Selection of Concept Detectors for Video Search by Ontology-Enriched Semantic Spaces

Xiao-Yong Wei, Chong-Wah Ngo and Yu-Gang Jiang

***Abstract*—This paper describes the construction and utilization of two novel semantic spaces, namely Ontology-enriched Semantic Space(OSS) and Ontology-enriched Orthogonal Semantic Space (OS^2), to facilitate the selection of concept detectors for video search. These two semantic spaces are enriched with ontology knowledge, while emphasizing consistent and uniform comparison of ontological relatedness among concepts for query-to-concept mapping. OS^2, in addition to being a linear space like OSS, also guarantees orthogonality of the semantic space. Compared with other ontology reasoning measures, both spaces are capable of providing platforms that offer a global view of concept inter-relatedness, by allowing evaluation of concept similarity in metric spaces. We simulate OSS and OS^2 by using LSCOM concepts and experiment search effectiveness with VIREO-374 concept detectors. Empirical observations indicate that the proposed semantic spaces enable more effective selection of concept detectors than eight other existing ontology measures. OS^2, in particular, is better in providing a viable and reasonable solution for fusion of multiple concept detectors.**

***Index Terms* — Semantic space, ontology, concept-based video search, semantic detectors.**

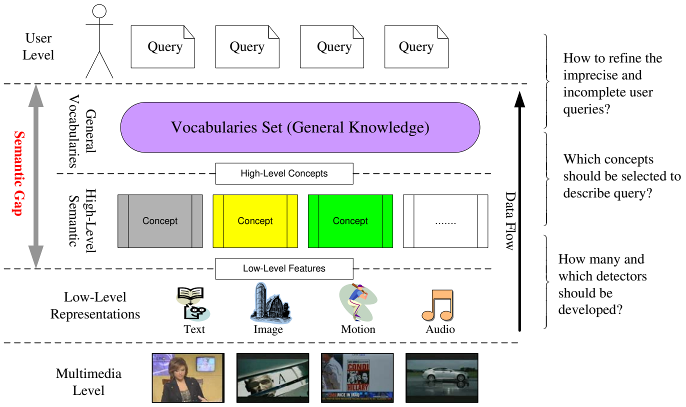

User Level
Query
Query
Query
Query
How to refine the imprecise and incomplete user queries?
Semantic Gap
General Vocabularies
Vocabularies Set (General Knowledge)
High-Level Concepts
High-Level Semantic
Concept
Concept
Concept
....
Data Flow
Which concepts should be selected to describe query?
Low-Level Features
How many and which detectors should be developed?
Low-Level Representations
Text
Image
Motion
Audio
Multimedia Level

Fig. 1. General framework of concept-based video search. The semantic gap between low-level features and user queries is bridged by a set of concept detectors enriched by general knowledge such as ontology.

I. Introduction

Semantic-based retrieval has been one of the long-term goals of multimedia computing. Traditional content-based approaches for deriving semantics, purely based on low-level features, such as color and texture, have shown their limitations in conquering the so-called "semantic gap". Modern approaches enable a semantic search by pooling a set of concept detectors (e.g., *car* and *building*) to extract semantics from low-level features, and thus forming a semantic space to facilitate high-level understanding of user queries [1]–[5]. Such search methodology is usually referred to as concept-based video search, as illustrated in Figure 1. The semantic gap from user queries to raw data is bridged with a pool of concepts enriched with general-purpose vocabularies, for instance, from ontology (e.g., WordNet) and external information (e.g., Internet). The ontology specifies the relationship among concept entities. Basically, a set of concept detectors is developed to represent high-level semantics. The detectors are classifiers learnt with training examples described by multi-modal features. Given a user query, the best set of concepts to describe the semantics of the query is reasoned through the vocabularies. A search list is then produced by ranking items (e.g., shots) according to their signal responses to the selected concept detectors.

Under the concept-based retrieval framework as depicted in Figure 1, an apparent issue is that, given a concept detector set, mapping ambiguity between queries and concepts needs to be carefully resolved. Consider, for instance, a query of "Find shots with snow", and a concept set with three detectors: *landscape*, *soccer*, *fire*. The concept similarities between {snow, *landscape*}, {snow, *soccer*} and {snow, *fire*} need to be properly reasoned in order to assign the best possible detectors with appropriate weights to answer the query. A common solution is to consider mapping through ontology reasoning [1], [3], [4], [6], [7], or more precisely selecting concepts, which minimize linguistic distance between the concepts and query terms. The mapping is normally done with a shared knowledge ontology such as WordNet [8], which is organized as a graph with nodes that represent concepts and edges that specify the relationships. Ontology reasoning normally involves only a local view of a subgraph structure where the two concepts under investigation reside. A fundamental question is: can the pairwise concept similarities measured based on the local view be effectively compared for selecting detectors? Such a reasoning technique does not allow uniform comparison of concept pairs, since the locally determined similarities, in principle, are not comparable from one concept pair to another.

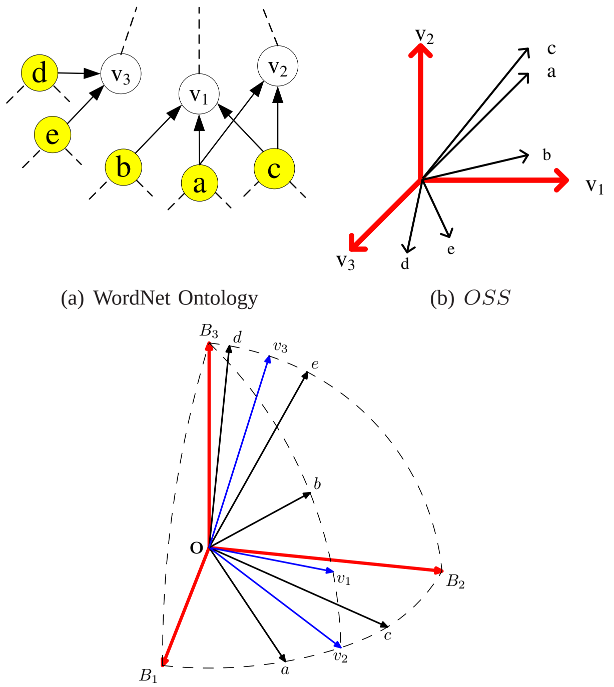
In this paper, we propose a novel construction of semantic space to measure concept similarity globally. In contrast to the conventional ontology reasoning, this space enables an *uniform* and *global* similarity measure of concepts. In this space, basis vectors are formed by modeling ontological relationship among concepts. Each concept is represented as a vector for similarity measurement purposes. Because ontology knowledge is taken into account when building the semantic

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).

space, we call the space “ontology enriched”. We propose two variants of the semantic space by considering orthogonality property of the space. The first space is named Ontology-enriched Semantic Space (OSS), originally presented in our recent work [9]. The second space is called Ontology-enriched Orthogonal Semantic Space (OS^2). With reference to figures 2(b) and 2(c), OSS is a linear space spanned with bases formed by a set of selected concept vectors. OS^2 is similar to OSS, but the bases are not formed by the concepts themselves. Instead, the basis vectors are computed by spectral decomposition in order to guarantee the orthogonality of the semantic space.

Figure 2 illustrates the major ideas of reasoning concept similarity in WordNet ontology, OSS and OS^2. Let concepts a to e as children and v_1 to v_3 as ancestors. In Figure 2(a), using the conventional ontology measures such as Resnik [10], the concept pairs (a, b) and (a, c) could be the same, although (a, c) shares another ancestor v_2 and intuitively should be more alike. On the other hand, the similarity scores of (d, e) and (a, b) cannot be reasonably compared as they reside in different parts of the ontology which carry different statistic and structural information. In brief, the reasoning is determined *locally* without a global ontological view. The uniform comparison of concept similarity scores cannot be conducted. OSS and OS^2, in contrast, project each concept as a vector in their semantic spaces for *global* and *uniform* concept similarity measures. In Figure 2(b), for instance, OSS is formed by selecting the ancestors v_1 to v_3 as the basis vectors. The concepts a-e are then linearly projected to OSS as vectors for concept similarity measure. OS^2, as shown in Figure 2(c), emphasizes space orthogonality and computes basis vectors (B_1, B_2, B_3) by spectral decomposition. With the bases, the semantic spaces in figures 2(b) and 2(c) guarantee consistency in comparing the concept pairs (a, b), (a, c) and (d, e), by keeping a global view of the concept relatedness to the basis vectors. Comparing both semantic spaces, OS^2, being an orthogonal space, has higher expressive ability because redundancy among the basis vectors is kept at a minimum. The orthogonality property could effectively prevent employment of basis vectors which might be correlated and ultimately results in certain subspaces dominating the whole semantic space. This property is important such that each basis has an equal contribution to the measurement of concept similarity.

