.
- F_A_VIREO.baseline_vk_3: local feature alone - multiple detectors.
- F_A_VIREO.baseline_vk_cm_4: average fusion of local feature and global feature.

Known-Item Search (KIS):

In this new task, there is a shift in the search requirement compared to previous years' search tasks, from general queries with multiple solutions to specific queries with single solution. This has rendered previously proposed algorithms ineffective. In this first attempt, our objective is to

observe the effectiveness of different modalities (metadata, automatic speech recognition (ASR) and concepts) towards known-item search. Evaluation result shows that for known-item search, textual-based modalities are useful, where the metadata is the most effective while ASR plays a complementary role. We submitted four runs for the fully automatic settings as follows:

- F_A_YES_vireo_run1_metadata_asr_1: metadata + ASR
- F_A_YES_vireo_run2_metadata_2: metadata only
- F_A_YES_vireo_run3_asr_3: ASR only
- F_A_YES_vireo_run4_concept_4: concept only

Content-Based Video Copy Detection (CCD):

We submitted three runs based on our video-only detection framework. Since we only conduct video-only copy detection, for the video+audio detection task, we simply submit the same detection results. The three submissions are defined as follows:

- Vireo.m.BALANCED.srpe: employ Hamming Embedding (HE), Enhanced Weak Geometric Consistency Checking (EWGC), Scale-Rotation Invariant Pattern Entropy (SR-PE) and 2D Hough Transform (2D HT)
- Vireo.m.BALANCED.srpeflip: repeat the same procedure as the previous run but with the flipped queries. The retrieval results are then linearly fused with the previous run before being fed into 2D HT.
- Vireo.m.NOFA.srpeflip: same as the second run except that a higher threshold is chosen.

1 Semantic Indexing

This year, we experiment two approaches, random walk and agreement seeking, for ontology-based concept fusion. First, we adopt the reranking framework proposed by Hsu [1] to refine the initial detection scores. Random walk is conducted on a context graph where each node represents a concept while an edge reflects the similarity between two concepts. Second, in the agreement-based approach, ontology reasoning is used to select a set of semantically related concepts. The ranked lists for the selected concepts are then consolidated into a final ranked list through linear fusion.

1.1 Baseline Detectors Using Local and Global Features

For baseline, we use the Bag-of-Word (BoW) representation derived from local keypoint features since it has been consistently one of the most effective features for concept detection. Our BoW representation framework is similar to that of our TRECVID 2007 system [3], where two 500-D component BoW feature vector generated from two different keypoint detectors are concatenated to form a single 1000-D feature vector. For more details on this BoW representation, please refer to [3, 4, 6].

We only extract grid-based color moments (CM) as global feature. For color moment, each keyframe is partitioned into 5 × 5 grids, and the first three moments are computed on the Lab

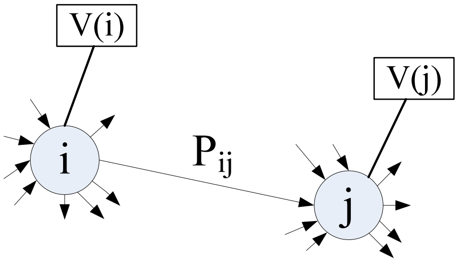

Figure 1: Example of score transition over nodes; $v(i)$ and $v(j)$ are initial scores for node i and j; p_{ij} is the transition probability from i to j.

color space over each grid. Concatenating the features from all grids forms a vector of 255 dimensions. Finally, two SVMs are trained for each concept using the local and global features, respectively.

Given a testing keyframe[1], the SVM classifiers are applied on the corresponding features for prediction and the raw outputs of the SVMs are converted into posterior probabilities (concept detection score). The output of classifiers learnt on local feature is our first baseline. We then combine the detection scores from the two SVMs by average fusion as the second baseline.

1.2 Inferring Semantic Context Relationship for Score Refinement

For most concept detection systems, the semantic indexing is done by independently predicting a concept on a testing sample (keyframe) without considering the contextual information between different concepts. However, most of the concepts do not occur in isolation. Mining semantic context information for improving detection accuracy has triggered a lot of interest in multimedia community. The inter-concept relationships are two-fold: 1) semantic relationship, for example concept “car” is a kind of “vehicle”; 2) contextual relationship, which considers the co-occurrence between concepts, such as “car” and “road”.

