Fusing Semantics, Observability, Reliability and Diversity of Concept Detectors for Video Search

Xiao-Yong Wei
Department of Computer Science
City University of Hong Kong
Kowloon, Hong Kong
xiaoyong@cs.cityu.edu.hk

Chong-Wah Ngo
Department of Computer Science
City University of Hong Kong
Kowloon, Hong Kong
cwngo@cs.cityu.edu.hk

ABSTRACT

Effective utilization of semantic concept detectors for large-scale video search has recently become a topic of intensive studies. One of main challenges is the selection and fusion of appropriate detectors, which considers not only semantics but also the reliability of detectors, observability and diversity of detectors in target video domains. In this paper, we present a novel fusion technique which considers different aspects of detectors for query answering. In addition to utilizing detectors for bridging the semantic gap of user queries and multimedia data, we also address the issue of "observability gap" among detectors which could not be directly inferred from semantic reasoning such as using ontology. To facilitate the selection of detectors, we propose the building of two vector spaces: semantic space (SS) and observability space (OS). We categorize the set of detectors selected separately from SS and OS into four types: anchor, bridge, positive and negative concepts. A multi-level fusion strategy is proposed to novelly combine detectors, allowing the enhancement of detector reliability while enabling the observability, semantics and diversity of concepts being utilized for query answering. By experimenting the proposed approach on TRECVID 2005-2007 datasets and queries, we demonstrate the significance of considering observability, reliability and diversity, in addition to the semantics of detectors to queries.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Design, Algorithms, Performance, Experimentation

Keywords

Detector Selection and Fusion, Concept-based Video Search

1. INTRODUCTION

Using semantic detectors for video search has been known as one of the most interesting approaches in bridging the "semantic gap". Detectors are the set of classifiers which are trained to detect the presence of semantic concepts such as *car* and *building* in videos or images. In video search, given a user query, detectors which could reflect the semantics of query are chosen for query answering. The video segments which contain the desired concepts, as indicated by detectors, are then retrieved, ranked and returned to users. The selected detectors basically serve as a bridge to narrow the gap between user semantics and multimedia features. This retrieval methodology is often referred to as concept-based video search. A recent simulation study in [7] predicts that a large pool of concept detectors, fewer than five thousands, with mean average precision (MAP) of 10% can already achieve high accuracy of search performance.

The open issues underlying the search methodology, apparently, are the appropriate mapping and fusion of detectors to queries. Given a user query like "*Find shots of military personnel or soldier together with military vehicle or weapons*" and a large pool of detectors, even human will find it not easy to map and rank the importance of each selected detector to the query. Generally speaking, the query-to-detector mapping involves the semantic reasoning [6], domain knowledge [3], or even the spatial-time dependent information [19] at the time a query being issued. For example, the detector *armored vehicle* can be associated to the aforementioned query by reasoning, the detector *explosion* can be selected by noticing the co-occurrence of *explosion* and *weapon*, and the detector *flag* may be picked if knowing that a country is in war recently. A subsequent interesting question after selection is: how to fuse the set of ultimately selected detectors which could probably involve concepts like *armored vehicle*, *explosion*, *flag*, *soldier*, *armed person*, *military personnel*, *tank*, *entertainment*, and *weapon*? These detectors are likely correlated, either complementarily for having parenthood relationship (e.g., *tank* and *armored vehicle*), supportive of one another (e.g, *soldier* and *weapon*), or contrasting to each other (e.g., *explosion* and *entertainment*). Basically, being able to understanding and exploiting the nature of detectors could ideally derive effective fusion strategy, which is a main issue addressed in this paper.

In essence, the selection and fusion of detectors requires support from multiple sources of information including ontology, statistics, and also the reliability of concept detection. This paper considers different aspects of semantic detectors as following for fusion in large-scale video search:

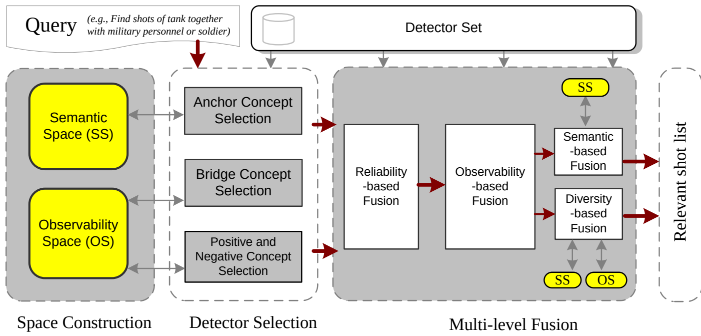


Figure 1: Overview of concept detector selection and multi-level fusion.

- *Semantics* refers to linguistic relatedness between concepts. A common approach for semantic measurement between query and detector is the use of ontology such as WordNet to reason their hyponym (is-a) relationship. The semantic reasoning, as pointed in [27], however relies only on a local view of sub-graph structure at where the concepts under investigation reside. Such reasoning technique does not allow uniform measurement of concept relatedness. We extend the work in [27] to build a semantic space (SS), which is a vector space facilitating uniform mapping of query to detectors. Different from [27], the proposed SS is an orthogonal space which also emphasizes the optimal coverage of semantic space built upon a vocabulary set which is given to describe the target video domain.

- *Observability* refers to the frequent occurrence of certain concepts in the target video domain. These concepts could correlate to each other, or even coexist in some unknown subspaces. These information is not directly observed from semantic reasoning. Similar to SS, we propose the building of a vector space, namely observability space (OS), to mine this piece of information. While SS is for narrowing the "semantic gap", OS addresses the bridging of "observability gap". Conventional correlation measures only provide local view of correlation among few concepts, and thus prohibit effective mining of correlated concepts in a global view. In contrast, OS offers a global and uniform view of how concepts co-occur in a vector space. In OS, given any two concepts, the subspace embedded by these concepts can be efficiently mined to infer useful detectors for video search. For instance, by knowing that *military vehicle* and *weapon* in the aforementioned query are also similar by observability analysis, a subspace of OS enclosed by the vectors of *military vehicle* and *weapon* can be formed. The set of detectors (e.g., *explosion*) which falls in this subspace could infer another piece of useful information for video search.

- *Reliability* refers to robustness of detectors in terms of detection performance. Intuitively, only robust detectors should be considered for query answering, or at least the robustness of detectors should be enhanced as best as possible before utilized for video search. To improve the reliability of a detector, we employ OS to uniformly determine the set of positively and negatively correlated detectors. The robustness of original detector is then enhanced by jointly fusing with the set of correlated detectors.

- *Diversity* refers to variety of detectors in answering query. For instance, in the aforementioned query, the set of selected detectors could be grouped into two types: person-related (e.g., *soldier*) and vehicle-related (e.g., *tank*). Considering each group of detectors separately during fusion, instead of each detector individually, could avoid the case that certain groups may have more selected detectors and thus bias the final search result. There are various ways of inferring diversity of detectors with respect to a query, for example by linguistic understanding of a query. In this paper, we measure diversity directly by similarity comparison in the proposed observability and semantic spaces.

