Ontology-Enriched Semantic Space for Video Search

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

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

ABSTRACT

Multimedia-based ontology construction and reasoning have recently been recognized as two important issues in video search, particularly for bridging semantic gap. The lack of coincidence between low-level features and user expectation makes concept-based ontology reasoning an attractive mid-level framework for interpreting high-level semantics. In this paper, we propose a novel model, namely ontology-enriched semantic space (OSS), to provide a computable platform for modeling and reasoning concepts in a linear space. OSS enlightens the possibility of answering conceptual questions such as a high coverage of semantic space with minimal set of concepts, and the set of concepts to be developed for video search. More importantly, the query-to-concept mapping can be more reasonably conducted by guaranteeing the uniform and consistent comparison of concept scores for video search. We explore OSS for several tasks including concept-based video search, word sense disambiguation and multi-modality fusion. Our empirical findings show that OSS is a feasible solution to timely issues such as the measurement of concept combination and query-concept dependent fusion.

Categories and Subject Descriptors

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

General Terms

Design, Algorithms, Performance, Experimentation

Keywords

Ontology, Semantic Space, Concept-based Video Search

1. INTRODUCTION

Semantic-based retrieval has been one of the long-term goals in multimedia computing. Traditional content-based approaches of deriving semantics *purely* based on low-level features have proven their limitations in conquering the so-called semantic gap. Modern approaches enable the semantic search by pooling a set of concepts and thus forming a semantic space to facilitate the high-level understanding of user queries and low-level features [2, 3, 18, 19, 22]. The search methodology is usually referred to as the *concept-based* video search, as illustrated in Figure 1. The sensory 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). Basically, a set of concept detectors is developed to represent the high-level semantics. The detectors are automatically learnt with training examples described by multi-modality features. Given a user query, the best set of concepts that can describe the semantic of 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.

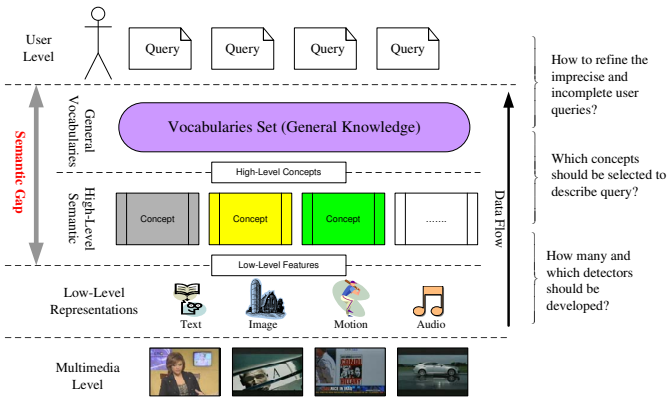

Figure 1: General framework of concept-based video retrieval.

Under the concept-based retrieval framework as depicted in Figure 1, there are several issues remain challenging. One fundamental problem is: which concept detectors should be and are feasible to be developed for search [6, 7, 15]? Ideally, the concept set could provide a high coverage of semantic space and is general and frequent enough so as to answer as many queries as possible [17]. On the other hand, given the concept set, the mapping ambiguity between queries and concepts need to be carefully resolved. A common solution is to consider the mapping through ontology reasoning [19, 22], or more precisely selecting the concepts, which minimize the linguistic distance with query terms. In the cur-

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

rent state-of-the-art, these two issues (concept development and query-concept mapping) are normally considered separately, where they are treated as two *independent* components. More specifically, the set of developed concepts is not exploited for query-concept mapping, and contradictorily, the mapping is *locally* determined in another semantic space spanned by a completely different (and much larger) concept space. This inconsistency indeed causes the similarity scores of query terms and concepts not directly comparable, resulting in less meaningful matching when finding the "best concepts" to interpret query semantic.

In this paper, we propose a novel model called Ontology-enriched Semantic Space (OSS) to jointly consider the two aforementioned issues. First, we address the *scalability* issue of which set of detectors should be developed. We argue that scalability should be grounded on the generalization power of detectors in spanning the semantic space. OSS is a linear space spanned by a set of basis concepts modeled with ontology knowledge. Each basis is deliberately arranged to cover an approximately equal portion of subspace in OSS. Under this arrangement, the space can be generalized to answer queries even for unseen concept detectors. Secondly, OSS provides a *computable* space where the query-concept mapping can be directly reasoned. Because the concept relations are inherently encapsulated via linguistic similarity, OSS allows a uniform and global way of choosing detectors for query answering.

The conventional query-concept mapping is conducted by linearly arranging the available detectors as a concept list. Given the query terms, the desired detectors are identified from the list with ontology reasoning or linguistic measure [19, 22]. Denote $\mathcal{Q} = \{q_1, \ldots, q_m\}$ as a text query with m terms, and $\mathcal{C} = \{c_1, \ldots, c_n\}$ as the list with n concepts. Typically a matrix $\mathbf{L}$ whose entry $l(i,j)$ represents the linguistic similarity between a query term q_i and a concept c_j is computed via ontology. The top-k concepts which receive the highest scores in $\mathbf{L}$ are then selected for retrieval. We argue that the similarities in entries $l(i,j)$ are not comparable, since each similarity represents a *local* decision computed from a branch (or sub-tree) of ontology. Each branch has properties such as information content and depth peculiar only to the context of a branch. Consequently, the similarities across branches are not uniform. Comparing the similarities inferred from two different branches (e.g., transport and fruit) each having different properties results in global inconsistence. This indeed causes the selection of top-k concepts arbitrary. Take Figure 2(a) as an example, let concepts a to e as children and v_1 to v_3 as ancestors. Using measure such as Resnik [20], the concept pairs (a,b) and (a,c) will be the same, although (a,c) sharing 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 measured as they reside in different parts of the ontology which carry different statistic and structural information.

OSS aims to provide a *computable* platform that allows uniform and global comparison of concept pairs. With reference to Figure 2(b), the semantic space is represented as a linear space spanned with basis concepts enriched with ontology knowledge. The bases of OSS can be viewed as the "key-concepts" of the original ontology space. Supposing the ancestors v_1 to v_3 of Figure 2(a) are selected as the basis concepts of OSS, then one can linearly project the concepts a-e to the metric space with ontology reasoning. Such framework indeed sights several opportunities. First, the basis concepts provide a high coverage of semantic space, and are probably the ones that should be developed if they are feasible to be built with the current technology. Secondly, in contrast to the examples in Figure 2(a), the space guarantees global consistency in comparing the concept pairs like (a,b), (a,c) and (d,e).