1.2.1 Random Walk on Context Graph

Considering the various issues in learning a concept detector, e.g., limited training samples, visual diversity and complexity, the performances of any single concept detector are not perfect. Indeed, the performance of a particular concept detector can be reinforced by regularizing its scores through a set of semantically-related concepts. For instance, the concept “car” always co-occurs with “road” and therefore the detection scores for these two concepts are expected to be strongly influenced by each other. We formulate this problem as a random walk process which propagates initial score on a pre-defined context similarity graph. The transition of a random walk process is governed by the element p_{ij} of matrix $\mathbf{P}$. As showed in Figure 1, the score of node i propagates to node j with a probability p_{ij}. The refined score of node j at

[1] For both SIN and KIS, we extract 3 keyframes from each test shot.

iteration k is defined as follows:

$$x_k(j) = \alpha \sum_{i \in B_j} x_{k-1}(i) p_{ij} + (1 - \alpha) v(j); \tag{1}$$

where $x_k(j)$ and $v(j)$ are the reranked score at iteration k and the initial score for image j, respectively, B_j is the set of nodes pointing to node j, and $\alpha \in [0, 1]$ is a parameter used to control the importance of transition score and initial score.

To construct the transition matrix, two different knowledge sources are employed. First, we adopt the Flickr context [2] which explores the context information generated by users of image sharing website. Given two concepts i and j, the Flickr context similarity (FCS) is derived from NGD using a Gaussian kernel, defined as

$$\mathrm{FCS}(i, j) = e^{-\mathrm{NGD}(i,j)/\rho}, \tag{2}$$

With the number of hits returned by Flickr, NGD estimates concept distance based on Kolmogorov complexity theory:

$$\mathrm{NGD}(i, j) = \frac{\max\{\log h(i), \log h(j)\} - \log h(i, j)}{\log N - \min\{\log h(i), \log h(j)\}}, \tag{3}$$

where $h(i)$ denotes the number of images associated with concept i in their context, and $h(i, j)$ denotes the number of images associated with both concepts i and j; N is the total number of images on Flickr. More descriptions and evaluations of FCS can be found in [2].

Second, we further infuse the ontology information provided by TRECVID 2010 into transition matrix. The ontology information defines two kinds of relationship between two concepts, i implies j or i excludes j, such as “actor” implies “person” and “indoor” excludes “outdoor”. If i implies j, in random walk process, the score of concept i should be transferred in its entirety to concept j. Conversely when i excludes j, the scores of concept i and j should not be allowed to influence one another..

With the above two kinds of knowledge sources, our transition probability is thus defined as:

$$p_{ij} = \begin{cases} 1, \text{ if } i \text{ implies } j; \\ 0, \text{ if } i \text{ excludes } j \text{ or } j \text{ excludes } i; \\ \mathrm{FCS}(i, j), \text{ otherwise} \end{cases} \tag{4}$$

1.2.2 Agreement Seeking on Concept Sets

In this framework, we try to leverage ranked lists from a set of neighbor concepts to improve the performance of an anchor concepts. Given a concept, the set of its $k = 3$ neighboring concepts is determined by FCS and ontology reasoning as described in setction 1.2. The set of ranked lists of the selected concepts are compared to seek the *agreement* such that videos which are ranked highly by multiple concepts are considered to be highly relevant. Specifically, for agreement, linear fusion is used to combine the scores from different concepts. Finally, regularization is performed to smooth these scores based on the video-to-video similarity derived from the textual modality.