Based on these four aspects of semantic detectors, we propose a novel fusion strategy as illustrated in the framework shown in Figure 1. Besides the space construction, the framework has two major parts: detector selection and multi-level fusion. In detector selection, detectors are selected separately from SS and OS. The detectors picked up from SS, named as anchor concept detectors, are semantically related to a query. The detectors chosen from OS, called as bridge concept detectors, are to bridge the observability gap of anchor detectors in the subspace of OS. During detector fusion, we consider different levels of fusion, each of which emphasizes one aspect of semantic detectors. In reliability-based fusion, positively and negatively correlated detectors are further selected from OS to enhance the robustness of anchor and bridge detectors. In observability-based fusion, the co-occurrence chance among concepts is exploited by fusing bridge detectors with their nearest anchor detectors. This not only enhances the reliability but also enriches the observability of anchor detectors, in a way that bridge and anchor detectors are co-operating to improve search result. Finally, the set of anchor detectors are either semantically fused or by further analyzing their diversity before fusion. The semantic-based fusion is conducted using SS, while diversity-based fusion involves combination of detectors in both semantic and observability spaces.

The remaining sections are organized as follows. Section 2 briefly surveys the existing related work. Sections 3 and 4 present the construction of the proposed semantic and observability spaces respectively. Section 5 describes the selection of detectors with SS and OS, while Section 6 presents the strategy for multi-level fusion of semantic detectors. Finally, Section 7 shows the experimental results for video search, and Section 8 concludes this paper.

2. RELATED WORK

Different from the traditional content-based retrieval, semantic search is enabled by pooling a set of concept detectors to bridge the gap between user semantics and low-level features. Two major efforts along this direction are the detection of semantic concepts and the utilization of concept detectors as "semantic filters" for video search. Since 2001, TRECVID (TREC Video Retrieval Evaluation) [25] sponsored by NIST has organized annual workshop, by releasing video benchmarks and system evaluations, for the related tasks including high-level feature extraction (HLFE) and video search. In HLFE, concept detectors are developed for video semantic annotation. In order to identify the right set of detectors to develop, collaborative efforts from various research organizations have been pooled in to assess the utility, observability and flexibility of concept detectors [17]. One typical example is the release of LSCOM (Large-Scale Concept Ontology for Multimedia) [17] which includes 834 semantic concepts and a collection of annotations (training examples) for 449 out of the 834 concepts. With LSCOM, two detector sets, Columbia-374 [29] and VIREO-374 [10], are also publicly released to share the sets of detectors developed based on the concepts in LSCOM. Another detector set commonly used is MediaMill-101 [5] which provides 101 concept detectors.

With the availability of various detector sets, concept-based video search is performed by assigning appropriate detectors to interpret query semantics. Various studies including [24, 19, 6, 2] have been reported regarding the usefulness of concept detectors for video search, compared to content-based search with low-level features and text-based search with keywords. Based on these studies, we can broadly categorize the existing works of query-to-detector mapping as: ontology-based [19, 6, 2], definition-based [6], statistic-based [3, 19], example-based [6, 16] and vector-based [18, 26].

In ontology-based mapping, text queries, which are usually short and imprecise, are mapped to detectors by using general purpose vocabulary such as WordNet. Specifically, detectors are chosen based on their semantic relatedness to query words. The relatedness is measured mostly based on traditional text-based ontology reasoning. For instance, Resnik [22] which utilizes information content to measure word-to-word relatedness is employed in [19, 6, 2]. Other ontology measures popularly used in video search include Wu and Palmer (WUP) [30], Jiang and Conrath (JCN) [9] and Lesk [15]. Text-based mapping, different from ontology-based approaches, compares query words directly with the text description associated to define concepts for identifying appropriate detectors. This mapping strategy is shown to be comparable to ontology reasoning in the experiments conducted by [6]. In statistic-based mapping, query words are further expanded with related terms prior to detector selection. The expanded terms, together with their weights, are learnt from training examples [3] or external information such as Internet [19]. In addition to text queries, multimedia-based queries in the forms of image/video examples may also be provided for search. With these queries, example-based mapping is often done by selecting the concept detectors which output high confidence to query examples, indicating the likelihood of corresponding concepts present in the queries. For instance, the approach in [6] selects the best confident detector for video search. For most of the existing approaches, when multiple detectors are selected, fusion of detectors is often done by weighting the significance of detectors towards queries. Depending on the category of mapping, most approaches assign weights to detectors, for example based on the ontology similarities of detectors to queries [19], or the detection scores of detectors to image/video examples [16].

Vector-based mapping is a relatively different strategy by constructing the semantic space or vector space for modeling concepts. The pioneering works in [18, 26] construct a vector space formed by the set of available concept detectors. In this space, a video shot is represented as a vector of model scores. The scores are computed based on the signal responses of detectors to shots. Contrasting to other mapping categories, no specific detector is selected, but rather all detectors are involved in the video search though each detector carries different weights. In [16], the idea of tf-idf originated from information retrieval, which weights the importance of a detector according to its appearance frequency, is adopted to further improve the search performance of vector space representation. Different from [16, 18, 26], the recent work in [27] constructs a semantic space by considering the all-pair ontology relationship between concepts. The constructed space is named as OSS (Ontology-enriched Semantic Space). Under this space, query words and detectors are represented as vectors, and the weights of detectors for performing fusion can be directly inferred from the space. Our proposed semantic space in this paper is indeed an improved version of OSS which also considers the orthogonality of space for optimal coverage of semantic concepts.

While the selection of detectors has been studied in various works, the fusion of detectors has not been actually paid much attention. Most fusion techniques, as mentioned in the previous paragraphs, are indeed formulated directly based on the underlying mapping strategies being used. The joint consideration of detector semantics, observability, reliability and diversity for effective video search has yet to be researched. As a fact to recognize the importance of this missing piece of works, there appear several works recently which analyze the properties of semantic concepts. For instance, the importance of utilizing concept co-occurrence is recognized in [11]. Different mining approaches including G-test [12], frequent itemset mining [28], shot clustering [12, 28] are experimented to identify useful concept patterns. An empirical study by [14] also confirms the co-occurrence information as a complementary view of concept semantics. The recent context-based concept fusion approach in [13] also attempts to discover and leverage the co-occurrence concept patterns for the re-ranking of search result. In this approach, based on the mutual information between detector scores and the pseudo-labels (positive/negative relevance) of shots, a fixed number of detectors are selected (75 in their experiments) and fused (based on detector scores) for boosting video search performance. Despite these works, the utility of semantic detectors, along the process of how these detectors are developed and in what ways the concepts co-occur in videos, for effective video search is still not fully understood. In particular, the fusion of various aspects of semantic detectors has not yet been seriously investigated.