The remaining sections are organized as follows. Section II briefly describes the works in current literature, in particular the ontology-based video search, related to our proposed works. Section III presents the construction of OSS and OS^2, and their properties. Section IV exploits the proposed semantic spaces for concept selection and fusion. Finally, sections V and VI present experiments on the construction of semantic spaces and the utilization of spaces for video retrieval respectively. Finally Section VII concludes this paper.



(a) WordNet Ontology

(b) OSS

(c) OS^2

Fig. 2. Reasoning concept similarity in (a) WordNet ontology: reasoning is conducted in a subgraph without a global view of the graph structure, (b) OSS: selected concepts (v_1, v_2, v_3) are represented as bases for vector-based concept similarity measure, and (c) OS^2: the bases (B_1, B_2, B_3) are computed by spectral decomposition to represent concept vectors.

II. Related Work

In the past few years, concept-based video retrieval has attracted numerous research attention. Two critical efforts are detection of semantic concepts and utilization of concepts as “semantic filters” for query answering. Since 2001, TRECVID (TREC Video Retrieval Evaluation) [11] sponsored by NIST has organized annual workshops to publicly release benchmarks and evaluations to support these efforts. Two tasks organized by TRECVID are high-level feature extraction (HLFE) and automatic video search. In HLFE, concept detectors are developed for video semantic annotation. In order to identify a right set of detectors to develop, collaborative efforts from various research organizations have been pooled in to assess the utility, observability and flexibility of the concept detectors [12]. One typical example is the release of LSCOM (Large-Scale Concept Ontology for Multimedia, *http://www.lscom.org/*) [12] 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 [13] and VIREO-374 [14], 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 [15] which provides 101 concept detectors.

With the availability of various concept detector sets, the automatic video search in TRECVID is often straightforward to perform based on the concept-based retrieval framework depicted in Figure 1. Various studies [2]–[4], [16] have been reported regarding the usefulness of concepts for video search, compared to search with low-level features and text keywords. The completeness, accuracy and utility of LSCOM concepts towards effective search performance is also investigated in [17]. Recently, the fundamental question of how many detectors are enough for effective video search is studied in [18]. In this work, it is reported that fewer than 5000 concepts, detected with minimal accuracy of 10%, is likely to provide satisfactory retrieval performance.

In this section, we begin by briefly describing the existing concept similarity measures for ontology reasoning in Section II-A. The related works in ontology-based video search will be further presented in Section II-B. A brief comparison of anchor-based selection approaches to our proposed semantic spaces will also be discussed in Section II-C.

A. Ontology Reasoning

Ontology reasoning is an ongoing research topic of linguistic computing [19]. Different measures have been proposed to evaluate relatedness of two concepts by querying ontologies such as WordNet for relatedness reasoning. The relatedness is normally based on ontology distance which utilizes the hyponym (is-a relationship) of concepts. With WordNet as an example, the is-a relationship can be viewed as a graph with nodes representing concepts and edges representing the concept relatedness. The distance between two concepts is dependent on information content (IC) and specificity of concepts, or path length from one concept to the other by traversing the edges. The IC is inversely proportional to the probability of a concept being observed. The specificity of a concept is defined by the depth of the concept in the graph, where depth is ordered according to the levels of is-a relationship. For instance, the concept *car* is under its ancestor *vehicle* and thus resides deeper than *vehicle* in WordNet.

Popular measures for concept similarity includes Leacock and Chodorow (LCH) [20], Wu and Palmer (WUP) [21], Resnik (RES) [10], Lin (LIN) [22], Jiang and Conrath (JCN) [23], Lesk [24], Gloss Vector (Vect) [25] and Pairwise Gloss Vector (VP) [25]. LCH and WUP use path length information, while the remaining measures utilize information content (RES, LIN, JCN) and definition of word sense (Lesk, Vect, VP). Denote D as the depth and I as the IC of a concept, L as the path length between two concepts, and p_{ij} as the common ancestor of concepts c_i and c_j. Some of these measures are defined as

$$\begin{aligned} LCH(c_i,c_j) &= -\log\frac{L(c_i,c_j)}{2\delta} && (1)\\ WUP(c_i,c_j) &= \frac{2D(p_{ij})}{L(c_i,c_j)+2D(p_{ij}))} && (2)\\ RES(c_i,c_j) &= I(p_{ij}) && (3)\\ LIN(c_i,c_j) &= \frac{2I(p_{ij})}{I(c_i)+I(c_j)} && (4)\\ JCN(c_i,c_j) &= \frac{1}{I(c_i)+I(c_j)-2I(p_{ij})} && (5) \end{aligned}$$

where δ denotes the maximum depth of WordNet. The IC is estimated based on the one-million-word Brown Corpus of American English [26]. Lesk utilizes the number of shared words (overlaps) in the definitions (glosses) of concepts. Vect represents concepts as gloss vectors using the co-occurrence information derived from glosses. The cosine similarity between gloss vectors is used to measure the concept relatedness. VP is similar to Vect, but different in the way it augments the glosses of concepts with adjacent glosses [25].

B. Ontology-based Video Search

Depending on the modalities of search queries (visual and/or text), there exist various ways to perform mapping from queries to concepts. For text queries, the approaches in [3], [4], [16] conduct mapping by the concept similarity measures as presented in Section II-A. In addition to ontology reasoning, some approaches also explore the mapping by comparing queries against the text descriptions associated with concepts [4], or to expand queries with related terms [1], [3]. The expanded terms as well as their weights are learnt from training examples [1] or external information such as Internet [3]. For queries with image or video examples, the mapping is often done by selecting the concept detectors which output high confidence to query examples, indicating the likelihood of corresponding concepts presented in the queries. When multiple detectors are selected, the weight of a detector is normally assigned based on the detection score of the detector to image/video examples [27], or the ontology similarity of the concept to text query [3].

A different strategy of query-to-concept mapping is via construction of semantic space or vector space for modeling concepts. The pioneering work in [5], [28] constructs a semantic space, or more precisely a vector space, formed by a set of available concept detectors. In this space, a retrieval item (e.g., shot) is represented as a vector of model scores. The scores are computed based on the signal responses of the detectors to the item. Contrasting to other approaches based on ontology reasoning [4], [16], no specific detector is selected, but rather all detectors are involved in the video search though each detector carries different weights. In [27], 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.

Conducting search based on ontology construction has also been previously studied in [4], [29]–[31]. The construction mostly involves manual mapping of visual elements to textual concept entities provided by shared vocabularies. In [29], WordNet is extended with visual tags describing properties such as visibility, motion and frequency of occurrence. In [31], based on WordNet and MPEG-7, a visual ontology is created by linking visual and general concepts. In view of the richness of human vocabularies and the need for domain experts in tagging or creating links, the scalability of these approaches still remains unclear. A relatively straightforward approach is recently proposed in [4] by directly attaching concept detectors to WordNet synsets. The semantically enriched detectors can thus utilize contextual information provided by WordNet. In addition to the ontologies built on the basis of general-purpose vocabularies, domain specific multimedia ontology is also investigated. For instance, in [32], two animal domain ontologies are constructed respectively for textual and visual descriptions for semantic search.

C. Anchor-based Selection

Considering the way that the proposed semantic spaces are built by selecting and constructing the bases, our work is also related to the anchor-based selection approaches [33]–[36]. In these approaches, anchor space is built by selecting a subset of objects from database as global reference axes. The selected objects are named as foci [33], anchor [34] or vantage points [35]. The main challenges of anchor-based selection are which and how many objects should be selected as anchors. The recent work in [33], for instance, proposes HF (Hull-of-Foci) algorithm for the selection of anchor objects. The idea of using anchors to build anchor-space has actually been used in various applications including database indexing [33], music classification [34], image retrieval [35], and animal sound classification [36].