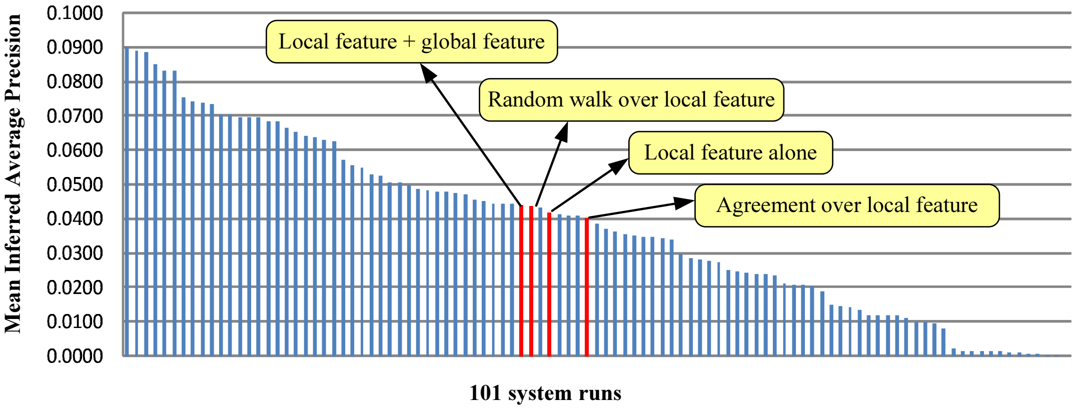

Figure 2: Mean average precision of all 101 SIN full version runs submitted to TRECVID 2010. Our submissions are marked in red.

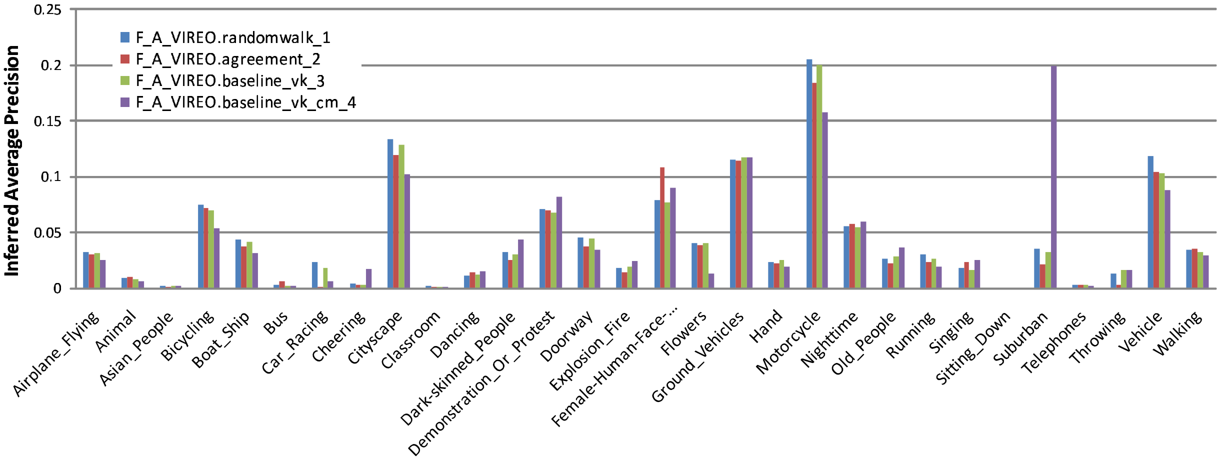

Figure 3: Per-concept performance of our submitted runs.

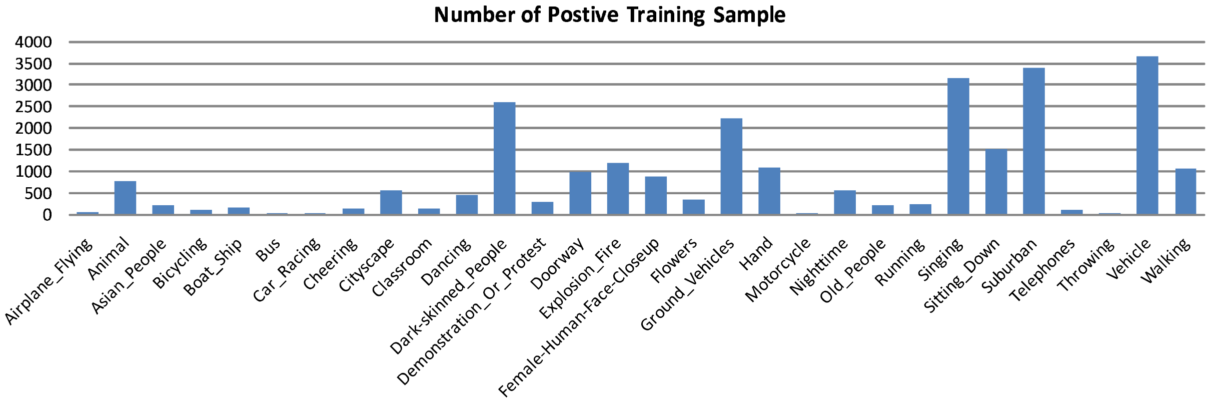

Figure 4: Number of positive training samples for evaluated concepts in TRECVID 2010