3. BUILDING SEMANTIC SPACE

Selection of detectors involves semantic reasoning of concept relation. Instead of relying on ontology such as WordNet to perform reasoning, we construct an orthogonal semantic space for uniform comparison of concept similarity. Given a vocabulary set $\mathcal{V} = [C_1, C_2, \ldots, C_n]$ of n concepts, we want to represent each concept C_i as a vector in the semantic space (SS). The basis vectors in SS are viewed as

$$\vec{C}_{b1} \times \vec{C}_{b2} \times \ldots \times \vec{C}_{bm} \longrightarrow \mathbb{R} \quad (1)$$

where the set $\mathbf{C} = [\vec{C}_{b1}, \vec{C}_{b2}, \ldots, \vec{C}_{bm}]$ with $m \leq n$ are the bases learnt from $\mathcal{V}$, which ideally can approximate the real world semantics $\mathbb{R}$ as best as possible. SS emphasizes the optimal coverage of semantic space based on the available set of concepts in $\mathcal{V}$. To learn $\mathbf{C}$, we first utilize ontology to measure the pair-wise semantic relatedness of concepts. The similarities of concepts are encapsulated in a matrix which will be further transformed to a new and compact semantic space for concept representation. The resulting SS is basically orthogonal and ontology-enriched, allowing representation of concepts in vector form and facilitating consistence measurement of concept relatedness.

3.1 Orthogonal and Ontology-Enriched SS

Similar to [27], we represent the concept relatedness using WUP similarity [30] with WordNet as the ontology:

$$WUP(C_i, C_j) = \frac{2D(p_{ij})}{L(C_i, C_j) + 2D(p_{ij}))} \quad (2)$$

where D returns the depth of a concept, and L gives the path length of two concepts in WordNet. Applying Eq. (2) for all concept pairs in $\mathcal{V}$ forms a matrix $\mathbf{R}$.

We build SS based upon the ontology-enriched matrix $\mathbf{R}$. Eq. (1) is estimated using $\mathbf{R}$ as follows

$$\mathbf{C}^T\mathbf{C} = \mathbf{R} \quad (3)$$

The equation can be solved by performing spectral decomposition to $\mathbf{R}$ such that

$$\begin{aligned} \mathbf{R} &= V\Lambda V^T \\ &= (V\Lambda^{\frac{1}{2}}V^T)^T(V\Lambda^{\frac{1}{2}}V^T) \end{aligned} \quad (4)$$

where Λ is a matrix with all the eigenvalues of $\mathbf{R}$ on its diagonal, and V is the corresponding eigenvector matrix. We employ Schur decomposition [8] for Eq. (3). By equations (3) and (4), the set of basis vectors is estimated as

$$\mathbf{C} = V\Lambda^{\frac{1}{2}}V^T \quad (5)$$

where the SS spanned by basis vectors in $\mathbf{C}$ is orthogonal while enriched by ontology knowledge learnt from $\mathbf{R}$. Comparing to the semantic space in [27] which simply performs clustering on $\mathbf{R}$ and employing the medoid concepts as bases, the proposed SS emphasizes the optimal coverage of semantic space by building an orthogonal space.

A footnote to the building of SS is that the redundancy of $\mathbf{R}$ should be kept in minimum so as to guarantee the computational stability of spectral decomposition. In our approach, we achieve this by performing clustering on the column vectors of $\mathbf{R}$. Each vector indeed corresponds to the WUP similarity of a concept to all other concepts in $\mathcal{V}$. Subsequently, the size of $\mathbf{R}$ is reduced by removing redundant column vectors or equivalently omitting the redundant concepts in $\mathcal{V}$, prior to the computation of Eq. (4).

3.2 Concept Representation

Having the orthogonal SS, the available concepts in $\mathcal{V}$ can be easily projected and represented as vectors of dimension $m \leq n$. For unknown concept $C_u \notin \mathcal{V}$, the concept vector $\vec{C}_u$ is predicted as

$$\begin{aligned} \mathbf{C}^T\vec{C}_u &= \vec{R}_u \\ \vec{C}_u &= (\mathbf{C}^T)^{-1}\vec{R}_u \end{aligned} \quad (6)$$

where $\vec{R}_u$ is a vector obtained by computing the WUP of $\vec{C}_u$ to the concepts in $\mathcal{V}$. In Eq. (6), we employ Moore-Penrose pseudoinverse [21] to compute the inversion $(\mathbf{C}^T)^{-1}$.

With this representation, the semantic relatedness between two concepts can be directly measure with cosine similarity:

$$Semantic(C_i, C_j) = \frac{\vec{C}_i \cdot \vec{C}_j}{|\vec{C}_i||\vec{C}_j|} \quad (7)$$

Comparing to the traditional measures such as WUP which reasons similarity based on local view (sub-graph) of ontology, Eq. (7) provides a global view of semantic relatedness in vector space.

4. LEARNING OBSERVABILITY SPACE

Complementary to SS, observability space (OS) gives cues to concepts of how they co-occur in video domain. Concepts with high similarity in OS are not necessarily also semantically similar in SS. For instance, *car* and *road* are not semantically related in ontology but often co-occur in video shots. To build OS, we adopt similar procedure as SS to learn concept observability. The main difference is that the matrix $\mathbf{R}$ is computed by learning from the pair-wise concept co-occurrence, instead of semantic relatedness.

We employ Pearson product-moment (PM) correlation to compute concept observability:

$$PM(C_i, C_j) = \frac{\sum_{k=1}^{|\mathcal{T}|}(O_{ik} - \mu_i)(O_{jk} - \mu_j)}{(|\mathcal{T}| - 1)\sigma_i\sigma_j} \quad (8)$$

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. Applying Eq. (8) to the concept pairs in vocabulary set $\mathcal{V}$, we form an observability matrix $\mathbf{O}$, which is similar to how $\mathbf{R}$ is formed in SS. $\mathbf{O}$ will be further decomposed as in Eq. (4) to compute a new transformed space of observability. Using the same procedure as in building SS, OS which is orthogonal and spanned with basis vectors computed as in Eq. (5) is constructed.

An important property of OS, similar in spirit to SS, is the offering of a 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 $\mathbf{O}$. In other words, comparing observability of any two concepts is globally, instead of locally, measured in OS. Similar to Eq. (7), the observability score of two concepts is computed as

$$Observability(C_i, C_j) = \frac{\vec{C}_i \cdot \vec{C}_j}{|\vec{C}_i||\vec{C}_j|} \quad (9)$$

The observability similarity measured in OS indeed has advantages that the erroneous effect due to missing annotation can be minimized. As a matter of fact, which is also reported in [11], 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 somehow be discovered if *snow* is always annotated together with *mountain*, and *mountain* is happened to be labeled with *outdoor* in some sample shots.

5. CONCEPT SELECTION

Given a query and a pool of detectors, we select two groups of detectors from semantic and observability spaces respectively. The first group of detectors, denoted as $\mathcal{A}$, refers to anchor concepts which describe the semantics of query terms. The second group, denoted as $\mathcal{B}$, is composed of bridge concept detectors which refer to the concepts found in the subspaces of OS formed by some of anchor concepts in $\mathcal{A}$. The detectors in $\mathcal{A}$ and $\mathcal{B}$ are chosen with the aim of bridging the semantic and observability gaps between the query and multimedia features.