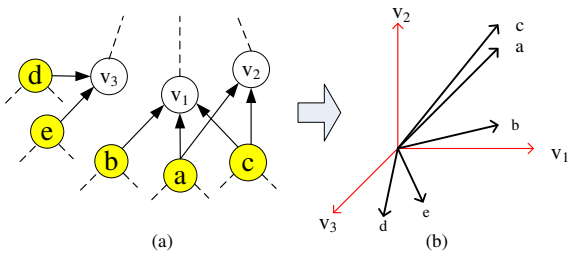

Figure 2: Reasoning with ontology (a) and OSS (b).

An intuitive explanation of OSS is that the space is linearly constructed to model the available set of concepts. The expressive power of OSS is linguistically spanned with a set of basis concepts, which is easier to generalize, not only to the available concept detectors but also to the unseen concepts. With OSS, we explore several search related tasks including concept selection and modality fusion in this paper. The major contributions of our works are briefly summarized as follows:

- *Scalability*: Building detectors for all concepts is impossible and not necessarily [7, 15]. A practical question is which detectors should be developed given the information at hand. Compared to recent works in [15], OSS provides another novel view of selecting concepts which have higher generalization ability in query answering.
- *Query-concept mapping*: With OSS, the mapping is no longer a local similarity comparison. Global consistency is ensured so that the selection of concepts becomes meaningful.
- *Multi-modality fusion*: A by-product of OSS is concept clusters. We demonstrate that the clusters can be exploited effectively for fusing the outcomes of concept-based search (visual) and ASR search (text), by taking into account the reliability of concept detectors.
- *Query disambiguation*: User queries are mostly ambiguous. We explore OSS to predict the search intention by finding the exact senses of query terms.

The remaining of this paper is organized as follows. Section 2 describes the related works. Section 3 proposes OSS, while Section 4 explores OSS for tasks such as concept development and query-concept mapping. Section 5 presents the experimental results. Finally, Section 6 concludes this paper and pinpoints future directions.

2. RELATED WORK

Concept-based video retrieval has recently attracted a new spurt of research attention, attributed to its potential in

bridging semantic gap. Two critical efforts are the detection of semantic concepts and the utilization of concepts as "semantic filters" for query answering. The current activities in LSCOM [17], MediaMill-101 [24] and TRECVID [25] further boost the common interest in building a large-scale ontology suitable for multimedia search and annotation. With these efforts and activities, encouraging findings have been reported regarding the usefulness of concepts for video search, compared to search with low-level features and text keywords [3, 19, 22, 23].

While encouraging, the issues of building concept ontology and performing query-concept mapping remain open and unsolved [3, 15, 23]. Multimedia and visual based ontology construction has been previously addressed in [8, 9, 16, 22, 26]. The construction mostly involves the manual mapping of visual elements to textual concept entities provided by shared vocabularies. In [9], WordNet is extended with visual tags describing properties such as visibility, motion and frequency of occurrence. In [8], 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 [22] 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 in [16, 26]. In [26], two animal domain ontologies are constructed separately for textual and visual descriptions. The study indicates that the ontologies are useful for image retrieval. Different from the existing ontology construction [8, 9, 22], our approach utilizes concepts to construct a semantic space which is computable and more viable for query-concept mapping.

Depending on the types (visual or text) of queries, the mapping from queries to concepts can be performed with detectors [2, 18] or resources such as ontology [19, 23, 22], text description [22] or co-occurrence statistic [2, 19]. For queries with image or video examples, the mapping is equivalent to pattern recognition problem. The responses of detectors basically indicate the likelihood of corresponding concepts present in queries. In [22], the best confident detector is selected, while in [2] a vector space model indicating the probabilities of concepts is utilized for search. For text queries, the mapping is usually performed through ontology reasoning which includes two steps: word sense disambiguation (WSD) and concept selection. In WSD, the exact senses (meanings) of query terms are estimated by checking the possible sense combinations through vocabularies. A popular algorithm for WSD is Lesk algorithm [1, 5] which automatically extracts the actual sense of each term. By knowing the senses of query terms, various ontology similarity measures can be directly employed for computing the association between terms and concepts. Popular measures include Resnik [20], JCN [11] and WUP [27] which consider ontological properties such as the specificity and information content of a concept, and the linguistic path length between two concepts. In addition to ontology reasoning, other approaches for mapping text queries are to compare queries against the text descriptions associated with concepts [22] or to expand queries with related terms [19]. The expanded terms as well as their weights can be learnt from training examples [2] or external information such as Internet [19]. Our proposed approach is mainly based on ontology reasoning. Both concepts and query terms are viewed as vectors (or points) in OSS for similarity (or distance) comparison.

3. MODELING SEMANTIC SPACE

Intuitively the abstract space of real world $\mathbb{R}$ can be viewed as the space spanned by low-level feature space ($\mathcal{L}$) and word or concept space ($\mathcal{W}$), i.e.,

$$\mathcal{L} \times \mathcal{W} \longrightarrow \mathbb{R} \tag{1}$$

We estimate the semantic space with the space spanned by concepts. Denote $c_i \in \mathcal{W}$ as the i^{th} concept, the semantic space is described by

$$\vec{c}_1 \times \vec{c}_2 \times \ldots \times \vec{c}_\infty \longrightarrow \mathbb{R} \tag{2}$$

where $\vec{c}_i$ is a basis concept. Each concept c_i is associated with a detector d_{c_i} (or classifier) learnt with low-level features as follows

$$d_{c_i} : \mathcal{L} \longrightarrow c_i \tag{3}$$

Generally modeling $\mathbb{R}$ with the known concepts and their detectors is computationally intractable due to the richness of human vocabulary and the limited computing power. Our aim here is to approximate $\mathbb{R}$ with the available set of concepts, and form a computable space that allows effective and meaningful comparison with unseen vocabularies.

3.1 OSS Construction

Given n concepts, the semantic space in Eqn (2) is constructed by computing the pair-wise similarity of each concept pair. The similarity is based on the ontology distance which utilizes the is-a structural relationship among 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 the specificity of concepts and the path length from one concept to the other by traversing the edges. 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. We employ WUP [27] to measure the ontology similarity of a concept pair c_i and c_j. WUP considers the speficity, path length and common ancestor of concepts. Let p_{ij} as the lowest common ancestor of c_i and c_j, the similarity is

$$WUP(c_i, c_j) = \frac{2D(p_{ij})}{L(c_i, c_j) + 2D(p_{ij}))} \tag{4}$$

where D returns the depth of a concept, and L gives the path length of two concepts.