1.3 SIN Results and Analysis

For SIN, we submitted four runs, two baseline runs (Run 3–4) and two reranked runs using the proposed approaches (Run 1–2). Figure 2 shows the mean average precision (MAP) performances of all 101 submissions where our runs are marked in red color. Due to time constraint, we have ignored the spatial information which tends to be helpful in TRECVID 2009 [5]. This has resulted in a noticable drop in the performance of all four runs which lie around the median among all submissions. In general, semantic indexing (SIN) on a web video dataset is more difficult compared to previous years where the best result of this year is 0.09, which is much worse than last year (around 0.22). Figure 3 and 4 further detail the average precision (AP) of our four submissions and the number of positive training samples for each evaluated concepts, respectively. The performance of our baseline fluctuates from concept to concept. Generally, concepts with sufficient training samples, e.g., "vehicle" and "ground vehicle", can achieve higher AP than concepts that do not, e.g., "bus" and "throwing". The other factor that impacts the performance of concept detector is the diversity of the training sample. Despite having more training samples, the performance of concepts that are too generic or possess very diverse visual appearance, e.g., "singing" and "hand", is worse than concepts which are more narrowly defined or with well-correlated appearance, e.g., "airplane_flying" and "boat_ship".

Compared to the baseline which uses local feature alone, the performance improves only slightly from 0.041 to 0.043 by reranking on the context graph constructed using Flickr context similarity and ontology relationship. One reason is because the ontology reasoning provided by TRECVID for the 130 concepts is too weak to make useful inference. The depth of an hierarchical chain of concepts rarely exceeds two in the ontology and the concepts mainly converge at several 'generic' concepts, i.e., "person", "outdoor" and "indoor". In addition, the re-ranking framework assumes that the baseline results, though not perfect, have sufficient positive samples in the initial ranked list for analysis. Unfortunately, this is not the case for the baseline results, thus leading to very limited improvement for the random walk framework and even degradation of performance for the agreement-based framework. However, it is noteworthy that when a concept, e.g., "vehicle", with an acceptable baseline performance and sufficient supporting concepts, random walk does achieve apparent improvement from 0.103 to 0.118.

2 Known-Item Search

2.1 Text-based Search

For metadata search, we extracted the following information associated with the videos if they are present, i.e., "title", "description", "subject", "keywords", and "shotlist" (arranged based on their frequency in test set). For video transcript, we use the automatic speech recognition (ASR) result donated by LIMSI and Vecsys Research [15]. Depending on the run, the text from the metadata and ASR are concatenated to form a single document. The query is then compared against the document set. For similarity measure, instead of using traditional method

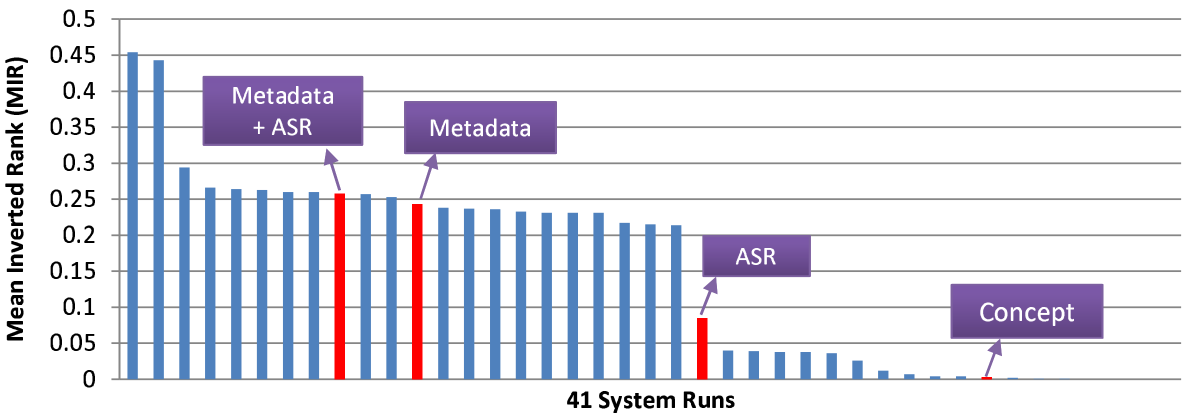

Figure 5: Mean inverted rank of all 41 KIS runs submitted to TRECVID-2010. Our submissions are marked in red.