III. Modeling Semantic Space with Ontology

This section presents the construction of ontology-enriched semantic space. Given a vocabulary set $\mathsf{V} = [c_1, c_2, \ldots, c_n]$ of n concepts, we want to represent each concept c_i in a vector form in the semantic space. Denote $\mathbf{C}$ as the n-by-n concept matrix which captures the vectors, defined as

$$\mathbf{C} = [\vec{c}_1, \vec{c}_2, \ldots, \vec{c}_i, \ldots, \vec{c}_n] \tag{6}$$

where $\vec{c}_i$ is a n-dimensional vector representing concept c_i. With $\mathbf{C}$, the semantic space can be estimated as

$$\vec{c}_1 \times \vec{c}_2 \times \ldots \times \vec{c}_n \longrightarrow \mathbb{R} \tag{7}$$

where ideally the n concept vectors together form the bases that approximate the real world space $\mathbb{R}$. To estimate the semantic space, there exist two major issues: the estimation of the concept matrix $\mathbf{C}$, and the orthogonality and compactness of the semantic space.

A. Constructing OSS

OSS, originally proposed in [9], aims to make the semantic space in Eqn (7) as compact and complete as possible. Similar to the anchor selection approaches, OSS achieves the aim by identifying a subset of concepts in $\mathbf{C}$ appropriate to serve as the basis vectors of the semantic space. To estimate $\mathbf{C}$, OSS computes the ontological relationships of n concepts by measuring their pairwise similarity. The WUP measure in Eqn (2) is employed to compute the similarity of each concept pair. This forms the matrix $\mathbf{R} = [r_{ij}]_{n\times n}$ where each component r_{ij} represents the similarity of a concept pair (c_i, c_j). $\mathbf{R}$ basically approximates $\mathbf{C}$ and encapsulates the all-pair WUP similarities of n concepts. Each column vector r_i in $\mathbf{R}$ outlines the similarities of the concept c_i to other concepts.

To minimize the redundancy among concepts, OSS adopts clustering approach which groups the n concept vectors in $\mathbf{R}$ into $m < n$ clusters, and then selects one medoid from each cluster to form the set of basis vectors. The OSS is thus spanned by m medoid concepts. With m bases, the matrix $\mathbf{R}$ is reduced to $\hat{\mathbf{R}}$ of m-by-m size. In OSS, each concept can be easily represented as a vector in m dimensions, by measuring the WUP similarity of the concept to m medoids. An advantage of OSS is that the bases are interpretable with each basis represented by a semantic concept. Nevertheless, the space is not strictly orthogonal, and the basis vectors are thus somewhat correlated.

B. Constructing OS^2

OS^2 aims to construct an orthogonal semantic space to depict Eqn (7). Similar to OSS, OS^2 first assumes the concept matrix $\mathbf{C}$ can be modeled with matrix $\mathbf{R}$ computed with WUP measure. Further assuming that each concept vector r_i in $\mathbf{R}$ is normalized, we can have

$$\mathbf{C}^T\mathbf{C} = \mathbf{R} \tag{8}$$

To solve $\mathbf{C}$, spectral decomposition [37] is applied to $\mathbf{R}$:

$$\begin{aligned} \mathbf{R} &= V\Lambda V^T \\ &= (V\Lambda^{\frac{1}{2}}V^T)^T(V\Lambda^{\frac{1}{2}}V^T) \end{aligned} \tag{9}$$

where Λ is a matrix with all the eigenvalues of $\mathbf{R}$ on its diagonal, and V is the corresponding eigenvector matrix. As a consequence, a particular solution that can describe $\mathbf{C}$ is

$$\mathbf{C} = V\Lambda^{\frac{1}{2}}V^T \tag{10}$$

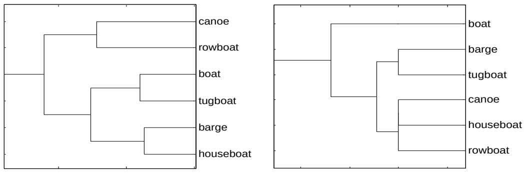
With the spectral decomposition, the semantic space formed by OS^2 is orthogonal and spanned with the eigenvectors which are computed by Schur decomposition [37] in our approach. The concept vectors in $\mathbf{C}$ are obtained directly via the transformation in Eqn (10). Comparing to OSS, the axes of OS^2 are not represented by the original concept vectors. Instead, each basis vector is the linear combination of concept vectors and orthogonal to each other.

1) Representing Unseen Concept: Because each basis vector is not directly interpretable, representing the unseen concepts not found in the vocabulary set V is not as straightforward as OSS. Given a concept $u \notin \mathsf{V}$, the corresponding concept vector $\vec{u}$ is predicted as

$$\begin{aligned} \mathbf{C}^T\vec{u} &= \mathbf{R}_u \\ \vec{u} &= (\mathbf{C}^T)^{-1}\mathbf{R}_u \end{aligned} \tag{11}$$

where $\mathbf{R}_u$ is a n-dimensional vector, representing the ontological relatedness of u to the n concepts in V with WUP similarity. Note that the matrix $\mathbf{C}^T$ might be singular as the concepts in V are not completely independent (e.g., concept *car* vs. *vehicle*), causing the $\vec{u}$ not having a unique solution. In our case, we solve the inversion of $\mathbf{C}^T$ with generalized inverse [38]. The Moore-Penrose pseudoinverse is adopted which can always make sure that Eqn (11) is solvable.

2) Minimizing Concept Redundancy and Re-estimate OS^2: While the solution of OS^2 with spectral decomposition and generalized inverse sounds feasible, theoretically the solution could be accurately estimated only when the initial given set of vocabulary V contains no redundant concept. By checking the concept sets such as the ones provided in LSCOM and MediaMill-101, there always exist redundant concepts, which make spectral decomposition unstable. Moreover, the approximation error will be further amplified by Moore-Penrose



(a) (b)

Fig. 3. Partial view of LSCOM dendrograms created by OS^2 with: (a) the original vocabulary set V; (b) a compact vocabulary set $\hat{\mathsf{V}}$. When concept redundancy is minimized in $\hat{\mathsf{V}}$, the general concept *boat* resides at a higher abstract level of the hierarchy in (b) than (a).

pseudoinverse, causing the prediction of unseen concepts imprecise.

To tackle this problem, OS^2 also adopts the clustering approach, prior to spectral decomposition, to group concepts while finding the optimal number of clusters in V. The aim is to reduce concept redundancy and then use a more compact concept set $\hat{\mathsf{V}}$ to estimate the semantic space. This process is similar to OSS where the medoid of each cluster is picked to formed the new set $\hat{\mathsf{V}}$ of $m < n$ concepts. A reduced matrix $\hat{\mathbf{R}}$ of m-by-m size is computed, and then decomposed via Eqn (9). The coordinate system of OS^2 is then estimated and represented with the eigenvectors of Eqn (9). The semantic space is spanned by $m < n$ basis concepts, and thus is compact and relatively efficient when predicting the unseen concepts using Eqn (11). Figure 3(b) illustrates the advantage of estimation with $\hat{\mathsf{V}}$ by showing a partial view of the LSCOM dendrogram created by OS^2 with $\hat{\mathsf{V}}$. Compared to the original dendrogram created with V in 3(a), the new dendrogram is more intuitive. For instance, the concept *boat* is correctly merged at a higher abstract level of the dendrogram in 3(b) than in 3(a).

C. Properties of OSS and OS^2

1) Metric Space: Since the spaces formed by OSS and OS^2 are linear, many known metrics can be employed to characterize distance. In this paper, we use cosine similarity for measuring the relatedness of concept vectors. Given two concepts u and v, the cosine similarity between them is

$$Sim(u, v) = \frac{\vec{u} \cdot \vec{v}}{|\vec{u}||\vec{v}|} \tag{12}$$

Note that the concept similarity is not only based on the ontology relationship between concepts u and v, but is also with respect to their relatedness to the medoid concepts obtained through clustering. Compared to OSS, OS^2 has the extra advantage that concept vectors are uniformly measured in the orthogonal space.

Compared to other ontology measures such as Resnik [10] and WUP [21], both OSS and OS^2 are metric spaces that allow the consistent comparison of concept similarities. It is not hard to show that other measures violate metric properties. Take the graph structure in Figure 4 as an example, the path length of $(b, a) + (a, c) \leq (b, c)$ violates triangle inequality. Similarly, suppose each node is attached with information content (IC), then $IC(e)+IC(d) \geq IC(f)$. Since IC is used as a similarity measure and inversely proportional to distance, IC based approach is also not a metric.