Based on Eqn (4), a n-by-n matrix $\mathbf{O}$ which encapsulates the all-pair similarities of n concepts are computed. Each row of $\mathbf{O}$, denoted as $\mathbf{o}_i$, outlines the similarities of concept i with other concepts. The n-dimensional vector $\mathbf{o}_i$ is viewed as a concept vector $\vec{c}_i$, i.e., we set

$$\vec{c}_i = \mathbf{o}_i \tag{5}$$

To make the semantic space (SS) complete and compact, the basis concepts need to be estimated. Assuming SS is a linear

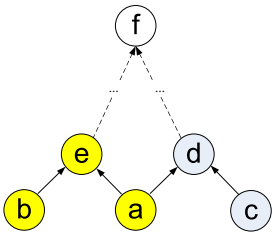

Figure 3: Path length based and information content based ontology measures are not metric, e.g., $(b,a)+(a,c) \leq (b,c)$.

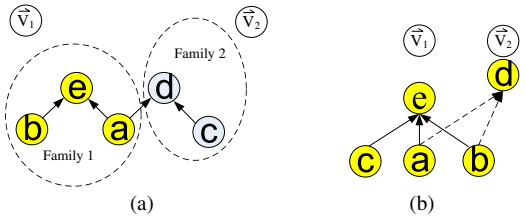

Figure 4: Measuring the concept similarity in WordNet with WUP. (a) The similarity of (a,b) is the same as (a,c), although a and b reside in a branch (Family-1) different from c (Family-2), and thus should have higher similarity. (b) The concept pairs (a,b), (a,c), (c,b) have the same WUP similarity, although a and b have another common ancestor d in addition to e, and thus should be more similar.

space, the bases can be found with techniques such as manifold learning, factor analysis and clustering techniques. Here we adopt clustering for estimation in order not to transform the space and to directly select a subset of concept vectors as the basis concepts. This would facilitate computation and make the space interpretable with each basis representing a concept, although other techniques like principal component analysis (PCA) can ensure the orthogonality of bases. Since the new space formed by $r < n$ selected basis concepts is ontology-enriched, we name the semantic space OSS – a linear approximation of semantic space enhanced with ontology relationship. With r bases, each concept is projected to OSS and described as a vector of r dimensions.

3.2 Metric in OSS

Since OSS is a linear space, many known metrics can be employed to characterize distance. We use cosine similarity for measuring the relatedness of concept vectors. Given an unseen concept c_u, a r-dimensional vector $\vec{c}_u$ is first formed by measuring the WUP similarity of c_u with r basis concepts c_b. The vector $\vec{c}_u$ can be represented as a linear expansion

$$\vec{c}_u = \sum_{b=1}^{r} \omega_b \vec{c}_b \tag{6}$$

where ω_b is the WUP similarity between c_u and c_b. The cosine similarity between two concepts c_u and c_i is then measured as

$$Sim(c_u, c_i) = \frac{\vec{c}_u \cdot \vec{c}_i}{|\vec{c}_u||\vec{c}_i|} \tag{7}$$

Note that the similarity is not only based on the ontology relationship between concepts c_u and c_i, but is also with respect to their relatedness to other basis concepts in OSS.

Compared to other ontology measures such as Resnik [20] and WUP [27], OSS is a metric space. It is not hard to show that other measures violate metric properties. Take the graph structure in Figure 3 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.

3.3 OSS versus WordNet

To fully reveal the benefit of OSS, we contrast the major difference of measuring concept similarity in OSS and in the original ontology space (WordNet). Figure 4 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 reside at the same level of depth, and concepts e and d are the ancestors. In Figure 4(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 OSS, suppose v_1 and v_2 are the basis concepts, where $\vec{v}_1$ is more related to Family-1 while $\vec{v}_2$ is more related to Family-2. By Eqn (7), 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 the basis of $\vec{v}_1$ and $\vec{v}_2$ as presented in Eqn (6). Similarly in Figure 4(b), the concepts pairs (a,b), (a,c) and (c,b) all have the same similarity with WUP, although a and b are more related because of sharing another common ancestor. Assuming the concept d is close to the basis $\vec{v}_2$ while concept e is close to $\vec{v}_1$, we can easily prove that $Sim(a,b) > Sim(a,c)$ in OSS.

In brief, the concept similarity in OSS is *globally* determined with the aid of basis concepts. While in WordNet, most linguistic reasoning methods utilize the local structure (depth, path length, specificity) peculiar to a sub-graph for measuring similarity. Consequently, a uniform and objective comparison of similarity scores obtained from different sub-graphs of WordNet becomes difficult.

4. EXPLORING OSS

With OSS as a computable platform, we explore several search related tasks including concept development, query-concept mapping and multi-modality fusion in this section.

4.1 Concept Development

Developing concept detectors is generally a time consuming task due to the need for collecting and annotating training samples. As a consequence, building detectors for all human-known concepts is unrealistic, but determining the kinds of concepts to be developed becomes a timely and practical issue. In [15], the empirical study indicates that the priority should be given to frequent concepts and scene-based concepts which could benefit most search queries. In OSS, the priority determination is based on the generalization power of concepts, by modeling the inter-concept relationship. The generalization can be identified based on the suitability of a concept being selected as a basis to fill the semantic space.

To illustrate the idea, we use MediaMill-101 concepts [24]

Table 1: Examples of basis concepts.

Basis	Cluster Members
water	river, water, waterfall
vehicle	tank, bicycle, vehicle
golf	soccer, football, golf
entertainment	racing, cycling, sports, entertainment

as an example to show the selection of basis concepts. Initially the matrix **O** which describes the ontological relationship among concepts is computed. Each concept is then represented as a high dimensional vector as in Eqn (5). The concept vectors are hierarchically clusters with agglomerative algorithm and consequently form a dendrogram as illustrated in Figure 5(a). With our approach, the correlation among concepts is nicely captured as observed in Figures 5(b)-(d) which highlights some groupings in the dendrogram. Basically concepts related to vehicle (5(b)) and sport (5(d)) are correctly grouped. To select the bases of MediaMill-101 concepts, we employ the inconsistent coefficient [10] to find the best possible concept clusters in the dendrogram (details in Section 5.1). The concepts nearer to cluster centroids are then selected from each cluster as the bases of OSS.

Table 1 shows the few selected basis concepts and their clusters. To enhance the generalization power of a semantic space, theoretically one should develop the detectors for basis concepts, which guarantee a high coverage. Nevertheless, practically some basis concepts are harder to build (e.g., *entertainment*) or less feasible than its member (e.g., *golf* versus *soccer*). Under these cases, the bases can still be utilized as references in the semantic space, while other developed concept detectors can model their ontological relationship with reference to the bases. Compared with [15] which concludes one should develop frequent and scene-based concepts, the MediaMill-101 bases picked by our approach are mostly general concepts (e.g., *water* and *vehicle*), implying more training examples for concept development. While this might be a good news that the selected basis concepts are easier to build, there is still a fundamental issue that general concepts normally include more varieties in appearance and thus could be harder to develop (e.g., *vehicle* versus *car*).