Table 1: The total number of detected items within the top 1, 10 and 100, as well as mean inverted rank performance at top 100 (MIR@100) for all runs.

RunID	Description	Top1	Top10	Top100	MIR@100
Run 1	Metadata+ASR	57 (19%)	105 (35%)	154 (51%)	0.258
Run 2	Metadata	54 (18%)	98 (33%)	148 (49%)	0.243
Run 3	ASR	19 (6%)	33 (11%)	58 (19%)	0.085
Run 4	Concept-based	0 (0%)	4 (1%)	13 (4%)	0.030

like TF-IDF, our text search model adopts Okpai [13] using the application interface is provided by Lemur [14].

2.2 Concept-based Search

For concept-based search, we employ the Orthogonal Ontology-enriched Semantic Space (OS^2) [16, 7] to perform concept-to-query mapping. OS^2 is constructed by (a) constructing a semantic space using concepts as basis by ontological reasoning through WordNet [18] and then (b) performing spectral decomposition to transform the semantic space into a space with orthogonal bases. The space is learnt on the set of 130 concepts selected for this year's Semantic Indexing (SIN) task, which is much smaller than what we use for last year's TRECVID search task (374 concepts). For detection scores for these concepts, we use the CU-VIREO scores [17] donated by Columbia University. By OS^2, the top-3 nearest neighbor concepts of a query are extracted. For each keyframe in a video, the detection scores of the selected concepts are linearly fused and the final video score is given by the largest detection score among its keyframe pool.

2.3 Result Analysis

Figure 5 shows the performance of our systems compared to other runs for automatic search whereas Table 1 shows the number of detected items within the top 1, 10 and 100 ranked videos,

Table 2: MIR@100 for different information in the video metadata.

Modality	MIR@100
Description	0.1935
Title	0.1543
Subject	0.0810
Keywords	0.0049
Shotlist	0.0011

Table 3: Example queries with erroneous query-to-concept mapping. The concepts in the query are highlighted in bold.

ID	Query	Selected Concepts
9	Find the video of a view of a calm **stream** with **rocks** and green **moss**	plant, explosion, dancing
17	Find the video of **President Bush** standing near **sea vessels** with **Coast Guard members** talking about his pride of the **Coast Guard**, **immigration**, and **security issues**.	crowd, meeting, court
276	Find the video of a man **bowling** and the titles "Signifying nothing" and "**Bowling** for **Clitonbine**" shown on the video.	chair, golf, politics

as well as the mean inverted rank performance at top 100 (MIR@100) for all submitted runs. The text-based modalities (metadata and ASR) is able to deliver good retrieval performance for KIS where in average, the positive videos are found in the 4th position (MIR=0.258) in the ranked list of our best run (metadata+ASR). We further evaluate the individual performance of the information in the metadata and the result is given by Table 2. *Title* and *description* are the most reliable metadata information because they are found in more than 95% of the videos whereas the others in less than 50% of the videos. In general, the textual information, including ASR, complement each other well when combined into into a single document because they increases the chances of the query terms to be mapped to the terms in the positive document.

Concept-based search is clearly no longer effective for known-item search where the difficulties occurs at two levels. First, at the query-to-concept mapping layer, the items defined in the queries are too specific to be mapped properly to the 130 pre-defined concepts. The problem is further aggravated by the smaller number of available concept detectors this year. Table 3 gives several example queries with erroneous query-to-concept mapping. The concepts found in these queries cannot be mapped to concept dictionary either because of the small concept dictionary size or the concepts in the query are too specific (e.g., "President Bush" in Query 17). Second, the performance of the concept detectors from semantic indexing (SIN) is poor this year where the best run only has a mean inferred average precision (infAP) of 0.09 compared to previous year with infAP of around 0.22). Table 4 shows some examples with such problem. For these

Table 4: Example queries with some valid query-to-concept mappings but low concept score detection. The concepts in the query are highlighted in bold.