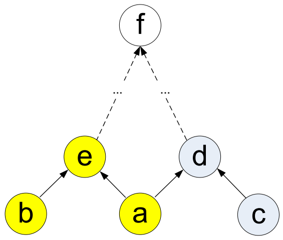
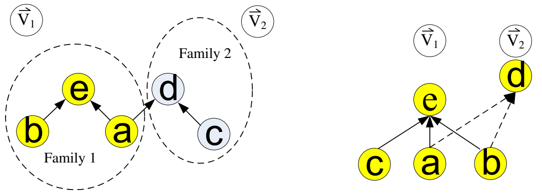


Fig. 4. Path length based ontology measures are not metric. For instance, the distance from concept b to c is equal to the length of path $b \rightarrow$e$\rightarrow \ldots \rightarrow f \rightarrow \ldots \rightarrow d \rightarrow c$. Obviously $(b, c) \geq (b, a) + (a, c)$.



(a) (b)

Fig. 5. Measuring the concept similarity in WordNet with WUP. (a) The similarity of (a, b) is the same as (a, c), although a and b reside in a branch (Family-1) different from c (Family-2), and thus should have higher similarity. (b) The concept pairs (a, b), (a, c), (c, b) have the same WUP similarity, although a and b have another common ancestor d in addition to e, and thus should be more similar.

2) Comparison to WordNet: To fully reveal the benefit of OSS and OS^2, we contrast the major difference of measuring concept similarity in these two spaces and in the original ontology space (WordNet). Figure 5 illustrates two typical cases where the linguistic-based similarity measures such as WUP fail in distinguishing the relatedness between concepts. For ease of elaboration, we assume concepts a, b and c resides at the same level of depth, and concepts e and d are the ancestors. In Figure 5(a), the concept a shares the same WUP similarity with both b and c, although c resides in a family different from a and b. With OS^2 (or OSS), suppose v_1 and v_2 are the medoid concepts in the set $\hat{\mathsf{V}}$ (or basis vectors in OSS), where $\vec{v}_1$ is more related to Family-1 while $\vec{v}_2$ is more related to Family-2. By Eqn (12), we can easily show that $Sim(a, b) > Sim(a, c)$. This is simply because the concept vectors $\vec{a}$, $\vec{b}$ and $\vec{c}$ are compared on a space that accounts the inter-concept relatedness. Similarly in Figure 5(b), the concepts pairs (a, b), (a, c) and (c, b) all have the same WUP similarity, although a and b are more related because of sharing another common ancestor. Assuming the concept d is close to $\vec{v}_2$ while concept e is close to $\vec{v}_1$, we can easily prove that $Sim(a, b) > Sim(a, c)$ in OSS and OS^2.

In brief, the concept similarity in either OSS or OS^2 is *globally* measured with the constructed semantic axes. While in WordNet, most linguistic reasoning methods utilize the local structure (depth, path length, specificity) peculiar to a sub-graph for measuring similarity. Consequently, an uniform and objective comparison of similarity scores obtained from different sub-graphs of WordNet becomes difficult.

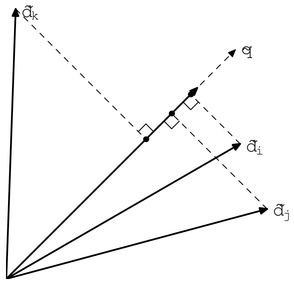


Fig. 6. Assigning weights to the top-3 similar detectors $\{d_i, d_j, d_k\}$ of the query term q. The weight of a detector is equal to the cosine similarity of the detector to query term in the semantic space.

IV. Concept-based Video Search by OSS and OS^2

Given a text query $\mathsf{Q} = \{q_1, \ldots, q_m\}$ of m terms and the detector set $\mathsf{D} = \{d_1, d_2, \ldots, d_n\}$ of n concepts, we measure the pairwise term-to-concept similarity. The detectors are then ranked according to their similarities to the query terms. The top-k most related detectors to the query are subsequently selected for concept-based video search. Because whether in OSS or OS^2 the detectors and query terms are represented as vectors (see Figure 6), we can adopt the same strategies for concept selection and fusion in both spaces.

A. Concept Selection and Fusion

The similarity between a query term q and a detector d_j is computed via the cosine similarity in Eqn (12). The top-1 detector is straightforwardly selected as

$$\hat{d} = \operatorname*{argmax}_{d_j \in \mathsf{D}} Sim(q, d_j) \ \forall q \in \mathsf{Q} \tag{13}$$

The selection of top-k detectors are conducted in a similar way as Eqn (13), by picking up the k^{th} most related detector one at a time. By having k detectors, a fundamental issue is the fusion of detectors, specifically how to assign weights to different detectors, for retrieval. Because OSS and OS^2 are linear spaces, the weights can be determined by simply equaling their values to the similarities of terms and concepts. With Figure 6 as an example, supposing the top-3 selected concept detectors are $\{d_i, d_j, d_k\}$, and the most similar query term to these detectors is q. The weight assigned to a detector is equal to its cosine similarity to q. In other words, the weight is inversely proportional to the angle between query vector and concept vector. The smaller the angle, the larger the weight. In Figure 6, the detector d_i is assigned the highest weight, followed by d_j and d_k.

Let I be a retrieval item (e.g., shot) and T be the set of top-k detectors. The similarity of a query Q to I is determined by the weighted linear fusion of detectors as follows

$$Sim(\mathsf{Q}, I) = \sum_{d_i \in \mathsf{T}} Sim(q_i, d_i) \times Score(d_i, I) \tag{14}$$

where q_i is the most similar query term to d_i, $Sim(q_i, d_i)$ is the weight assigned to the detector d_i, and $Score(d_i, I)$ is the output score of d_i when detecting the corresponding concept on item I.

B. Word Sense Disambiguation (WSD)

A query term is normally associated with multiple senses or meanings. The exact sense of a term can be inferred by knowing the contextual relationship of neighboring terms in a query. For example, *map* has two senses in WordNet: *graphic map* or *mapping function*. Given the query "*map of Iraq*", *map* is assigned to the former sense by knowing *Iraq* is a country. Word sense disambiguation (WSD) is a query preprocessing technique commonly used for inferring the word sense and predicting the search intention of queries which are short and imprecise [24].

We formulate WSD as a greedy search approach which can be implemented directly in OSS and OS^2. The approach estimates the actual sense of a term q_i jointly with other senses of terms in the query Q. Suppose each term has p senses, there are m^p ways of interpreting Q. Greedy search is adopted to find a combination that maximizes the overlap of senses for all terms in Q. With OS^2 as example, the approach is implemented by representing each sense with Eqn (11) and then measuring the similarity of senses via Eqn (12). Denote s_i^k as the sense of q_i in k^{th} combination, the actual query sense $\hat{\mathsf{Q}} = \{\hat{s}_1, \ldots, \hat{s}_m\}$ is computed as

$$\hat{\mathsf{Q}} = \operatorname*{argmax}_{1 \le k \le m^p} \phi(k) \tag{15}$$

where

$$\phi(k) = \sum_{i=1}^{m} \sum_{j=i+1}^{m} Sim(s_i^k, s_j^k) \tag{16}$$

The query $\hat{\mathsf{Q}}$, which associates the predicted sense of each term, is then used for concept selection and fusion as presented in Section IV-A.

V. Experiment-I: Construction of Semantic Space

The aim of this section is to experiment with the construction of OSS and OS^2 for effective video search. In particular, the selection and computation of basis vectors are evaluated. Comparison to the anchor-based selection approach in [33] is also given to verify the effectiveness of the clustering algorithm adopted in OSS and OS^2.

In Section V-A, we use LSCOM concepts as the vocabulary set for the construction of semantic space. The test set of TRECVID 2006 video dataset [11] is further used to verify the search effectiveness in sections V-B and V-C. The video archive consists of about 150 hours (79,484 reference shots) of broadcast videos collecting from multi-lingual sources including English, Chinese and Arabic languages. Twenty-four search topics (see Table III), together with their ground-truth provided by TRECVID 2006, are used as queries. We only use the text queries of search topics for experiments, imagining that most searchers use to perform search with a short description of words.

For semantic concepts, we use VIREO-374 concept detectors [14] trained using TRECVID 2005 development set. 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 detector set of 244 concepts. In the experiments, we test the selection of single and multiple concepts per search topic respectively. The retrieved items (shots) are ranked according to their score to the selected concept detector(s). The search performance is then evaluated with mean average precision (MAP), where AP is defined as

$$AP = \frac{1}{\min(R,k)} \sum_{j=1}^{k} \frac{R_j}{j} I_j \tag{17}$$

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. We set k=1000, following the standard of search task in TRECVID. MAP is the mean AP over all search topics.