4.2 Query-to-Concept Mapping

Given a text query $\mathcal{Q} = \{q_1, \ldots, q_m\}$ with m terms, the task is to rank n concepts in the set $\mathcal{C}$ according to their importance and relevancy to the query. The top-k concept detectors can then be utilized for video search. With OSS, the similarity between a query term q_i and a concept c_j is computed via Eqn (6) and Eqn (7). The top-1 concept, for instance, is selected as

$$\hat{c} = \underset{c_j \in \mathcal{C}}{\operatorname{argmax}} \, Sim(q_i, c_j) \;\; \forall q_i \in \mathcal{Q} \tag{8}$$

4.3 Word Sense Disambiguation (WSD)

A query term q_i normally carries multiple senses (meanings). WSD is to estimate the actual sense of q_i jointly with other senses of terms in query $\mathcal{Q}$. Suppose there are m terms and each term has p senses, there are m^p ways of interpreting $\mathcal{Q}$. A greedy approach commonly adopted in WSD is to find a combination that maximizes the overlap of senses for all terms in $\mathcal{Q}$. With OSS, the greedy approach can be easily

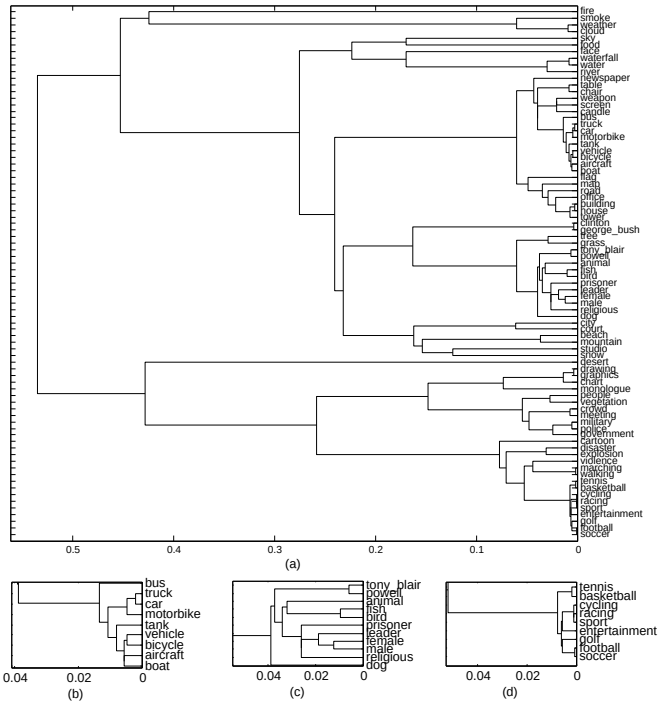

Figure 5: Dendrogram of MediaMill-101 detectors.

implemented by measuring the projection of senses to basis concepts. Denote s_i^k as the sense of q_i in k^{th} combination, the actual query sense $\hat{\mathcal{Q}} = \{\hat{s}_1, \ldots, \hat{s}_m\}$ is computed as

$$\hat{\mathcal{Q}} = \underset{1 \le k \le m^p}{\operatorname{argmax}} \, \phi(k) \tag{9}$$

where

$$\phi(k) = \sum_{i=1}^{m} \sum_{j=i+1}^{m} Sim(s_i^k, s_j^k) \tag{10}$$

The query $\hat{\mathcal{Q}}$ is then used for performing query-concept mapping. Basically, WSD is a query preprocessing step aiming to predict the search intention of queries which are usually short and imprecise.

4.4 Multi-modality Fusion

OSS guarantees the consistent measurement of concept similarity. In OSS, concepts can be effectively clustered according to their ontological relatedness. Figure 6 shows the 2-dimensional distribution of MediaMill-101 concepts in OSS with multi-dimensional scaling (MDS). Apparently, these concepts form few clusters sparsely distributed in the space. Our aim here is to explore the correlation of these concept clusters with multi-modality features, and subsequently utilize the correlation of cluster-modality pairs for fusion. The task is similar to query-class dependent fusion [28], where each cluster represents a class appropriate for answering a group of queries with similar type. In our case, a query $\mathcal{Q}$ is projected to OSS to locate one or multiple clusters, and the clusters provide information on how to fuse multi-modality features by examining the correlation of cluster-modality pairs.

We adopt the fuzzy synthetic evaluation [21] to estimate the correlation between concept clusters and multi-modality

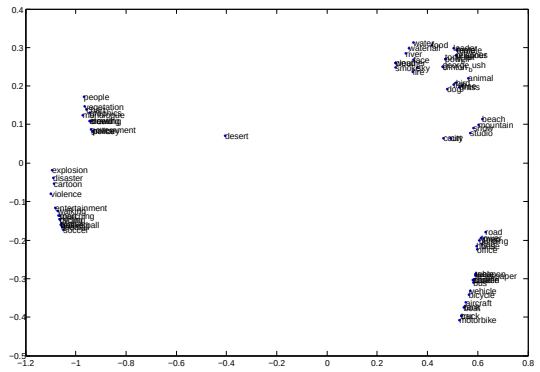

Figure 6: Two dimensional view of MediaMill-101 detectors in OSS.

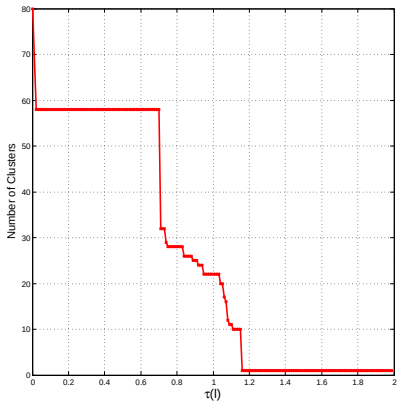

Figure 7: Basis selection: searching best possible clustering through the inconsistent coefficients of dendrogram.

features. Let $\mathcal{U} = [u_1, u_2, \ldots, u_n]$ as the order set with n clusters, and $\mathcal{V} = [v_1, v_2, \ldots, v_m]$ as the order set with m modalities. We want to estimate the fuzzy correlation of two sets, i.e., $\mathbf{R} \in \mathscr{F}(\mathcal{V} \times \mathcal{U})$, where $\mathbf{R}$ is a relation matrix and $\mathscr{F}$ is a fuzzy set. The matrix $\mathbf{R}$ is in the form of

$$\mathbf{R} = \begin{bmatrix} \mathbf{r}_{11} & \mathbf{r}_{12} & \mathbf{r}_{13} & \ldots & \mathbf{r}_{1n} \\ & & & \ldots & \\ \mathbf{r}_{m1} & \mathbf{r}_{m2} & \mathbf{r}_{m3} & \ldots & \mathbf{r}_{mn} \end{bmatrix} \quad (11)$$

where $\mathbf{r}_{ij}$ specifies the relation of modality i to cluster j. The relation matrix $\mathbf{R}$ can be estimated with training or subjective pairwise evaluation. We adopt the first scheme and the details will be discussed in the experiment (Section 5.4).