ID	Query	Selected Concepts
9	Find the video showing a **man** with full black **beard** and black **turban**	beard, asian, nighttime
225	Find the video of **baby girl** wearing a **yellow shirt** reading **book** on **wood floor**	girl, infant, chart
281	Find the video with a **horse**, **house**, **woman** and **child** feeding **chickens**, and the **narrator** saying, "**ladies** and **gentlemen** this is a **flying saucer**".	horse, girl, bird

queries, the detection score scores of the selected concepts are too low for the positive videos and as a result, they are mostly ranked out of the top 100 either because of a weak model or the selected concepts are too general to pinpoint any specific video. In short, KIS throws a totally new set of challenges that needs to be resolved by current concept-based search techniques.

3 Content-Based Copy Detection

3.1 Video-only copy detection

Our submitted runs for CCD are mainly based on our recent works on near-duplicate video search [11, 12]. Figure 6 illustrates the overview of our framework, which is composed of an online retrieval and an offline indexing part. For the offline indexing part, one frame is sampled every 2.5 second. Keypoints are extracted by Harris-Laplcian detector and described with SIFT [9]. A visual dictionary of 20K words is generated and each keyframe is represented using the Bag-of-Words (BoW) feature [8]. A inverted file is then employed to index the keyframe set.

For the online retrieval part, the query is processed by sampling one frame per 1.25 second. Each sampled frame is represented with BoW, and searched against the inverted file. To alleviate the information loss due to vector quantization, we impose both visual (Hamming Embedding (HE) [10]) and geometric verification (Enhanced Weak Geometric Consistency Checking (EWGC) [12]) on the visual word matches. We further incorporate Scale-Rotation Invariant Pattern Entropy (SR-PE) [11] as the post-processing to alleviate false matches introduced by BoW retrieval. Together with EWGC, SR-PE is employed to perform reciprocal validation since EWGC and SR-PE are two different ways in estimating the same linear transformation parameters (scale and rotation). Finally we employ 2D Hough transform (HT) to aggregate the scores from corresponding frames and localize the copy segments. Since we only conduct video-only copy detection, for video+audio detection task, we simply submit the same detection results for the corresponding queries.

We submitted three runs based on our video-only detection framework. The configurations of these three runs are detailed as follows.

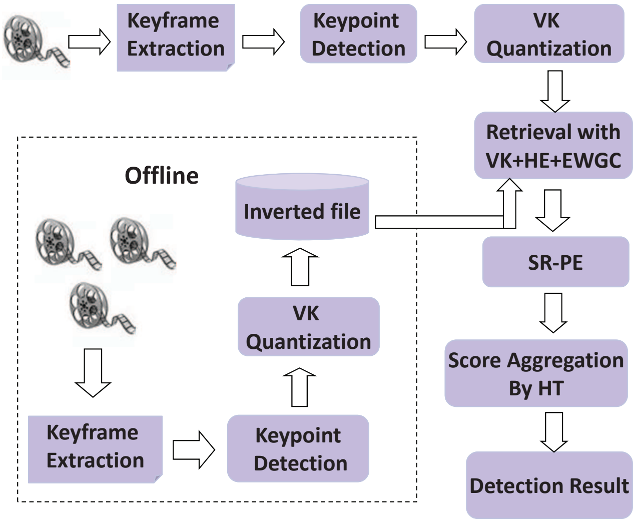

Figure 6: Video-only copy detection framework.

Vireo.m.BALANCED.srpe: This run employs HE, EWGC, SR-PE and 2D HT, as indicated in Figure 6. The similarity between the query and retrieved reference frames are measured by the multiplication of BoW similarity and the number of correct matches identified by SR-PE. These similarity values between query and reference frames are later aggregated by 2D HT. For each query, we only return the top-one retrieved video whose scores are above a threshold as our final submission.

Vireo.m.BALANCED.srpeflip: Repeating the same procedure as the previous run, we flip the query and perform the detection a second time. The retrieval results are then linearly fused with previous run before being fed into 2D HT. Similar to previous run, we return only the top one retrieved videos as our final submission.

Vireo.m.NOFA.srpeflip: This run is the same as the second run except that we choose a higher threshold to prune out more false alarms.

3.2 Reciprocal Validation

Based on the linear transformation parameters (scale and rotation) recovered from different estimation models in SR-PE and EWGC, we can perform reciprocal validation (RV) on visual word matches during the post-processing stage. In [11], affine transformation parameters, namely scale ($\hat{s}$) and rotation ($\hat{\theta}$), are estimated for a matched keypoint pair pq by referring to another matched keypoint pair $p'q'$ (assuming that point p and p' are in image I while q and q' are in image J). The scale $\hat{s}$ between image I and image J can be approximated by the distance ratio between $\overrightarrow{pp'}$ and $\overrightarrow{qq'}$ while the rotation $\hat{\theta}$ can be approximated by the angle between $\overrightarrow{pp'}$ and $\overrightarrow{qq'}$.