5.1 Query-to-Concept Semantic Mapping

Let a user query be $Q = \{q_1, q_2, \ldots\}$ where q_i is a word carrying semantic meaning. We consider only nouns and gerunds of a query, assuming that noun mostly indicates the name of place, thing or person, and gerund describes an action (e.g., *walking*, *running*) of event. With SS, each q_i is represented as a vector via Eq. (6). Given a detector set $\mathcal{D}$, the most relevant detector to q_i is retrieved as an anchor concept of Q. Ultimately, by considering all the query terms, we have

$$\mathcal{A} = \bigcup_{q_i \in Q} \underset{C_j \in \mathcal{D}}{\operatorname{argmax}} \{Semantic(q_i, C_j)\} \tag{10}$$

where $\mathcal{A}$ includes a set of detectors, each refers to an anchor concept being semantically selected to interpret Q. In Eq. (10), the mapping from words to detectors is performed on the basis of one-to-one. In other words, the number of chosen detectors is at most equal to the amount of words in Q, i.e., $|\mathcal{A}| \leq |Q|$.

5.2 Detector Mining in OS

The detectors in $\mathcal{A}$ emphasize the semantic aspect of Q. Some of them may frequently appear together in a target video domain. These detectors could be projected to OS to observe the frequency of co-occurrence among each other. The detectors which are clustered in OS could be further utilized to mine another set of detectors, namely bridge concept detectors, for query answering in video search. For instance, by knowing the detectors *car* and *road* are close in OS, one could probably mine another more useful detector *car_on_road* by traveling along the subspace formed by the concept vectors of *car* and *road*.

Given anchor concepts, we first perform clustering to group concepts which are close in OS. This forms several subspaces of OS, in which each corresponds to a cluster of anchor concepts. For a cluster with two concepts, its subspace is a parallelogram enclosed by the two concept vectors. For a cluster of more than two concepts, the subspace is basically a cube or super-cube bounded by the concept vectors. Geometrically, it is not difficult to mine bridge concept detectors, by simply determining whether their concept vectors reside in the super-cube of a cluster. Nevertheless, due to rounding error when building OS, it is possible that the vector of a bridge detector lies outside the subspace of its cluster. Thus, we instead identify bridge concepts based on the rule that given any two anchor concepts in a cluster, the observability score of a bridge concept to any of the anchor concepts should be larger than the two anchor concepts themselves. This rule can be easily implemented with a function $M(.)$ which compares the observability scores of concepts. Ultimately, bridge concept detectors are identified as

$$\mathcal{B} = \bigcup_{\mathcal{G}_A \in \Omega} \left\{ \bigcap_{\{A_i, A_j\} \subseteq \mathcal{G}_A, A_i \neq A_j} \{C | M(C, A_i, A_j), C \in \mathcal{D}\} \right\} \tag{11}$$

where Ω denotes the set of anchor concept clusters, and $\mathcal{D}$ denotes the detector set. $M(.)$ confirms whether a bridge concept detector C lies within two given anchor concepts A_i and A_j, by comparing the observability scores of $\{C, A_i\}$, $\{C, A_j\}$ and $\{A_i, A_j\}$.

6. MULTI-LEVEL DETECTOR FUSION

In addition to the $\mathcal{A}$ and $\mathcal{B}$ detector sets, we also introduce the positive ($\mathcal{P}^+$) and negative ($\mathcal{N}^-$) detector sets in this section. The fusion of concept sets, i,e., $\{\mathcal{A}, \mathcal{B}, \mathcal{P}^+, \mathcal{N}^-\}$, is conducted in a progressive manner with one set of detectors being fused one at a time. At the lowest level of fusion, $\mathcal{P}^+$ and $\mathcal{N}^-$ are combined to enhance the reliability of detectors in set $\mathcal{B}$. With one level up, the reliability-enhanced detectors will be further fused in observability space to support set $\mathcal{A}$. Ultimately, detectors in $\mathcal{A}$ are fused in semantic space for query answering. The multi-level fusion effectively exploits the different facets of detectors, including reliability, observability, semantic and diversity, at different levels of fusion.

6.1 Reliability-based Fusion

Both positive and negative concepts are known to be useful for semantic detection and search [11, 13, 7]. On one hand, positively correlated concepts, each with its own detection reliability, could assist each other in boosting concept detection performance. On the other hand, negatively correlated concepts could provide effective evidence to confirm or to decline the existence of other concepts. Aiming to enhance the detection reliability, we utilize the sets of $\mathcal{P}^+$ and $\mathcal{N}^-$ detectors as supporting concepts to boost the detection performance of $\mathcal{A}$ and $\mathcal{B}$.

Given a concept $C \in \mathcal{A}$ or $\mathcal{B}$, the top-k most positively and negatively correlated concepts to C are picked from OS. Let $\mathbf{D}$ as a function for concept detection, and $\mathbf{D}(C)$ as the C-concept detector. $\mathbf{D}(C)$ outputs a list which is composed of the detection scores of retrieved items. The reliability of $\mathbf{D}(C)$ is improved by

$$\begin{aligned} \hat{\mathbf{D}}(C) &= \mathbf{D}(C) \\ &+ \frac{1}{|\mathcal{P}^+|} \sum_{P_i \in \mathcal{P}^+} Observability(P_i, C) \times \mathbf{D}(P_i) \\ &- \frac{1}{|\mathcal{N}^-|} \sum_{N_i \in \mathcal{N}^-} Observability(N_i, C) \times \mathbf{D}(N_i) \end{aligned} \tag{12}$$

where the improved detector $\hat{\mathbf{D}}(C)$ is a linear fusion of original detector with a set of positive and negative detectors.

The weight of a detector is dependent on the observability score based on Eq. (9). In our current approach, we select top-3 positive and top-3 negative concepts for each detector in $\mathcal{A}$ and $\mathcal{B}$ for fusion.

It is worth to notice that bridge and positive detector sets both exploit the co-occurrence of concepts, although in different ways. The former is to bridge the observability gap of anchor concepts, while the later is mainly for enhancing detection performance. For example, the detector *urban_building* is a positive detector of *road* but not a bridge detector of *car* and *road*. This is because *urban_building* only show strong correlation with *road* but not *car*.