A. *Constructing OSS and OS^2*

We adopt agglomerative hierarchical clustering algorithm [39] to find the best set of concepts to construct OSS and OS^2. The initial vocabulary set V is formed by the concepts from LSCOM. We select 572 concepts from LSCOM and include them in V, by discarding those concepts not defined in WordNet or being synonym of the existing concepts. The actual senses of the selected concepts in WordNet are then manually assigned based on visual impression. For ease of evaluation, we only assign one sense to each concept, although multiple sense assignment is possible.

By the agglomerative hierarchical clustering algorithm [39], a dendrogram of 572 concepts is formed. We employ the inconsistency coefficient [39] to find the best possible concept clusters in the dendrogram. Denote l as a link connecting two clusters, the inconsistency coefficient $\tau(l)$ of the link is computed as

$$\tau(l) = \frac{len(l) - \mu(l)}{\sigma(l)} \tag{18}$$

where $len(l)$ is the length of link l, defined as the centroid distance between two clusters connected by l. The $\mu(l)$ and $\sigma(l)$ specify the average length and standard deviation of all links under l respectively. The coefficient $\tau(l)$ basically characterizes the tightness of a grouping under the link l, by comparing its length with all links under this grouping. The lower the value of τ, the more similar the concepts under the link. At the lowest level of dendrogram, $\tau(l) = 0$ since only two concepts are under l.

Figure 7 shows the number of clusters (y-axis) whose links are below a given coefficient value (x-axis). The result indicates that the best possible case happens when there are 366 concept clusters, where the $\tau(l)$ increases slightly from 0 but with a dramatic jump of 572 to 366 concepts. Table I shows few examples of the 366 concept clusters and their medoids. The medoid concepts are selected as the basis vectors of OSS, while for OS^2 the medoids form the reduced vocabulary set $\hat{\text{V}}$ for spectral decomposition in Eqn (9). The next subsection will further investigate the impact of concept clusters to the search performance.

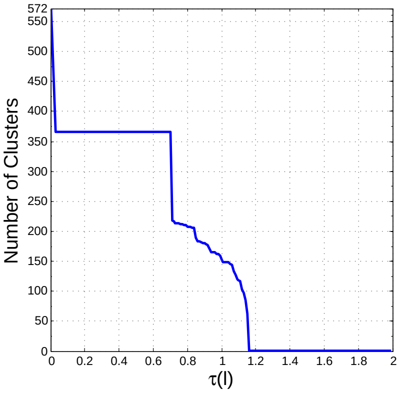


Fig. 7. Obtaining different number of concept clusters by thresholding the inconsistency coefficient $\tau(l)$ at different values. The best number of clusters happens when the curve remains steady at point where there are 366 clusters.

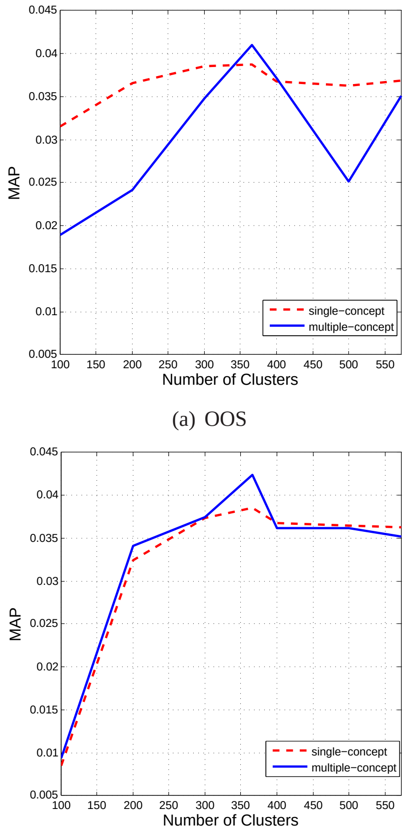
TABLE I
EXAMPLES OF THE SELECTED MEDOID CONCEPTS BY AGGLOMERATIVE HIERARCHICAL CLUSTERING AND INCONSISTENCY COEFFICIENT.

Medoid	Cluster Members
building	greenhouse, office building, music hall, theater, bathhouse, barn, boathouse, pumping station, farm building, observatory, hotel, bridge, bridge, viaduct, overpass
vehicle	tractor, armored vehicle, car, motorcycle, pickup truck, limousine, truck
leader	cheer leader, tribal chief
store	supermarket, butcher shop, shopping mall, retail store
boat	sailboat, rowboat, houseboat, tugboat, canoe, barge boat
room	cabin, conference room, ballroom, room, classroom, art gallery lobby, bar, dining room, kitchen, emergency room, library, bathroom, living room
battleship	frigate, aircraft carrier, submarine
sport	tennis, basketball, ice skating, swimming, car racing, skiing, gymnastics

B. *Impact on Video Search Performance*

To verify that the selection of 366 concept clusters is the best possible choice for OSS and OS^2, we compare the search performance in terms of MAP by varying the number of concept clusters. Figures 8(a) and 8(b) show the MAP of 24 TRECVID search topics against different choices from 2 to 572 concept clusters[1] for OSS and OS^2 respectively. We experiment both single and multiple concept selection. For multiple concepts, the top-3 detectors are selected for query answering. As indicated in the figures, the search performance basically improves when more medoids are included to learn the semantic space. The MAP reaches the highest when the number of axes is equal to 370 and 361 for OSS and OS^2 respectively. The performance starts to drop from this point onwards when more medoids are considered. The results of peak at 370 and 361 are slightly deviated from the ideal theoretical peak of having 366 medoid concepts. The results are not surprised since the performance to certain extent is also dependent on the reliability of detectors. The empirical evidence, overall, shows that 366 clusters, or slightly deviated from this number, are enough to represent the 572 concepts in LSCOM.

[1] In Figure 8, the step size of the number of concept clusters is set to 25. The step size is further refined to 1 in the range of 300 and 400 in order to find the best search performance.



(a) OOS

(b) OS^2

Fig. 8. Experimenting search performance when different numbers of concept clusters are used for constructing semantic space. The best performances of OSS and OS^2 are achieved when there are 370 clusters.

The performance of OSS could be explained by the space completeness. Underestimating the number of axes results in the lack of bases to span the semantic space. The incompleteness causes the deficiency of vector representation in the space. Overestimating the number of basis vectors, on the other hand, could results in over emphasis of certain concepts which are correlated. Similarly for OS^2, the space should be completed with abundant concepts. Furthermore, concept redundancy ought to be minimized. Otherwise, the concept vectors might not be properly predicted with Eqn (11). From the results in Figure 8, the performances of OSS and OS^2 are similar and attain the best, for single and multiple concept selections, when there are abundant concepts with less correlation selected to build the semantic space.

Considering when there is lack of concepts to span the semantic space, OSS shows better MAP for single concept selection, as indicated in Figure 8 before reaching the peak performance. This difference is mainly due to their fundamental considerations: OSS is built purposely with the real concepts as their bases, but this is not the case for OS^2. Due to the requirement of space transformation, OS^2 could suffer from computational instability if there are no abundant concepts available in the vocabulary set $\hat{\mathrm{V}}$. On the other hand, OS^2 has higher ability in offering consistent query-concept similarity for the fusion of multiple concept detectors. Particularly, if the number of basis vectors is over-estimated, OS^2 is able to show more stable and better performance than OSS. This is mainly due to the advantage of having orthogonal bases where the redundancy is tackled during the stage of space transformation. OSS, without taking into account the space orthogonality, deteriorates considerably when redundant concept clusters are included, as indicated in Figure 8(a).

TABLE II
COMPARING HIERARCHICAL CLUSTERING AND HF ALGORITHM IN CONSTRUCTING THE SEMANTIC SPACE FOR VIDEO SEARCH.

	Basis	Single concept		Multiple concept	
	vectors	OSS	OS^2	OSS	OS^2
Clustering	366	0.0384	0.0385	0.0410	0.0424
Hull of Foci (HF)	347	0.0316	0.0341	0.0351	0.0374

C. Comparison to Anchor-based Selection Algorithm

In constructing OSS and OS^2, there are various ways of selecting concepts which ultimately form the semantic spaces. In this subsection, we verify the choice of adopting hierarchical clustering in OSS and OS^2, in comparison with the "Hull of Foci" (HF) algorithm recently proposed in [33]. HF is basically a greedy search algorithm to select a number of anchors from a given dataset as global reference points. In OSS, for instance, the anchors can be directly treated as the basis vectors of the semantic space. In [33], the number of anchors is estimated by approximating the intrinsic dimension of a dataset. In our implementation, we employ the algorithm in [40] to approximate the intrinsic dimension.