Alternatively, by EWGC [12], the transformation parameters scale ($\tilde{s}$) and rotation ($\tilde{\theta}$) for the matched pair, pq can also be estimated based on the keypoints characteristic scales and dominant orientations.

Table 5: Video-only performances of two BALANCED runs.

	Opt.NDCR				**Opt.F1**			
	srpe	srpeflip	m.median	m.best	srpe	srpeflip	m.median	m.best
T1	1.107	1.107	1.037	**0.005**	0.000	0.000	0.581	**0.974**
T2	1.107	1.107	1.026	**0.088**	0.000	0.000	0.449	**0.976**
T3	0.799	0.784	0.998	**0.001**	0.589	0.584	0.829	**0.976**
T4	1.107	1.107	1.003	**0.003**	0.000	0.000	0.824	**0.980**
T5	0.880	0.791	0.998	**0.003**	0.471	0.555	0.819	**0.981**
T6	0.679	0.679	1.003	**0.014**	0.622	0.622	0.822	**0.985**
T8	0.832	0.718	0.988	**0.050**	0.587	0.626	0.870	**0.981**
T10	0.924	0.848	0.988	**0.076**	0.508	0.548	0.811	**0.980**

Table 6: Video-only performances of NOFA run.

	Opt.NDCR			**Opt.F1**		
	srpeflip	m.median	m.best	srpeflip	m.median	m.best
T1	108.071	15.89	**0.086**	0.000	0.761	**0.974**
T2	108.071	90.19	**0.06**	0.000	0.607	**0.976**
T3	0.835	0.988	**0.056**	0.559	0.9	**0.976**
T4	107.995	1.00	**0.073**	0.486	0.856	**0.980**
T5	0.880	0.998	**0.039**	0.516	0.877	**0.981**
T6	0.679	1.003	**0.052**	0.650	0.896	**0.985**
T8	0.880	0.988	**0.079**	0.537	0.898	**0.981**
T10	0.856	0.988	**0.082**	0.542	0.894	**0.980**

$$\widetilde{s} = 2^{(s_q - s_p)}, \tag{5}$$

$$\widetilde{\theta} = \theta_q - \theta_p, \tag{6}$$

where s_p, s_q and θ_p, θ_p are the characteristic scales and dominant orientations for keypoints p and p, respectively.

To qualify as a valid match, $\hat{s}$ must be similar to $\tilde{s}$, while $\hat{\theta}$ must be similar to $\tilde{\theta}$. Conversely, this does not hold true for false matches. Therefore, pruning noisy visual word matches is achieved by deleting matches whose discrepancies ($\triangle = max\{|\hat{\theta} - \widetilde{\theta}|, |\hat{s} - \tilde{s}|\}$) from the estimated parameters are higher than a threshold. As $\hat{s}$, $\hat{\theta}$ and $\tilde{s}$, $\tilde{\theta}$ are estimated independently, such kind of reciprocal validation can be quite effective.

3.3 Experiment Result

Table 5 and 6 show the performances of three runs, measured in terms of NDCR and F1, compared to the best and median audio+video results[2]. Overall, the three submitted runs show unsatisfactory performances compared to the best run. For the transformations T3, T5, T6, T8 and T10, the performances of all three runs are still above the median level. As seen from Table 5, by fusing the retrieval results from original and flipped queries, the performances of *BAL-*

[2] The best and median results are estimated from the results officially released by TRECVID.

ANCED.srpeflip on these five types of transformations have seen considerable improvements. One possible reason is because when symmetric scenes and objects are flipped, the keypoints for true-positive segments are preserved and enhanced while the keypoints for negative segments will undergo more significant change and are thus less correlated. Our approach fails to identify video copy on transformations T1 (camcoding), T2 (Pic-in-Pic) and T4 (re-encoding with high compressing rate). This is mainly because BoW retrieval simply misses most of the true-positive frames in the reference set. This is due to the significant drop of video quality in the queries where many corresponding visual words are indexed into different bins in the inverted table. To verify our judgement, we performed brute force point-to-point matching between the query and ground-truth frames for the failure cases. Even when we engage a much robust matching algorithm, it turns out that only a small number of correct point matches can be found. To handle these types of video copies with severe visual transformations, further incorporating the audio information is an appropriate choice.