6.2 Observability-based Fusion

The reliability-enhanced detectors of bridge concepts, i.e., $\hat{\mathbf{D}}(C)$, are further used to support the anchor concept set $\mathcal{A}$. Basically $\mathcal{B}$ consists of the set of concepts found in OS which could bridge the observability gap of concept pairs in $\mathcal{A}$. Similar to Eq. (12), we linearly fuse the bridge concepts together with anchor concepts to enhance the observability of concept pairs in $\mathcal{A}$. As explained in Section 5.2, concepts in a pair are semantically similar and always co-occur. For each $B \in \mathcal{B}$, we assign B to the nearest anchor concept $A \in \mathcal{A}$ in OS. For an anchor concept A associated with one or more B concepts, the observability-based fusion is conducted as

$$\begin{aligned}\bar{\mathbf{D}}(A) &= \hat{\mathbf{D}}(A) \\ &+ \frac{1}{|N(A)|}\sum_{B_i \in N(A)} Observability(B_i, A) \times \hat{\mathbf{D}}(B_i)\end{aligned} \quad (13)$$

where $N(A)$ is the set of bridge concepts whose nearest anchor concept is A. Detector $\bar{\mathbf{D}}(A)$ is an enriched version of the reliability-enhanced detector $\hat{\mathbf{D}}(A)$, by further considering the linear fusion of bridge concepts weighted with their observability scores. Eq. (13) is indeed similar to Eq. (12), though the aim is beyond the reliability enhancement. The observability of $\bar{\mathbf{D}}(A)$ is enriched by $B_i \in N(A)$ detectors, where their reliability has in turn been enhanced by their corresponding positive and negative concept detectors.

6.3 Semantic-based Fusion

The fusion of anchor concept detectors is performed directly in SS by employing the set of $\bar{\mathbf{D}}(A)$ detectors. The fusion is basically based on the semantic similarity of anchor concepts to query Q. Let I as a video shot, the similarity of Q to I is determined as

$$Sim(Q, I) = \sum_{A_i \in \mathcal{A}} Semantic(q_i, A_i) \times Score(\bar{\mathbf{D}}(A_i), I) \quad (14)$$

where q_i is a query term which maps to concept A_i. $Score(.)$ outputs the probability of finding A_i on I when the detector $\bar{\mathbf{D}}(A_i)$ is applied. The detection probability is then weighted based on the semantic relatedness of q_i and A_i measured in SS.

6.4 Diversity-based Fusion

Weighting detectors individually in SS has the disadvantage that the diversity of detectors is not addressed during fusion. For instance, given the set of anchor concepts $\mathcal{A}$={*person*, *face*, *police*, *newspaper*}, the search results will be biased by the first three concepts related to people. The diversity-based fusion makes use of the anchor concept clusters, which is a by-product of finding bridge concepts introduced in Section 5.2, to fuse detectors according to the clusters they belong to. The basic idea is first to combine the scores of a video shot, i.e., $Score(\bar{\mathbf{D}}(A_i), I)$, outputted by the detectors of an anchor cluster. The final score of I is then computed by fusing the combined scores from different clusters.

Let $\mathcal{G}_A$ be an anchor concept cluster formed in OS. The score of a shot, denoted as $L(I)$, computed based on the set of detectors in $\mathcal{G}_A$ is defined as

$$L(I) = \sum_{A_i \in \mathcal{G}_A} Observability(A_i, M_A) \times Score(\bar{\mathbf{D}}(A_i), I) \quad (15)$$

where

$$M_A = \frac{1}{|\mathcal{G}_A|}\sum_{A_i \in \mathcal{G}_A} A_i \quad (16)$$

Q is not involved in Eq. (15) since query terms q_i cannot be represented in OS. Instead, a "virtual concept" which is actually the mean vector M_A of an anchor cluster is utilized to weight the score of a detector. The weight is determined based on the similarity of A_i to M_A in OS.

Let Ω as the set of anchor clusters, the ultimate similarity between query Q and shot I is determined by fusing the scores $L(I)$:

$$DSim(Q, I) = \sum_{\mathcal{G}_A \in \Omega} \max_{A_i \in \mathcal{G}_A} \{Semantic(q_i, A_i)\} \times L(I) \quad (17)$$

where the similarity measure is conducted in SS. The weight of $L(I)$ is set by using the maximum semantic similarity among the pairs of q_i and A_i in an anchor cluster $\mathcal{G}_A$.

7. EXPERIMENT

7.1 Dataset and Evaluation

We conduct experiments using datasets (TV05, TV06, TV07) from TRECVID 2005 to 2007 [25] respectively, involving a total of 72 search queries. TV05 and TV06 datasets are composed of news videos in English, Chinese and Arabic. TV07 dataset contains mainly the documentary videos from the Netherlands. We use only the testing datasets for experiments, where there are 85 hours (45,765 shots), 150 hours (79,484 shots) and 50 hours (18,142 shots) of videos for 2005 to 2007 datasets respectively. Each dataset comes along with 24 search queries covering the search of person-thing, event, place, name entity, or any combination of them. Table 1 lists some of the 72 queries used in our experiments. We only use text queries, which are mostly short and abbreviated with few words, for testing.

For detector set, we use VIREO-374 [10] which is composed of detectors for 374 LSCOM semantic concepts. The detectors are trained using TRECVID 2005 development set based on the annotations provided by LSCOM. Each detector is associated with three SVM classifiers trained with local interest point features, grid-based color moment and wavelet texture respectively. The outputs of three classifiers are combined as the detection score with average fusion. We remove those detectors that have different description in LSCOM and WordNet, resulting in a set of 244 detectors.

Table 1: Examples of search queries.

ID	Query
156	Tennis players on the court, both players visible at the same time
169	One or more tanks or other military vehicles
173	One or more emergency vehicles in motion
181	US President George W. Bush, Jr. walking
183	Water with one or more boats or ships
187	One or more helicopters in flight
188	Something burning with flames visible
192	A greeting by at least one kiss on the cheek
193	One or more smokestacks, chimneys, or cooling towers with smoke or vapor coming out
199	A person walking or riding a bicycle
201	A canal, river, or stream with some of both banks visible
205	A train in motion
206	Hills or mountains visible
207	Waterfront with water and buildings
212	A boat moves past
214	A very large crowd of people
215	A classroom scene with one or more students
218	One or more people playing musical instruments such as drums, guitar, flute, keyboard, piano, etc.
219	The Cook character in the Klokhuis series
220	Grayscale shots of a street with one or more buildings and one or more people

In the experiments, the retrieved items (shots) are ranked according to their score to the selected concept detectors. The search performance is then evaluated with mean average precision (MAP), where AP is defined as

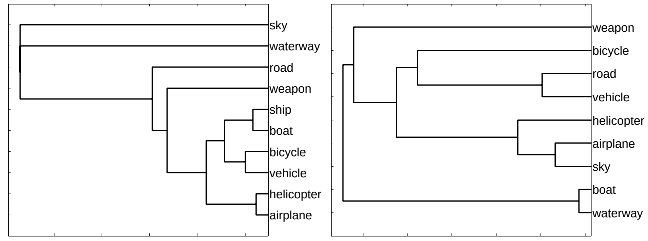
$$\mathrm{AP} = \frac{1}{\min{(R,k)}} \sum_{j=1}^{k} \frac{R_j}{j} I_j \tag{18}$$

where R is the number of relevant shots to a search topic, R_j is the number of relevant shots in the top-j retrieved shots, and $I_j = 1$ if the shot ranked at j^{th} position is relevant and 0 otherwise. MAP is the mean AP over all search queries. In the experiments, unless otherwise stated, we set $k = 1000$ following TRECVID evaluation.