Table II lists the search performance of employing hierarchical clustering and HF algorithm for constructing OSS and OS^2. For HF, there are 347 anchors being selected to form the basis vectors. As shown in Table II, for both single and multiple concept selections, the hierarchical clustering outperforms HF algorithm. We investigate the results and find that hierarchical clustering indeed has a better capability in removing concept redundancy in our application. For instance, the concepts *military personnel* and *military* are both selected as anchors by HF but not by hierarchical clustering.

VI. EXPERIMENT-II: CONCEPT-BASED VIDEO SEARCH

In this section, we study the search performance of OSS and OS^2 from three different aspects: effectiveness of concept fusion, influence of detector set, and comparison to eight other ontology reasoning measures used in the literature. OSS and OS^2 are constructed based on the results presented in Section V-A, where there are 366 concept clusters being selected to build the semantic spaces.

In order to have more sample queries for experiments, we use the testing sets of TRECVID 2005 and 2006 in this section. This results in a total of 48 testing queries. The topics, ranging from ID 149 to 196, are listed in Table III. The topic-ID is named and assigned by TRECVID. Note that the topics 149-172 are conducted on TRECVID 2005 dataset, while the topics 173-196 are conducted on TRECVID 2006 dataset. There are 85 (150) hours of videos and 45,765 (79,484) reference shots in the testing set of TRECVID 2005 (2006) dataset.

TABLE III

SEARCH TOPICS (TOPICS 149 TO 172 ARE FROM TRECVID 2005; TOPICS 173-196 ARE FROM TRECVID 2006).

ID	Topic
149	Condoleeza Rice
150	Iyad Allawi, the former prime minister of Iraq
151	Omar Karami, the former prime minister of Lebannon
152	Hu Jintao, president of China
153	Tony Blair
154	Mahmoud Abbas, also known as Abu Mazen, prime minister of the Palestinian Authority
155	A graphic map of Iraq, location of Bagdhad marked - not a weather map
156	Tennis players on the court, both players visible at the same time
157	People shaking hands
158	A helicopter in flight
159	George Bush entering or leaving a vehicle, he and vehicle both visible at the same time
160	Something on fire with flames and smoke visible
161	People with banners or signs
162	One or more people entering or leaving a building
163	A meeting with a large table and more than two people
164	A ship or boat
165	Basketball players on the court
166	One or more palm trees
167	An airplane taking off
168	A road with one or more cars
169	One or more tanks or other military vehicles
170	Tall building
171	A goal being made in a soccer match
172	An office setting, i.e., one or more desks/tables and one or more computers and one or more people
173	One or more emergency vehicles in motion
174	One or more tall buildings
175	People leaving or entering a vehicle
176	Soldiers, police, or guards escorting a prisoner
177	Daytime demonstration or protest with at least part of one building visible
178	US Vice President Dick Cheney
179	Saddam Hussein with at least one other person's face at least partially visible
180	Multiple people in uniform and in formation
181	US President George W. Bush, Jr. walking
182	Soldiers or police with one or more weapons and military vehicles
183	One or more boats or ships
184	People seated at a computer with display visible
185	One or more people reading a newspaper
186	A natural scene
187	One or more helicopters in flight
188	Something burning with flames visible
189	A group including least four people dressed in suits, seated, and with at least one flag
190	At least one person and at least 10 books
191	At least one adult person and at least one child
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
194	Condoleeza Rice
195	One or more soccer goalposts
196	Scenes with snow

TABLE IV

COMPARISON OF CONCEPT FUSION STRATEGIES FOR VIDEO SEARCH USING VIREO-374 DETECTOR SET.

	OSS	OS^2	Borda voting	Detection reliability
TRECVID 2006	0.0410	0.0424	0.0266	0.0104
TRECVID 2005	0.1186	0.1235	0.0286	0.0205

TABLE V

SEARCH PERFORMANCE WITH COLUMBIA-374 DETECTOR SET.

Concept selection	TRECVID 2006		TRECVID 2005	
	Single	Multiple	Single	Multiple
OSS	0.0211	0.0205	0.0554	0.0773
OS^2	0.0248	0.0261	0.0563	0.0788

A. Comparison of Concept Fusion Strategies

To investigate the effectiveness of fusing multiple concepts in OSS and OS^2, two fusion strategies based on Borda voting and detection reliability are used as the baselines for performance comparison. In Borda voting, the rank positions of a shot retrieved by different detectors are summed as the score. In detection reliability, the reliability of a detector is used as the weight for fusion. There are various ways to determine the reliability of a detector. In our implementation, the weights of detectors are set equal to their APs estimated based on a subset of training data obtained from TRECVID 2005 development set. In the experiment, the top three detectors of a search topic are first selected and then the rank lists are produced respectively by four different fusion strategies. Table IV shows the comparison of various fusion strategies. OS^2 shows the best MAP for 48 search topics followed by OSS. Both fusion strategies show significantly better search performance than the baselines by Borda voting and detection reliability.

B. Influence of Detector Set

To study the influence of detectors towards the search performance, we conduct an experiment by using Columbia-374 [13], instead of VIREO-374, as the set of detectors. Columbia-374, trained using the same dataset as VIREO-374, use grid-based color moment and Gabor texture as features. Table V shows the MAP of using the Columbia-374 detector set for single and multiple concept selections. Obviously, the performance is significantly impacted by the choice of detector set. Comparing to the MAP of 0.0424 (TRECVID 2006) by VIREO-374 and of 0.0261 by Columbia-374 in the case of multiple concept selection with OS^2, the performance difference is indeed significant.

C. Comparison of Concept Similarity Measures

In this section, we compare OSS and OS^2 to eight other popular ontology measures: LCH [20], WUP [21], RES [10], LIN [22], JCN [23], Lesk [24], Gloss Vector (Vect) [25] and Pairwise Gloss Vector (VP) [25]. In the experiment, except OSS and OS^2, all measures employ Lesk algorithm [24] for word sense disambiguation. OS^2 estimates the actual senses of query terms in its own semantic space as presented in Section IV-B. For multiple concept selection, linear fusion as presented in Eqn (14) is employed. Depending on the ontology measure being used, the weight of a selected detector is set equal to its similarity to query.

Table VI shows the performance comparison of ten different measures. The search result indicates that OS^2 outperforms other measures, particularly for the multiple concept selection.

TABLE VII
AVERAGE PRECISION OF VARIOUS ONTOLOGY MEASURES FOR MULTIPLE CONCEPT SELECTION ON TRECVID 2005 AND 2006. THE BEST RESULTS ARE BOLD.