In terms of query processing and retrieval time, the *BALANCED.srpe* run takes approximately 152.7 seconds to complete the search for a query video of average time duration 71.7 seconds. For *BALANCED.srpeflip* and *NOFA.srpelip* runs, the average processing runtimes are exactly doubled.

4 Summary

This year, we explore integrating ontology context relationship for semantic indexing (SIN). We test two approaches, random walk on context graph and agreement seeking on neighboring concepts, for detection score refinement. Unfortunately, due to an unsatisfactory baseline result and weak reasoning relationship between concepts, the random walk approach only manages to achieve slight improvement and the agrement-based framework, which more aggressively re-rank the videos, degrades the performance. The poor performance for the baseline are attributed to the lack of training samples for certain concepts, the complexity of the concepts and the employment of a simple feature which neglects the spatial information. For known-item search, the text-based modality (metadata and ASR) is able to deliver good retrieval performance. Through combining all textual information into a single document, the chances of the documents to be mapped to query terms are enhanced since different kinds of textual information can complement each other well. In contrast, concept-based search is ineffective for known-item search. This is because the query is too specific to be mapped into a small dictionary of 130 pre-defined concepts. Moreover, the performance of concept detector from semantic indexing (SIN) is far from satisfactory to be able to support concept-based search.

For content based copy detection, our detection approach is solely based on visual contents of the video sequences represented by BoW. To alleviate the false alarms introduced by vector quantization, both visual and geometric verifications are adopted. According to our evaluation, the framework fails to detect video copies which undergo severe visual transformations such as camcoding, pic-in-pic and strong re-encoding. The keypoints feature becomes intolerant to these types of variations. In such context, to further consider the audio feature may be essentially helpful.

Acknowledgment

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 119709).

References

[1] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking througth random walk over document-level context graph," in *ACM Multimedia*, 2007.

[2] 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.

[3] 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.

[4] 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.
VIREO-374: http://vireo.cs.cityu.edu.hk/research/vireo374/
VIREO-WEB81: http://vireo.cs.cityu.edu.hk/vireoweb81/

[5] C.-W. Ngo, Y.-G. Jiang, X.-Y. Wei, W.-L. Zhao, Y. Liu, J. Wang, and S.-A. Zhu, "VIREO/DVMM at trecvid 2009: High-level feature extraction, automatic video search, and content-based copy detection," in *TRECVID workshop*, 2009.

[6] Y.-G. Jiang, J. Yang, C.-W. Ngo, A. G. Hauptmann, "Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study," in *IEEE Trans. on Multimedia*, 12(1):42–53, 2010.

[7] X.-Y. Wei, C.-W. Ngo and Y.-G. Jiang, "Selection of Concept Detectors for Video Search by Ontology-Enriched Semantic Spaces," in *IEEE Trans. on Multimedia*, 10(6):1085–1096, 2008.

[8] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," 2003, pp. 1470–1477.

[9] D. Lowe, "Distinctive image features from scale-invariant keypoints," in *International Journal on Computer Vision*, 60(2):91–110, 2004.

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

[11] W.-L. Zhao and C.-W. Ngo, "Scale-Rotation Invariant Pattern Entropy for Keypoint-based Near-Duplicate Detection," in *IEEE Trans. on Image Processing*, 18(2):412–423, 2009.

[12] W.-L. Zhao, X. Wu and C.-W. Ngo, "On the Annotation of Web videos by Efficient Near-duplicate video Search," in *IEEE Trans. on Multimedia*, 12(5):448–461, 2010.

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

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

[15] J. L. Gauvain, L. Lamel, and G. Adda. “The LIMSI Broadcast News Transcription System,” in *Speech Communication*, 37(1-2):89-108, 2002.

[16] X.-Y. Wei, C.-W. Ngo, “Ontology-Enriched Semantic Space for Video Search,” in *ACM Multimedia*, 2007.

[17] Y.-G. Jiang, A. Yanagawa, S.-F. Chang, C.-W. Ngo, “CU-VIREO374: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection,” *Columbia University ADVENT Technical Report #223-2008-1*, Aug. 2008.

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