The relation $\mathbf{R}$ basically determines a fuzzy transformation $T_{\mathbf{R}}$ from $\mathcal{U}$ to $\mathcal{V}$, described as

$$T_{\mathbf{R}} : \mathscr{F}(U) \longrightarrow \mathscr{F}(V) \quad (12)$$

A query $\mathcal{Q} = \{q_1, q_2, \ldots\}$ is converted to a fuzzy vector $\mathcal{P} = [p_1, \ldots, p_n]$ by applying the membership function $U(q_i)$ for every query term. The function U assigns each q_i a belonging score to clusters via Eqn (7) in OSS, and then outputs $\mathcal{P}$ with p_i describes the similarity of clusters to query term q_i. The vector $\mathcal{P}$ is further transformed with fuzzy mapping as follows

$$T_{\mathbf{R}}(p_i) = p_i \circ \mathbf{R} \in \mathscr{F}(V) \quad \forall p_i \in \mathscr{F}(U) \quad (13)$$

where $\circ$ is a fuzzy composition operation. The transformation produces a vector $\mathcal{W} = [w_1, \ldots, w_m]$ with w_i specifies the weight (importance) of i^{th} modality with respect to the given query $\mathcal{Q}$. The vector $\mathcal{W}$ can be employed directly for the linear fusion of multi-modalities.

5. EXPERIMENTS

We use the automatic search task of TRECVID 2005 video dataset [25] for experiment evaluation. The video archive contains one-month's (November 2004) broadcast videos collecting from multi-lingual sources including English, Chinese and Arabic languages. The testing data, containing 79,484 shots (about 85 hours of videos), is used for assessing the performance of OSS in video search. Twenty four search topics (see Table 4), together with their ground-truth provided by TRECVID'05, are used for the experiments. We only use the text queries of search topics, imagining that most users perform search with a short description of words.

For the semantic concepts, we use MediaMill-101 concept detectors [24]. We remove the concepts not defined in WordNet, resulting in 80 detectors. The actual senses of each concept in WordNet are manually assigned based on visual impression. For ease of evaluation, we only assign one sense to each concept, although multiple sense assignment is possible in OSS. In the experiments, we test the selection of single and multiple concepts per search topic respectively for retrieval. The retrieved items (shots) are ranked according to their scores to the selected concept detector(s). The search performance is then evaluated with mean average precision (MAP), where AP is defined as

$$AP = \frac{1}{\min(R, k)} \sum_{j=1}^{k} \frac{R_j}{j} I_j \quad (14)$$

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 of all search topics.

5.1 Basis Selection vs. Search

We first demonstrate the construction of OSS by showing the selection of basis concepts versus search performance based on MediaMill-101 detectors. Agglomerative hierarchical clustering is initially performed to group the concepts, and Figure 5(a) shows the dendrogram. To select the OSS basis concepts, we employ the inconsistency coefficient [10] to find the best possible concept clusters in the dendrogram. Denote l as a link connecting two clusters, the coefficient $\tau(l)$ of the link is computed as

$$\tau(l) = \frac{len(l) - \mu(l)}{\sigma(l)} \quad (15)$$

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

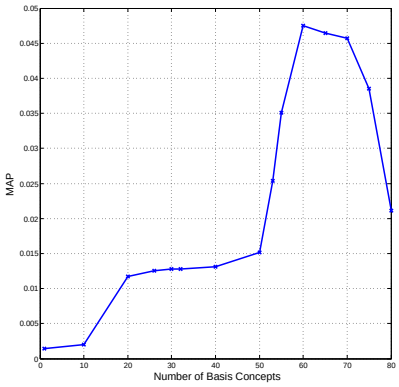

Figure 8: Performance of video search on different basis selections.

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 58 concept clusters, where the $\tau(l)$ increases slightly from 0 but with a dramatic jump of 80 to 58 concepts. Table 1 shows few basis concepts and their cluster members. Basically, a basis concept is the most centrally located concept in the cluster.

To verify that the choice of bases under the dendrogram of 80 concepts, we conduct an experiment to measure the search performance by varying the number of selected bases. Figure 8 shows the MAP of 24 search topics against different choices of OSS bases. Each search topic is assigned one concept in this experiment. The search performance improves when more bases are included to span the semantic space. The MAP reaches the highest when the number of bases is equal to 58. The performance starts to drop from this point onwards when more bases are considered. The result indeed aligns with the observation obtained from $\tau(l)$ coefficient which indicates 58 clusters are enough to represent the 80 concepts in MediaMill-101.

The performance of OSS could be theoretically explained by the completeness and independence of the space. Underestimating the number of bases 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 bases, on the other hand, affects the independence of basis vectors. The correlation of concept vectors could not be properly measured with the inclusion of the redundant information. Due to the use of clustering, the selected basis concepts are not strictly orthogonal to each other and are asymmetrically distributed. As the number of bases increases beyond a certain limit, the asymmetric distribution can actually bias the similarity measurement of concept vectors.

5.2 Word Sense Disambiguation

Prior to video search, a pre-processing step is to infer the actual senses of query terms. In this experiment, we explore OSS for word sense disambiguation as presented in Section 4.3. We compare the performance of OSS with Lesk algorithm [1] which is commonly adopted for sense disambiguation. Table 2 summarizes the comparison of experimenting 24 TRECVID search topics composing of 70 query terms. The topics form a total of 852 possible sense combinations. Both OSS and Lesk algorithm estimate the best combination of sense for each topic. The task is basically performed by finding the combination that maximizes the all-pair similarities of senses. In OSS, each sense is represented as a vector, and the similarity of two senses is directly computed with Eqn (7). In Lesk algorithm, the similarity is based on the amount of overlap (in words) between the definitions of two senses. The performance of OSS and Lesk is judged based on the percentage of senses being correctly predicted. In the experiment, the groundtruth of each term-sense assignment is manually judged. The result in Table 2 indicates that both OSS and Lesk can correctly predict 80% of actual senses. The performance of OSS is slightly better than Lesk algorithm. The result indeed shows the benefit of OSS − the ontology enriched space achieves competitive performance as Lesk algorithm even though the definitions of senses are not utilized for similarity comparison.