7.2 Space Construction

This sub-section mainly details the construction of semantic and observability spaces. Examples will be given to illustrate how SS and OS are utilized for effective search. We use LSCOM concepts as the available vocabulary set. Among the concepts, 572 of them are selected and included in $\mathcal{V}$ for space construction, by discarding those concepts not defined in WordNet or being synonym of the existing concepts.

In building SS, the matrix $\mathbf{R}$ in Eq. (3) is first formed by computing the pair-wise WUP similarity of all concepts in $\mathcal{V}$. To facilitate the decomposition of $\mathbf{R}$ in Eq. (4), we further reduce the size of $\mathbf{R}$ by performing clustering. As in [27], we employ agglomerative algorithm to first hierarchically clusters the column vectors in $\mathbf{R}$ as a dendrogram. Note that each column vector ($\vec{R}_i$) could actually represent a concept vector, where each component (r_{ij}) of the vector means the WUP similarity between concepts C_i and C_j. The inconsistency coefficient [1], which specifies the degree of tightness between two sub-graphs which are linked by an edge in a dendrogram, is then used as a cue to group the set of column vectors into clusters. An advantage of using inconsistency coefficient is that the number of clusters is not a required parameter for clustering. The medoid vectors of clusters, each corresponds to a concept in the vocabulary set $\mathcal{V}$, are subsequently selected to form $\mathbf{R}$. This ultimately reduces the redundancy of concepts in $\mathcal{V}$ from 572 to 366 concepts, implying that the size of $\mathbf{R}$ is reduced to 366×366. By decomposing $\mathbf{R}$ using Eq. (4), we obtain the set of 366



(a) Semantic space (b) Observability space

Figure 2: Grouping of concepts in the dendrograms generated by SS (a) and OS (b).

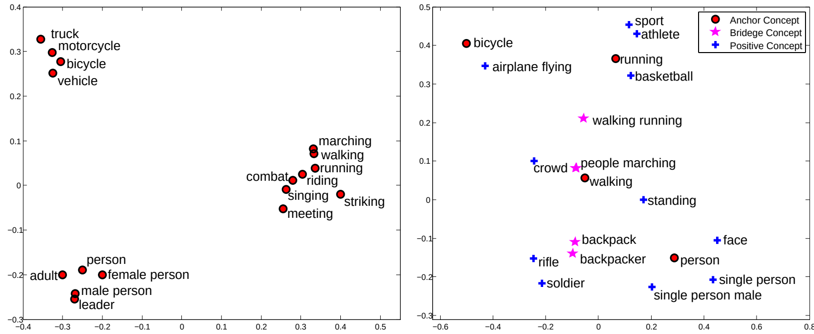
basis vectors to represent SS.

The building of OS is similar to SS, except that the matrix $\mathbf{O}$ is formed by Eqn (8). We use the available annotations in LSCOM as the training set $\mathcal{T}$ to compute Eq. (8). Similar to SS, the size of $\mathbf{O}$ is also reduced by performing agglomerative clustering. This ultimately leads to a set of 253 basis vectors to represent OS. Figure 2 shows two dendrograms generated during the construction of SS and OS respectively. The dendrograms illustrate the different merging processes of vehicle-related and scene-related concepts. In Figure 2(a), the concepts are grouped separately at different stages into scene, and ground, water and air vehicle related groups. While in Figure 2(b), the grouping is based on the co-occurrence of vehicle and scene related concepts. For example, *sky* is grouped together with *airplane* and *helicopter* at the early stage of merging. The forming of the dendrograms indeed indicates the very different nature of SS and OS, which also illustrates how semantics and observability can complement each other in selecting detectors for query answering.

Figure 3 shows an example of how concepts are distributed in semantic and observability spaces. For visualization purpose, the spaces are projected to 2D using multi-dimensional scaling technique. In 3(a), there are basically three clusters of concepts, respectively, related to person, vehicle and event. A very different view of concept distribution is seen in 3(b) when considering their observability. For example, the person-related concepts such as *crowd* are close to *walking*. In Figure 3, when the search query ID-199 is issued, the anchor concept detectors selected from Figure 3(a) are {*person*, *walking*, *running*, *bicycle*}. Note the the concept *riding* is not a detector in VIREO-374 and thus is not selected. These detectors are further grouped into two anchor concept clusters, $\mathcal{G}_1$ = {*person*, *walking*, *running*} and $\mathcal{G}_2$ = {*bicycle*}, in 3(b). The bridge concept detectors such as *walking running*, *people marching* and *backpacker* are subsequently selected from 3(b) based on the subspace formed by $\mathcal{G}_1$. The selection of the detectors *people marching* and *backpacker* indicates the presence of novel concepts to the query. The concepts cannot be semantically inferred from 3(a) but probably useful for retrieval due to their co-occurrence relationship with the anchor detectors. Besides bridge concepts, positive concept detectors such as *sport*, *crowd* and *face* will be further selected from 3(b) to enhance the detection reliability for this particular query.

7.3 Video Search Performance

We compare four types of detector selection and fusion strategy proposed in this paper for video search. The runs being experimented are:



(a) Semantic space (b) Observability space

Figure 3: Mapping detectors to query ID-199 " Find shots of a person walking or riding a bicycle": (a) distribution of concepts in SS, (b) selecting detectors from OS based on the anchor concepts chosen from SS.

- S-only: Semantic-based fusion using anchor concept detectors selected from SS.
- S+O: Anchor and bridge concept detectors selected respectively from SS and OS are used for observability-based and semantic-based fusion.
- S+OR: Consider anchor, bridge, positive and negative concept detectors for reliability, observability and semantic based fusion.
- S+ORD: Similar to S+OR, but consider diversity-based fusion instead of semantic fusion in Eq. (14).

7.3.1 *Top-k Search Performance*

Table 2 shows the MAP of four runs on different years of TRECVID datasets. To better analyze the performance, we list the MAP over the top $k = 30$, 50, 100 and 1000 ranked shots. Constant improvements are basically observed throughout the experiments when more aspects of detectors are considered. In most cases, an improvement of 5% to 8% is observed when one more aspect of detectors is taken into account. The improvement is particularly obvious for TV06 and TV05 datasets when considering AP of top 30 retrieved shots. For both datasets, an improvement of about 30% of MAP is observed when comparing S+ORD to S-only. Comparatively, less improvement (22%) is observed in TV07. This is indeed not completely surprise because the detector set VIREO-374 is trained mostly based on news videos annotated in LSCOM while TV07 is mainly composed of documentary videos. When considering more top-k retrieved shots, the improvement of MAP, nonetheless, become more obvious for TV07 than TV06 and TV05. From our analysis when browsing through the retrieved shots, more relevant shots are observed within top-100 list in TV05 and TV06. While in TV07, the relevant shots tend to spread throughout the search list. When considering MAP of the first 1000 retrieved shots, S+ORD gets an improvement of 35% compared to S-only in TV07, and of around 10% and 3% improvement in TV06 and TV05 respectively.