ID	OSS	OS^2	LCH	WUP	RES	LIN	JCN	Lesk	Vect	VP
149	0.0002	0.0003	0.0002	0.0002	0.0003	0.0002	9E-05	**0.0006**	8E-05	3E-05
150	0.0001	**0.0015**	0.0001	9E-05	**0.0015**	0.0001	0.0000	0.0000	8E-05	0.0000
151	0.0007	0.0008	**0.0012**	0.0008	8E-05	0.0011	0.0002	0.0007	0.0006	0.0006
152	0.0239	**0.028**	0.0239	0.0095	0.0079	0.0236	0.0077	0.0021	0.016	0.002
153	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	3E-05	**0.0003**
154	0.0000	1E-05	4E-05	1E-05	1E-05	4E-05	9E-05	4E-05	3E-05	**0.0001**
155	0.027	**0.0289**	0.0177	0.0135	0.0241	0.0261	0.0185	0.0203	0.0211	0.0166
156	0.5844	**0.6245**	0.2953	0.1907	0.1845	0.1763	0.5844	0.5844	0.1736	0.2404
157	**0.0056**	0.0054	0.0047	0.0053	0.0044	4E-06	7E-06	0.0003	2E-06	0.0009
158	0.1853	**0.1912**	0.1796	0.1664	0.0536	0.0000	0.0000	0.19	0.1868	0.1901
159	**0.0016**	**0.0016**	0.0008	0.0008	0.001	**0.0016**	0.0005	**0.0016**	0.0002	0.0007
160	**0.0271**	0.027	0.0224	0.0174	0.0197	0.0198	0.0226	0.0234	0.0262	0.0214
161	0.0942	**0.0976**	0.0155	0.0145	0.0115	0.0213	0.0216	0.0001	2E-05	0.0421
162	0.0014	**0.0017**	0.001	0.0017	0.0008	0.0005	0.0011	0.0011	0.0002	0.0009
163	**0.0481**	**0.0481**	0.0244	0.0247	0.0253	0.0322	0.0258	0.0464	0.0004	0.033
164	**0.1549**	**0.1549**	0.1518	0.1425	0.1368	0.1117	0.1528	0.127	0.058	0.1546
165	0.5221	**0.5343**	0.471	0.412	0.3944	0.4437	0.5343	0.5204	0.2675	0.4421
166	**0.0064**	**0.0064**	0.0046	0.0037	0.0057	0.0032	0.0059	0.0058	0.0053	0.0058
167	0.0393	**0.0397**	0.0368	0.0393	0.0304	0.0132	0.0178	0.0203	0.0214	0.0277
168	**0.1979**	**0.1979**	0.1479	0.1399	0.1263	0.1737	0.1438	0.1516	0.0703	0.1979
169	**0.0761**	**0.0761**	0.0761	0.0761	0.0747	0.0761	0.0761	0.0503	0.0761	0.0699
170	0.1122	0.1122	0.047	**0.1127**	0.0103	0.039	0.0493	0.0506	0.0303	0.0476
171	0.6086	**0.6583**	0.539	0.4954	0.0426	0.0000	0.0000	0.4701	0.3493	0.5309
172	0.1281	0.1281	**0.1337**	0.1204	0.0833	0.1121	0.0851	0.0407	0.0585	0.129
173	0.0066	0.0081	0.0081	0.0074	0.0097	0.0140	0.0074	**0.0140**	**0.0140**	0.0133
174	**0.0656**	0.0607	0.0649	**0.0655**	**0.0656**	0.0276	0.0282	0.0435	0.0289	0.0531
175	0.0042	0.0063	0.0051	0.0044	0.0042	0.0058	**0.0072**	0.0036	0.0000	0.0042
176	0.0009	0.0010	0.0010	0.0010	0.0010	0.0009	0.0010	**0.0011**	**0.0011**	0.0010
177	0.0011	0.0018	0.0021	0.0014	0.0011	0.0002	0.0003	0.0015	0.0004	**0.0042**
178	0.0081	0.0065	0.0076	0.008	0.0081	0.0066	0.0067	0.0003	0.0001	**0.0088**
179	0.0000	**4E-07**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
180	0.0003	0.0003	**0.0004**	0.0000	0.0000	6E-06	2E-05	0.0000	0.0000	0.0000
181	**0.0024**	0.0013	0.0016	0.0024	0.0024	0.0017	0.0011	0.0006	0.0006	0.0007
182	**0.0376**	**0.0376**	**0.0376**	**0.0376**	0.0265	0.0271	0.0297	0.0297	0.0376	0.0258
183	0.0269	**0.0285**	0.0283	0.0256	0.0252	0.0152	0.0278	0.0299	0.0223	0.0287
184	0.0026	0.0042	0.0035	0.0028	0.0026	0.0040	**0.0045**	0.0029	0.002	0.0042
185	0.0713	0.0755	0.0572	0.0276	0.0906	0.0752	**0.0986**	0.0970	0.0033	0.0854
186	0.0266	0.0241	0.0241	0.0078	0.0107	**0.0305**	0.0117	0.0241	0.0246	0.0227
187	0.0210	0.0267	0.0095	0.0158	0.0011	0.0044	0.0240	0.0258	0.0257	**0.0290**
188	3E-06	7E-06	8E-06	1E-05	3E-06	0.0122	0.0012	1E-05	**0.0114**	6E-05
189	0.0201	0.0256	0.0224	0.0207	0.0201	**0.0309**	0.0306	0.0292	0.0062	0.0292
190	0.0001	0.0003	**0.0005**	6E-06	4E-06	3E-06	0.0000	0.0000	0.0001	6E-06
191	2E-05	2E-05	2E-05	0.0025	2E-05	2E-05	**0.0025**	2E-05	**0.0025**	2E-06
192	0.0090	**0.0167**	0.0139	0.0126	0.0137	0.0163	0.0164	0.0137	0.0000	0.0132
193	0.0384	0.0384	0.0384	0.0384	0.0384	0.0384	0.0384	0.0384	0.0384	0.0384
194	0.0000	**0.0001**	4E-05	0.0000	0.0000	3E-05	9E-05	3E-05	**0.0001**	3E-05
195	0.6085	**0.6208**	0.6181	0.6100	0.6035	0.5929	0.5929	0.5941	0.4291	0.5128
196	**0.0332**	**0.0332**	0.0306	0.0176	0.0122	0.0149	0.0244	0.0335	0.0210	0.0031

TABLE VI
SEARCH PERFORMANCE OF TEN ONTOLOGY MEASURES. THE BEST RESULTS ARE BOLD.

Concept selection	TRECVID 2006		TRECVID 2005	
	Single	Multiple	Single	Multiple
OSS	0.0384	0.0410	**0.1056**	0.1186
OS^2	**0.0385**	**0.0424**	**0.1056**	**0.1235**
LCH [20]	0.0363	0.0406	0.0963	0.0914
WUP [21]	0.0363	0.0379	0.0968	0.0828
RES [10]	0.0364	0.0390	0.0650	0.0516
LIN [22]	0.0363	0.0383	0.0650	0.0531
JCN [23]	0.0364	0.0398	0.0650	0.0728
Lesk [24]	0.0363	0.0410	0.0962	0.0962
Vect [25]	0.0384	0.0279	0.1040	0.0567
VP [25]	0.0359	0.0366	0.0755	0.0898

This again demonstrates the capability of OS^2 in offering pairwise concept similarities appropriate for the fusion of detector outputs. Comparing the MAP, multi-concept selection strategy basically improves the search performance of all measures (except Vect) over single-concept. While Vect shows very competitive performance in single concept selection, the MAP degrades significantly when multiple concepts are considered. The similarity value given by Vect appears to be less reliable and can be easily distorted with noise in WordNet.

Table VII further details the search performance of each measure in multiple concept selection. Among the 48 search topics, OS^2 obtains the best performance in 23 topics, followed by OSS which performs the best in 12 topics. For certain categories of search topics, such as topics involve name entity (e.g., ID-149) and motion or event (e.g., ID-173), the advantage of having semantic axes is not apparent. Among all the search queries, topics ID-156 and ID-195 have the most positive influence towards the overall performance. By removing these two queries from experiments, OS^2 and OSS still obtain the best MAP performance.

Table VIII lists the top-3 selected VIREO-374 detectors by OS^2. By manually browsing the detectors selected by various ontology measures, OS^2 is always able to pick up the semantically appropriate top-3 detectors. The AP of few topics (ID: 179, 180, 188, 190, 191, 194), nevertheless, is lower than 0.001. There are several reasons. For topics including name entities like topics 178, 179 and 194, text search is more appropriate than concept-based search in general. In addition, as we do not consider detector reliability, fusing with unreliable detectors will also degrade the performance. For instance, consider the detector *tie* in topic-180 and *firefighter* in topic-188. While both detectors are correctly picked, the detection performance is too low for expecting a reasonable search result. On the other hand, the specificity and coverage of concept detectors, which to certain extent can affect search performance, is not considered in our work. For instance, in topic-190, all the selected concepts are related to *person* but not *book*. The search performance can be improved, if by knowing that *book* is more specific than *person*, or by having a mechanism to select the set of detectors which are more diverse. These issues are outside the scope of this paper, but will be included in our future studies.

TABLE VIII
THE TOP THREE DETECTORS SELECTED BY OS^2 ON SEARCH TOPICS OF TRECVID 2006.