Table 2: Comparison of WSD

	Query Terms	Correct Sense	Accuracy
Lesk [1]	70	56	80.00%
OSS	70	58	82.85%

Table 3: Mean average precision (MAP) of video search with different ontology measures (single concept selection).

LCH	WUP	RES	LIN	JCN	OSS
0.0213	0.0213	0.0418	0.0104	0.0104	0.0486

5.3 Concept-based Video Search

We compare OSS with five other popular ontology measures: LCH [12], WUP [27], RES [20], LIN [14] and JCN [11] for video search. The first two measures use path length information, while the last three utilize information content [20]. In the experiments, we use WordNet as the ontology for all the measures. Denote D as the depth and I as the information content 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. The measures are defined as

$$LCH(c_i, c_j) = -\log \frac{L(c_i, c_j)}{2\delta} \quad (16)$$

$$WUP(c_i, c_j) = \frac{2D(p_{ij})}{L(c_i, c_j) + 2D(p_{ij}))} \quad (17)$$

$$RES(c_i, c_j) = I(p_{ij}) \quad (18)$$

$$LIN(c_i, c_j) = \frac{2I(p_{ij})}{I(c_i) + I(c_j)} \quad (19)$$

$$JCN(c_i, c_j) = \frac{1}{I(c_i) + I(c_j) - 2I(p_{ij})} \quad (20)$$

where δ is the maximum depth of WordNet. The information content is estimated based on the one-million-word Brown Corpus of American English [4]. For OSS, as presented in Section 5.1, a total of 58 basis concepts is selected. Each concept in OSS is thus represented as a 58-dimensional vector. The concept vectors are compared via cosine similarity as in Eqn (7).

Table 3 shows the performance comparison of six different measures with single-concept selection on 24 search topics.

Table 4: Comparison of Semantic Measures

		WUP		RES		OSS	
ID	**Topic**	**Detector**	**AP**	**Detector**	**AP**	**Detector**	**AP**
149	Condoleeza Rice	face	0.0007	face	0.0007	face	0.0007
150	Iyad Allawi	face	0.0001	face	0.0001	leader	8E-05
151	Omar Karami	face	0.0009	face	0.0009	face	0.0009
152	Hu Jintao	face	0.0005	face	0.0005	face	0.0005
153	Tony Blair	face	0.0006	face	0.0006	face	0.0006
154	Mahmoud Abbas	face	0.0025	face	0.0025	face	0.0025
155	Graphic map of Iraq with Baghdad marked	map	0.0069	map	0.0069	map	0.0069
156	Two visible tennis players on the court	court	0	tennis	0.6624	tennis	0.6624
157	People shaking hands	people	0.0018	people	0.0018	people	0.0018
158	Helicopter in flight	aircraft	0.0111	cloud	0.003	aircraft	0.0111
159	George Bush entering or leaving vehicle	face	0	face	0	vehicle	4E-05
160	Something on fire with flames and smoke	grass	0.0006	smoke	0.0111	smoke	0.0111
161	People with banners or signs	people	0.0009	people	0.0009	people	0.0009
162	People leaving or entering a building	people	0.0002	people	0.0002	people	0.0002
163	A meeting with large table and people	people	0.0097	meeting	0.0251	people	0.0097
164	A ship or boat	boat	0.0672	weather	0.0003	boat	0.0672
165	Basketball players on the court	court	0.0001	basketball	0.1529	court	0.0001
166	One or more palm trees	tree	0.0034	dog	0	tree	0.0034
167	An airplane taking off	aircraft	0.0082	aircraft	0.0082	aircraft	0.0082
168	A road with one or more cars	car	0.0756	car	0.0756	car	0.0756
169	One or more military vehicles	military	0.0365	vehicle	0.0187	vehicle	0.0187
170	A tall building	building	0.0276	building	0.0276	building	0.0276
171	A goal being made in a soccer match	soccer	0.2541	football	0.0007	soccer	0.2541
172	Office setting	office	0.0029	office	0.0029	office	0.0029
	MAP		0.0213		0.0418		0.0486

In the experiment, except OSS, all measures employ Lesk algorithm for word sense disambiguation. OSS is capable of estimating the actual senses in its own semantic space as presented in Section 5.2. The search result indicates that OSS outperforms other measures in terms of MAP performance. Interestingly the information content based measures exhibit very different performance, where RES relies on only the lowest common ancestor performs better than LIN and JCN. On the other hand, path length based approaches like LCH and WUP are not performed as well as RES. Our analysis shows that these five measures are less reliable and can be easily distorted with noise introduced in WordNet.

Table 4 lists the detailed performances of WUP, RES and OSS on the 24 search topics. Basically path length approaches like WUP are straightforward but sensitive to the outcomes of word sense disambiguation. When incorrect senses (e.g., Topic-160) are assigned, inappropriate concept detectors (e.g., *grass*) will be selected. Information content-based approaches like RES, on the other hand, are sensitive to the statistics of corpus. For instance, the concept *soccer* is not selected in Topic-171 simply because the information content of soccer is 0 in the corpus. OSS, using a completely different methodology, does not suffer from these shortcomings. The basis concepts provide a relatively robust measure by modeling the inter-concept relationship. The modeling makes the measure less sensitive to word disambiguation while guaranteeing global consistency of similarity scores. Comparing with the recent results ($MAP = 0.0485$) in [22] where there is a pool of 363 detectors for the 24 search topics, our results are indeed encouraging.

Table 5 shows the MAP performance of different approaches when the best three concepts are selected for query answering. Table 6 further lists the first three detectors selected by OSS. The MAP performances of all approaches, particularly the LCH, WUP and JCN, are improved when comparing to single concept selection. Overall, OSS still exhibits the best performance. However, the performances of few topics degrade. It is partially because these topics have less than three related concepts. Selecting multiple concepts may introduce irrelevant detectors (e.g., the *male* and *female* in Topic-166) and worsen the precision. On the other hand, the concept selection does not take into account the reliability of detectors. Including more detectors could probably degrade the performance supposing detectors with less reliability are selected (e.g., the *chair* in Topic-172).

Table 5: Mean average precision (MAP) of video search with different ontology measures (multiple concept selection).