To verify whether the performance improvement is by chance, we also conduct significance test. The test is performed based on MAP of top $k = 1000$ retrieved shots on the three datasets. We employ the randomization test [23] suggested by TRECVID, where the target number of iterations used in the randomization is set to $100,000$. At the 0.05 level of significance, S+ORD is significantly better than

Table 2: Search performance on different years of TRECVID datasets. AP-30 means MAP over the top $k = 30$ ranked shots.

Set	Run	AP-30	AP-50	AP-100	AP-1000
TV07	S-only	0.213	0.125	0.078	0.029
	S+O	0.231	0.136	0.083	0.035
	S+OR	0.243	0.155	0.086	0.038
	S+ORD	**0.261**	**0.163**	**0.087**	**0.039**
TV06	S-only	0.374	0.234	0.114	0.044
	S+O	0.416	0.250	0.123	0.044
	S+OR	0.443	0.273	0.138	0.046
	S+ORD	**0.492**	**0.295**	**0.159**	**0.049**
TV05	S-only	0.749	0.448	0.234	0.123
	S+O	0.843	0.527	0.291	0.125
	S+OR	0.904	0.555	0.327	0.126
	S+ORD	**0.973**	**0.581**	**0.343**	**0.127**

Table 3: MAP performance on different query types. The () indicates the number of queries.

	Event (31)	PT (19)	Place (14)	NE (12)
S-only	0.0461	0.0716	0.1346	0.0039
S+O	0.0464	0.0785	0.1383	0.0037
S+OR	0.0490	0.0811	0.1434	0.0037
S+ORD	**0.0507**	**0.0814**	**0.1439**	**0.0065**

three other runs, while S+OR is better than two other runs and S+O is also better than S-only.

7.3.2 *Performance based on Query Types*

To further consider the impact of detectors toward different types of search queries, we group the 72 tested queries into four major types: event, person+thing (PT), place and name entity (NE). The grouping is based on the query classification given by TRECVID [20]. Table 3 shows the search performances for MAP over top $k = 1000$ retrieved shots. Note that there are few queries belong to more than one class. For PT, we only show the MAP of person+thing only queries because nearly all queries given by TRECVID are related to person+thing.

Due to the limitation of space, we do not show the MAP of each query in Table 3. There are, however, several interesting observations worth to discuss when browsing through the detailed search results: 1) observability-based fusion shows constant improvement for PT and place type queries; 2) diversity-based fusion shows constant improvement mainly for event type queries; and 3) reliability-based fusion shows improvement for all query types except NE. For PT and place related queries, there are always abundant detectors in VIREO-374 which can be correctly mapped as anchor con-

Table 4: Average number of selected detectors for each query type.

	Anchor	Bridge	Positive	Negative
Event	2.74	3.03	5.13	8.52
Person+Thing	2.42	4.16	5.37	6.95
Place	2.69	3.78	5.62	7.46
Name entity	2.08	4.08	4	7.17
Average	2.57	4.01	5.28	7.54

cepts. This is one reason that PT and place queries overall achieve higher MAP. More importantly, there are also plenty of detectors which are identified as bridge concepts in OS during observability analysis. To show the statistics, Table 4 lists average number of detectors being picked up. On average there are around 4 bridge detectors selected per PT/place query, compared to about 2.5 anchor concept detectors selected in SS. These bridge concept detectors can always boost the search performance, indicating the effectiveness of observability-based fusion for PT and place queries.

For event type queries, there are normally less bridge concept detectors, on average 3 detectors, being assigned. These detectors comparatively are less helpful. As an example, for a PT query with anchor concept detectors *person* and *road*, the detector corresponding to the bridge concept *walking* will be chosen. For an event query with anchor concept detectors *person* and *walking*, the detector *people marching* will be selected as a bridge concept detector. Because *walking* and *people marching* already have certain degree of redundancy, the improvement due to *people marching* is less compared to the use of *walking* as a bridge detector to specify another more general anchor concept *people* in the PT query. For most event queries, on the other hand, diversity-based fusion is found to be useful. There are 18 event queries with more than one anchor concept cluster. Constant improvements are always observed when comparing S+ORD and S+OR for these queries. For the majority of PT and place queries, however, normally only one anchor cluster found. As a result, the improvement is less obvious. Among all the query types, there is no improvement observed in NE type queries. This is not surprise because most terms in NE queries are too specific to match any detector. For most cases, *face* is matched as anchor concept and *government Leader* is matched as bridge concept. Because it is already difficult to select appropriate anchor concepts, further selecting bridge, positive and negative concepts normally does not lead to improvement for NE queries.

For the 72 tested queries, on average there are approximately 20 detectors being chosen per query, as listed in Table 4. Among them, there are about 5 positive and 7 negative detectors being selected. For S+OR run, these relatively larger set of detectors successfully boosts the search performance. Particularly, the place queries, which tend to have plenty of useful indoor/outdoor detectors complementary each other for reliability-based fusion, achieve the most improvement. To verify the performance improvement of using different fusion strategies for different query types, we also conduct significance test. Table 5 shows the results using randomization test [23] at the 0.05 level of significance. Overall, reliability-based fusion is always proven to be useful. Furthermore, considering diversity in addition to semantics is effective for event type queries, while considering observability is shown to be significant for PT/place type queries.

Table 5: Significance test at 0.05 level ($X \gg Y$ indicates that X is significantly better than Y).

Query type	Fusion approach
Event	S+ORD $\gg$ S+OR $\gg$ S+O, S-only
Person-Thing	S+ORD $\gg$ S+O $\gg$ S-only;
	S+OR $\gg$ S+O
Place	S+ORD $\gg$ S+O $\gg$ S-only
	S+OR $\gg$ S+O
Name entity	S+ORD $\gg$ S+O; S+OR $\gg$ S+O

Table 6: Performance comparison of different measures across different years and different query types of TRECVID datasets.

	S+ORD	OSS	RES	JCN	WUP	Lesk
2007	**0.039**	0.025	0.014	0.016	0.018	0.015
2006	**0.049**	0.041	0.039	0.04	0.04	0.041
2005	**0.127**	0.119	0.052	0.073	0.083	0.096
Event	**0.051**	0.044	0.019	0.013	0.033	0.035
PT	**0.081**	0.063	0.051	0.052	0.058	0.049
Place	**0.144**	0.128	0.064	0.115	0.082	0.113
NE	**0.006**	0.003	0.002	0.001	0.002	5E-04

7.4 Comparison to Ontology Reasoning

To further confirm the merit of jointly fusing observability, reliability and diversity in addition to semantics, we also compare S+ORD with five other measures using only semantic reasoning. These measures include OSS [27], Resnik (RES) [22], JCN [9], WUP [30], and Lesk [15]. Except OSS which builds a semantic space, other measures basically use information such as glosses (Lesk), path length and depth (WUP) of concepts in WordNet for reasoning relationship of query and detector. For RES and JCN, information content, which is estimated based on the one-million-word Brown Corpus of American English [4], is also used. In the experiments, each measure selects the top-3 most similar detectors per query. The search results returned by detectors are then linearly fused. Depending on the measure being used, the weight of a detector is set equal to its similarity to query.