ID	Top-1	Top-2	Top-3
173	*car*	*truck*	*police*
174	*building*	*house of worship*	*office building*
175	*vehicle*	*business people*	*bicycle*
176	*police*	*soldier*	*prisoner*
177	*building*	*protester*	*office building*
178	*george bush*	*face*	*head of state*
179	*person*	*face*	*protester*
180	*group*	*business people*	*tie*
181	*walking*	*face*	*george bush*
182	*military*	*vehicle*	*soldier*
183	*boat*	*ship*	*rowboat*
184	*computer*	*business people*	*group*
185	*newspaper*	*business people*	*group*
186	*lake*	*mountain*	*field*
187	*helicopter*	*airplane*	*vehicle*
188	*fire*	*explosion*	*firefighter*
189	*flag*	*group*	*business people*
190	*person*	*individual*	*protester*
191	*adult*	*child*	*person*
192	*greeting*	*face*	*body part*
193	*smokestack*	*tower*	*smoke*
194	*face*	*george bush*	*head of state*
195	*soccer*	*football*	*baseball*
196	*snow*	*landscape*	*graveyard*

1) Significance Test: To verify whether the performances of OSS and OS^2 are by chance, we further conduct significance test. The test is based on the randomization test [41] suggested by TRECVID, where the target number of iterations used in the randomization is $100,000$. At the 0.05 level of significance, OSS and OS^2 are significantly better than all the other measures in terms of single and multiple concept selections. The only exception is Vect where the performance is indistinguishable from OSS and OS^2 for single concept selection. Comparing the two proposed semantic spaces, OS^2 is considered better than OSS for multiple concept selection at the 0.15 level of significance. There is no significant difference between OS^2 and OSS for single concept selection.

VII. Conclusion and Future Work

We have presented our approaches in constructing two variants of ontology-enriched semantic space: OSS and OS^2 for concept-based video search. Both spaces can guarantee a consistent way of comparing concept similarity scores when performing query-to-concept mapping. Using VIREO-374 detectors, experimental results over 235 hours videos on 48 search topics of TRECVID 2005 and TRECVID 2006 have indicated and confirmed the feasibility of OSS and OS^2 for large-scale video search. Compared with the traditional measures such as Resnik and WUP, both semantic spaces offer better search performance. Compared with OSS, OS^2 shows better performance in the fusion of multiple concept detectors due to the employment of orthogonal bases.

While encouraging, there are a couple of issues not being addressed in our current work and worth for future consideration. For instance, the occurrence of concepts, in addition to their ontological relatedness, can be explored for modeling a semantic space more viable for multimedia search. The co-occurrence statistics ideally hint the observability and discriminativeness of multimedia-based concepts. Incorporating this information could possibly enlighten the construction of a space that provides hint to select the most diverse and discriminant set of concepts for query answering. In addition to the positively correlated concepts, the set of negative concepts (e.g., indoor versus outdoor) is also a useful piece of information for the fast pruning of search results as presented in [42]. Whether and how the frequently, positively and negatively correlated concepts can be embedded in a semantic space for effective video search will be the topic of our future studies.

References

[1] M. Campbell, S. Ebadollahi, D. Joshi, M. Naphade, A. P. Natsev, J. Seidl, J. R. Smith, K. Scheinberg, J. Tešić, L. Xie, and A. Haubold, "IBM research TRECVID-2006 video retrieval system," in *TRECVID*, 2006, pp. 175–182.

[2] S. F. Chang, W. Hsu, W. Jiang, L. Kennedy, D. Xu, A. Yanagawa, and E. Zavesky, "Columbia university TRECVID-2006 video search and high-level feature extraction," in *TRECVID*, 2006, pp. 99–109.

[3] S.-Y. Neo, J. Zhao, M.-Y. Kan, and T.-S. Chua, "Video retrieval using high level features: Exploiting query matching and confidence-based weighting," in *Intl. Conf. on Image and Video Retrieval (CIVR)*, 2006.

[4] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring, "Adding semantics to detectors for video retrieval," *IEEE Transaction on Multimedia*, vol. 9, no. 5, pp. 975–986, 2007.

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

[6] A. Jaimes, B. L. Tseng, and J. R. Smith, "Modal keywords, ontologies, and reasoning for video understanding," in *Intl. Conf. on Image and Video Retrieval (CIVR)*, 2003, pp. 248–259.

[7] Y. Wu, B. L. Tseng, and J. R. Smith, "Ontology-based multi-classification learning for video concept detection," in *IEEE Intl. Conf. on Multimedia and Expo (ICME)*, vol. 2, 2004, pp. 1003–1006.

[8] C. Fellbaum, *WordNet: an electronic lexical database*, 1998.

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

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

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

[12] M. Naphade, J. R. Smith, J. Tešić, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE MultiMedia*, vol. 13, no. 3, pp. 86–91, 2006.

[13] A. Yanagawa, S. F. Chang, L. Kennedy, and W. Hsu, "Columbia university's baseline detectors for 374 LSCOM semantic visual concepts," Columbia University, Tech. Rep., 2007.

[14] 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.

[15] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *ACM Intl. Conf. on Multimedia (MM)*, 2006, pp. 421–430.

[16] C. G. M. Snoek, J. C. van Gemert, T. Gevers, B. Huurnink, D. C. Koelma, M. van Liempt, O. de Rooij, K. E. A. van de Sande, F. J. Seinstra, A. W. M. Smeulders, A. H. C. Thean, C. J. Veenman, and M. Worring, "The MediaMill TRECVID 2006 semantic video search engine," in *TRECVID*, 2006, pp. 277–290.

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

[18] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar, "Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news," *IEEE Transaction on Multimedia*, vol. 9, no. 5, pp. 958–966, 2007.

[19] A. Budanitsky and G. Hirst, "Evaluating WordNet-based measures of lexical semantic relatedness," *Computational Linguistics*, pp. 13–47, 2006.

[20] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," pp. 265–283, 1998.

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

[22] D. Lin, "Using syntactic dependency as local context to resolve word sense ambiguity," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 1997, pp. 64–71.

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

[24] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine code from an ice cream cone," in *the 5th Annual Intl. Conf. on Systems Documentation*, 1986, pp. 24–26.

[25] S. Patwardhan and T. Pedersen, "Using WordNet-based context vectors to estimate the semantic relatedness of concepts," in *Conf. of the European Chapter of the Association for Computational Linguistics (EACL)*, 2006.

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

[27] X. Li, D. Wang, J. Li, and B. Zhang, "Video search in concept subspace: A text-like paradigm," in *Intl. Conf. on Image and Video Retrieval (CIVR)*, 2007.

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

[29] A. Hoogs, J. Rittscher, G. Stein, and J. Schmiederer, "Video content annotation using visual analysis and a large semantic knowledgebase," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2003, pp. 327–334.

[30] H. Luo and J. Fan, "Building concept ontology for medical video annotation," in *ACM Intl. Conf. on Multimedia (MM)*, 2006, pp. 57–60.

[31] L. Hollink, M. Worring, and A. T. Schreiber, "Building a visual ontology for video retrieval," in *ACM Intl. Conf. on Multimedia (MM)*, 2005.

[32] H. Wang, S. Liu, and L.-T. Chia, "Does ontology help in image retrieval? A comparison between keyword, text ontology and multimodality ontology approaches," in *ACM Intl. Conf. on Multimedia (MM)*, 2006, pp. 109–112.

[33] C. T. Jr., R. F. S. Filho, A. J. M. Traina, M. R. Vieira, and C. Faloutsos, "The OMNI-family of all-purpose access methods: a simple and effective way to make similarity search more efficient," *The Intl. Journal on Very Large Data Bases (VLDB)*, vol. 16, no. 4, pp. 483–503, 2007.

[34] A. Berenzweig, D. P. W. Ellis, and S. Lawrence, "Anchor space for classification and similarity measurement of music," in *IEEE Intl. Conf. on Multimedia and Expo (ICME)*, 2003.

[35] J. Vleugels and R. C. Veltkamp, "Efficient image retrieval through vantage objects," *Pattern Recognition*, vol. 35, no. 1, pp. 69–80, 2002.

[36] M. Slaney, "Mixtures of probability experts for audio retrieval and indexing," in *IEEE Intl. Conf. on Multimedia and Expo (ICME)*, 2002.

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

[38] R. Penrose, "A generalized inverse for matrices," *Proceedings of the Cambridge Philosophical Society*, vol. 51, pp. 406–413, 1955.

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

[40] S. D. Bhavani, T. S. Rani, and R. S. Bapi, "Feature selection using correlation fractal dimension: Issues and applications in binary classification problems," *Applied Soft Computing*, vol. 8, no. 1, pp. 555–563, 2008.

[41] J. P. Romano, "On the behavior of randomization tests without a group invariance assumption," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 686–692, 1990.

[42] W. H. Lin and A. Hauptmann, "Which thousand words are worth a picture? Experiments on video retrieval using a thousand concepts," in *IEEE Intl. Conf. on Multimedia and Expo (ICME)*, 2006, pp. 41–44.