LCH	WUP	RES	LIN	JCN	OSS
0.0460	0.0533	0.0423	0.0344	0.0475	0.0543

During retrieval, the score of a retrieved item is computed as the linear sum of the concept detectors' responses weighted by their similarities to the given query. We notice that the setting of weight is an important factor for analyzing the performances of difference measures. For instance, due to the limited pool of detectors, the first three selected concepts of OSS, WUP and JCN are similar. However, because OSS is able to assign proper weights to concepts due to the use of basis concepts, the semantic importance of concepts towards queries can be better characterized. For instance, in Topic-161, WUP assigns higher weights to *people* and *house* (1.0) than *crowd* (0.9). Similarly, JCN assigns 1.0 to *people*, 0.99 to *house* but 0.112 to *crowd*. OSS gives a relatively reasonable weight combination (1.0 for *people*, 0.94 for *house*, 0.935 for *crowd*). The consistency in concept measurement, and thus the ability in assigning proper concept combination, indeed leads to the performance stability of OSS in both single and multiple concept selections.

Table 6: Multiple concept selection with OSS.

Topic ID	Selected Detectors			AP
	1st	2nd	3rd	
149	face	graphics	candle	0.0009
150	leader	face	female	0.0001
151	face	leader	female	0.0009
152	face	leader	female	0.0002
153	face	graphics	candle	0.0006
154	face	leader	female	0.0025
155	map	city	road	0.0182
156	tennis	court	basketball	0.6623
157	people	house	violence	9E-05
158	aircraft	boat	vehicle	0.0031
159	vehicle	bus	bicycle	0.0001
160	smoke	fire	grass	0.0006
161	people	house	crowd	0.0013
162	people	building	house	0.0001
163	people	table	meeting	0.0195
164	boat	aircraft	vehicle	0.0062
165	court	basketball	leader	0.1573
166	tree	female	male	0.0012
167	aircraft	boat	vehicle	0.0151
168	car	truck	bicycle	0.0871
169	vehicle	military	police	0.0423
170	building	house	tower	0.0236
171	soccer	football	golf	0.2587
172	office	studio	chair	0.0016

5.4 Multi-Modality Fusion

In this experiment, we demonstrate that OSS can be employed for multi-modality fusion, by considering the abilities of concept clusters in query answering. Through training, we estimate the importance of each concept cluster to two modalities: retrieval-by-ASR (text baseline) and retrieval-by-concept. The information gathered for each cluster is then adopted for the late fusion of two modalities. As presented in Section 4.4, the task is similar to query-class dependent fusion, where each concept cluster is in charge of answering a group of queries which requires a specific way of multi-modality fusion. However, different from conventional approaches [28], we do not perform query classification. Instead, we measure the significance of each concept cluster to a query directly in the ontology-enriched semantic space. All clusters are involved in determining the weights for fusion depending on their significance towards a query.

Based on the OSS constructed in Section 5.1, each concept is described as a 58-dimensional vector. We empirically set the number of concept clusters as 14 and apply k-means to divide the concepts into 14 partitions. We add in one extra cluster for "name entity" resulting in 15 concept clusters as listed in Table 7. With the clusters, we learn the relation matrix $\mathbf{R}$ in Eqn (11) using the TRECVID 2005 development set. Denote V as the retrieval-by-concept modality, and T as the retrieval-by-ASR modality. For V-modality, the weight $\mathbf{r}_{Vi}$ for cluster i is estimated based on the MAP of its detectors in the development set. For T-modality, the weight $\mathbf{r}_{Ti}$ for cluster i is estimated by the probability of finding their concepts in the speech transcripts of concept ground-truth. For instance in Cluster-10, the $\mathbf{r}_{V10}$ is the MAP performance of three detectors (*beach*, *mountain*, *snow*), while $\mathbf{r}_{T10}$ is the probability of observing the three words in the transcripts of shots containing the three concepts. One exception is Cluster-15 for name entity, where we set $\mathbf{r}_{V15} = 0$ and $\mathbf{r}_{T15} = 1$, assuming that there is no concept detector available for the majority of name entities.

Table 7: Concept clusters for multi-modality fusion.

ID	Concepts
1	cartoon, chart, drawing, graphics, monologue
2	crowd, government, meeting, military, people, police, vegetation
3	basketball, cycling, entertainment, football, golf, marching, racing, soccer,sport, tennis, walking
4	disaster, explosion
5	violence, face, food, river, water, waterfall
6	cloud, fire, sky, smoke, weather
7	female, leader, male, prisoner, religious
8	animal, bird, dog, fish, grass, tree
9	city, court, studio
10	beach, mountain, snow
11	building, flag, house, map, office, road, tower
12	aircraft, bicycle, boat, bus, car, motorbike, tank, truck, vehicle
13	candle, chair, newspaper, screen, table, weapon
14	desert
15	named entities

Given a query $\mathcal{Q}$, a fuzzy vector $\mathcal{P}$ is first obtained by applying Eqn (12). The vector is computed by measuring the similarity of each query term with the cluster centroids in the semantic space. With Eqn (13), $\mathcal{P}$ and $\mathbf{R}$ are combined through fuzzy composition. The transformation then produces the fusion weights $\mathcal{W} = [\mathcal{W}_V, \mathcal{W}_T]$ for retrieval by concept and ASR modalities respectively.

To verify the proposed approach, we compare our approach (OSS) with two heuristic linear fusion strategies: weighted average fusion (WAF) and pseudo query-class dependent fusion (PQF). In WAF, we empirically assign the fusion weights of 0.6 to ASR modality and 0.4 to concept modality. In PQR, for queries with name entity, the ASR modality is given a weight of 0.7 while the concept modality is given 0.3. Otherwise, we set the weights of both modalities as 0.5. The performances are compared against the text baseline (retrieval-by-ASR) implemented based on Lemur [13].

Table 8 shows the comparison of four different approaches. Basically all fusion techniques improve over the ASR baseline, indicating the usefulness of concept modality. Among these techniques, OSS achieves the highest AP for 13 search topics, most of them are non-name-entity queries. Overall, compared with the fixed-weight setting as in WAF and PQR, OSS achieves the most improvement (53.46%) over the baseline. This indicates that OSS, incorporating with fuzzy transformation, is capable of estimating appropriate fusion weights. Take Topic-156 and Topic-171 as examples, OSS shows significant improvement as the related concepts (Cluster-3) are semantically grouped, and the relation matrix $\mathbf{R}$ takes into account the reliability of detectors. We find that there are 6 topics, including 2 name entity queries, where OSS does not improve over baseline. We investigate the results and notice that the reason is indeed because we use cluster centroids for comparing the similarities with query terms. This somehow makes the fusion weights under or over estimating the importance of particular modality. We believe a better cluster-query similarity measurement can further boost the performance of OSS.