Table 6 shows the detailed experimental results, which are listed separately according to query types and datasets. Basically S+ORD constantly outperforms all other measures. The improvement, in addition to attributing to the ability of using more detectors (on average 20 against 3 per query), is also largely due to the success identification of relevant and novel concept detectors for fusion throughout the process. For example, in query ID-169, the relevant concept *armored vehicle* is selected as bridge detector, while the novel concepts *weapon* and *entertainment* are chosen as positive and negative detectors respectively. These detectors are not found by other measures. Comparing the performances based on query types, S+ORD basically shows much better MAP, except the NE type queries. Compared to OSS, improvements of 16% for event, 29% for PT and 13% for place queries are achieved. Compared to other popular ontology measures such RES, the improvements are even obvious (168% for event, 59% for PT and 125% for place queries). We also conduct significance test to confirm the improvements. At the 0.05 level of significance using randomization test, S+ORD is significantly better than five other measures.

8. CONCLUSION

We have presented our proposed approach in fusing the different aspects of semantic detectors for video search. Two vector representations, semantic and observability spaces, are constructed to assist the selection and fusion of detectors. The fusion is novelly conducted in a multi-level fashion where each level emphasizes one aspect of detectors. Experiments on TRECVID datasets demonstrate the significance of fusing the semantics, observability, reliability and diversity of detectors for effective video search. While the results also indicate that the proposed fusion strategies are not necessarily significant for all types of queries, we do show several interesting observations. These include the findings that considering diversity is helpful for event type queries, while observability analysis is useful for mining and combining bridge concept detectors which are useful for person+thing and place related queries. In addition, fusing positive and negative concept detectors can enhance the search robustness for all types of queries. Performance comparison with the popular ontology reasoning techniques such as Resnik also confirms the merit of our approach in utilizing various aspects of detectors for video search.

Our works in this paper indeed provide interesting studies to some of the fundamental questions in concept-based video search, e.g., “How many detectors to select for a query?”, “How to utilize different properties of detectors for effective video search?”. In our approach, the answer to the first question depends on the number of detectors that could be found based on the subspaces of observability, which are formed by those anchor concept clusters related to a query. In our experiments on TRECVID datasets, on average there are approximately twenty detectors that can be attached to a query. Comparing to some existing approaches which utilize as few as one to three detectors [6, 27] or a predefined number of detectors [13], our approach is able to adaptively identify more detectors which sometimes appear novel for query answering. For the second question, we consider four aspects of detectors which have been empirically shown to be effective for video search. One aspect not included in this paper is the frequency appearance of concepts. Generally, frequent concept implies more training examples and thus the chance of developing a robust detector is higher. In future, we will take into account this aspect for possible extension of our fusion strategies proposed in this paper.

9. ACKNOWLEDGMENTS

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 118906).

10. REFERENCES

[1] A. K. Jain and R. C. Dube. *Algorithms for Clustering Data*. 1988.

[2] C. G. M. Snoek and et. al. The MediaMill TRECVID 2006 semantic video search engine. In *TRECVID*, pages 277–290, 2006.

[3] M. Campbell and et. al. IBM research TRECVID-2006 video retrieval system. In *TRECVID*, pages 175–182, 2006.

[4] N. Francis and H. Kucera. *Frequency analysis of English usage: Lexicon and grammar*. 1982.

[5] C. G. M. Snoek and et. al. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *ACM Conf. on Multimedia (MM)*, pages 421–430, 2006.

[6] C. G. M. Snoek and et. al. Adding semantics to detectors for video retrieval. *IEEE Trans. on Multimedia*, 9(5):975–986, 2007.

[7] A. Hauptmann and et. al. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *IEEE Trans. on Multimedia*, 9(5):958–966, 2007.

[8] R. A. Horn and C. R. Johnson. *Matrix Analysis*. 1985.

[9] J. J. Jiang and D.W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Intl. Conf. Research on Computational Linguistics*, 1997.

[10] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Intl. Conf. on Image and Video Retrieval (CIVR)*, 2007.

[11] J. R. Kender. A large scale concept ontology for news stories: Empirical methods, analysis, and improvements. In *Intl. Conf. on Multimedia and Expo (ICME)*, 2007.

[12] J. R. Kender and M. R. Naphade. Visual concepts for news story tracking: Analyzing and exploiting the NIST TRECVID video annotation experiment. In *Intl. Conf. on Computer Vision and Pattern Recognition*, 2005.

[13] L. S. Kennedy and S.-F. Chang. A reranking approach for context-based concept fusion in video indexing and retrieval. In *Intl. Conf. on Image and Video Retrieval (CIVR)*, 2007.

[14] M. Koskela and A. F. Smeaton. An empirical study of inter-concept similarities in multimedia ontologies. In *Intl. Conf. on Image and Video Retrieval (CIVR)*, 2007.

[15] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine code from an ice cream cone. In *Annual Intl. Conf. on Systems Documentation*, pages 24–26, 1986.

[16] X. Li and et. al. Video search in concept subspace: A text-like paradigm. In *Intl. Conf. on Image and Video Retrieval (CIVR)*, 2007.

[17] M. Naphade and et. al. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[18] A. P. Natsev, M. R. Naphade, and J. R. Smith. Semantic representation, search and mining of multimedia content. In *ACM Conf. on on Knowledge Discovery and Datamining (SIGKDD)*, pages 641–646, 2004.

[19] S.-Y. Neo and et. al. Video retrieval using high level features: Exploiting query matching and confidence-based weighting. In *Intl. Conf. on Image and Video Retrieval (CIVR)*, 2006.

[20] P. Over, W. Kraaij, and A. F. Smeaton. TRECVID 2007 - overview. In *TRECVID*, 2007.

[21] R. Penrose. A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society*, 51:406–413, 1955.

[22] P. Resnik. Using information content to evaluate semantic similarity in taxonomy. In *Intl. Joint Conf. on Artificial Intelligence (IJCAI)*, 1995.

[23] J. P. Romano. On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85(411):686–692, 1990.

[24] S. F. Chang and et. al. Columbia university TRECVID-2006 video search and high-level feature extraction. In *TRECVID*, pages 99–109, 2006.

[25] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *ACM Intl. Workshop on Multimedia Information Retrieval*, 2006.

[26] J. R. Smith, M. Naphade, and A. P. Natsev. Multimedia semantic indexing using model vectors. In *Intl. Conf. on Multimedia and Expo (ICME)*, 2003.

[27] X.-Y. Wei and C.-W. Ngo. Ontology-enriched semantic space for video search. In *ACM Conf. on Multimedia (MM)*, 2007.

[28] L. Xie and S.-F. Chang. Pattern mining in visual concept streams. In *Intl. Conf. on Multimedia and Expo (ICME)*, 2006.

[29] A. Yanagawa and et. al. Columbia university's baseline detectors for 374 LSCOM semantic visual concepts. Technical report, Columbia University, 2007.

[30] W. Zhibiao and M. Palmer. Verb semantic and lexical selection. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 133–138, 1994.