6. CONCLUSION

We have presented OSS as a new computable platform for the uniform and consistent measurement of concept similar-

Table 8: Comparison of different fusion techniques.

Topic-ID	Baseline	WAF	PQF	OSS
149	0.0337	0.0493	0.0454	**0.0494**
150	0.0154	0.0125	0.0144	0.0152
151	0.2335	0.2411	0.2460	**0.2461**
152	0.2190	0.2817	0.2778	0.2711
153	0.2590	0.2575	0.2578	0.2563
154	0.1559	0.1553	0.1629	0.1597
155	0.000	0.0054	0.0028	**0.0067**
156	0.0063	0.283	0.4466	**0.4725**
157	0.0013	0.0014	0.0015	**0.0028**
158	0.0169	0.0227	0.0213	**0.0262**
159	0.0023	0.0056	0.0048	0.0043
160	0.0218	0.0236	0.0261	**0.027**
161	0.0357	0.0341	0.0326	0.0349
162	0.0018	0.0018	0.0019	0.0014
163	0.0105	0.0201	0.0239	0.0234
164	0.0892	0.0492	0.0268	0.0476
165	0.0165	0.1022	0.1424	0.1613
166	0.004	0.0038	0.0038	**0.0038**
167	0.000	0.0011	0.0017	**0.0017**
168	0.0453	0.0792	0.0949	**0.0952**
169	0.0729	0.0672	0.0522	0.0683
170	0.0057	0.0081	0.0126	**0.013**
171	0.1817	0.1953	0.1838	**0.2104**
172	0.0149	0.0145	0.0149	**0.0152**
MAP	0.0601	0.0798	0.0874	0.0922
Improve	-	32.78%	45.42%	53.46%

ity and combination. The platform, aiming at a high coverage of semantic space with a minimal concept set, shapes the ways of modeling concept inter-relatedness, while providing guideline for concept development. To show the feasibility of OSS, we explore and experiment several search related tasks including query-concept mapping and multi-modality fusion. Our findings show that, due to the uniform way of assessing similarity, OSS is a feasible solution for large-scale video search, concept combination, and query dependent fusion with concept clusters. Currently we assume that OSS exists in a linear space for computational reason. Whether a nonlinear space assumption is feasible for OSS remains an unanswered issue that worths further investigation.

A useful resource currently not explored in OSS is the co-occurrence statistics of concepts in video data. The statistics can be directly utilized for basis concept selection, amending the semantic space such that the co-occurred behavior can also be modeled. Under such circumstance, the space is enriched with both ontology semantic and statistics useful for video search. Developing the basis concepts in this space as detectors could be more realistic since the statistics indeed hint the utility and observability of the concepts. In addition to positively correlated concepts, the set of negative concepts (e.g., *indoor* versus *outdoor*) is also a useful piece of information for fast pruning in video search as presented in [15]. It is possible to have another "negatively correlated" semantic space, complementary to OSS, to allow fast filtering one on hand, and effective searching on the other hand. We will consider both aspects (co-occurrence and negative correlation) as the future extension of OSS.

7. 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).

8. REFERENCES

[1] S. Banerjee and et. al. An adapted lesk algorithm for word sense disambiguation using WordNet. In *Computational Linguistics and Intelligent Text Processing*, 2002.

[2] M. Campbell and et. al. IBM Research TRECVID-2006 video retrieval system. In *TRECVID*, 2006.

[3] S. F. Chang, W. Hsu, and et. al. Columbia University TRECVID-2006 video search and high-level feature extraction. In *TRECVID*, 2006.

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

[5] A. Haubold, A. Natsev, and M. R. Naphade. Semantic multimedia retrieval using lexical query expansion and model-based reranking. In *ICME*, 2006.

[6] A. Hauptmann. Towards a large scale concept ontology for broadcast video. In *CIVR*, 2004.

[7] A. Hauptmann, R. Yan, and W. H. Lin. How many high-level concepts will fill the semantic gap in video retrieval? In *CIVR*, 2007.

[8] L. Hollink, M. Worring, and A. T. Schreiber. Building a visual ontology for video retrieval. In *ACM MM*, 2005.

[9] A. Hoogs, J. Rittscher, G. Stein, and J. Schmiederer. Video content annotation using visual analysis and a large semantic knowledgebase. In *CVPR*, 2003.

[10] A. Jain and R. Dube. *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, NJ, 1988.

[11] J. J. Jiang. Semantic similarity based on corpus statistics and lexical taxonomy. In *Int. Conf. Research on Computational Linguistics*, 1997.

[12] C. Leacock and M. Chodorow. *Combining local context and wordnet similarity for word sense identification*. MIT Press, 1998.

[13] Lemur Toolkit. In *http://www.lemurproject.org/*, 2006.

[14] D. Lin. Using syntactic dependency as local context to resolve word sense ambiguity. In *ACL*, 1997.

[15] W.-H. Lin and A. Hauptmann. Which thousand words are worth a picture? experiments on video retrieval using a thousand concepts. In *ICME*, 2006.

[16] H. Luo and J. Fan. Building concept ontology for medical video annotation. In *ACM Multimedia*, 2006.

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

[18] A. Natsev, M. R. Naphade, and J. R. Smith. Semantic representation, search and mining of multimedia content. In *ACM 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 *CIVR*, 2006.

[20] P. Resnik. Using information content to evaluate semantic similarity in taxonomy. In *IJCAI*, 1995.

[21] T. Ross. *Fuzzy logic with engineering applications*. John Wiley, 2004.

[22] C. Snoek and et. al. Adding semantics to detectors for video retrieval. *IEEE Trans. on Multimedia*, 2007.

[23] C. Snoek, J. Gemert, and et. al. The MediaMill TRECVID 2006 semantic video search engine. In *TRECVID*, 2006.

[24] C. Snoek, M. Worring, J. Gemert, and et. al. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *ACM MM*, 2006.

[25] TRECVID. In *http://www-nlpir.nist.gov/projects/trecvid/*.

[26] H. Wang, S. Liu, and L.-T. Chia. Does ontology help in image retrieval - A comparison between keyword, text ontology and multi-modality ontology approaches. In *ACM Multimedia*, pages 109–112, 2006.

[27] Z. Wu and M. Palmer. Verb semantic and lexical selection. In *Annual Meeting of the ACL*, pages 133–138, 1994.

[28] R. Yan, J. Yang, and A. Hauptmann. Learning query-class dependent weights in automatic video retrieval. In *ACM MM*, pages 548–555, 2004.