Semantic Video Search by Exploiting Large-Scale Visual Concepts

Xiao-Yong Wei, Yu-Gang Jiang and Chong-Wah Ngo
Department of Computer Science
City University of Hong Kong, Kowloon, Hong Kong
{xiaoyong,yjiang,cwngo}@cs.cityu.edu.hk

ABSTRACT
This paper summarizes our recent research works on semantic video search. Our solution to this exciting topic is grounded on concept-based video search, which attempts to narrow the semantic gap by first indexing the video content with a large number of pre-trained semantic concepts, and then employing these concepts to respond user’s queries (query-by-concept). We developed novel and effective techniques used in both stages, i.e., concept detection and semantic query-to-concept mapping. Our findings lead to several research prototypes, including VIREO-374 concept detectors [1] and query-to-concept mapping [4], which show state-of-the-art performance in TRECVID 2008 concept detection and automatic video search.

1. INTRODUCTION
With the ever increasing popularity of digital videos, there is a need to develop techniques for searching large-scale video databases. The predominant user queries in this scenario are text, conveying certain semantic meanings of interest. However, computer can only extract low-level features (e.g., color, texture) from the videos, based on which it is hard to infer any semantic interpretation. This discrepancy is popularly referred to as the semantic gap.

Recently, as illustrated in Figure 1, a promising direction for bridging the gap is to introduce an intermediate layer between the semantic queries and computable low level features, namely high level semantic concept. The concepts are predefined so as to cover a wide range of popular topics that are useful for semantic search. Through developing automatic detectors for these predefined concepts, an incoming user query such as “find video clips with an airplane” can be easily answered by returning the top ranked results from a pre-developed detector of concept “airplane”. This video search framework is called concept-based video search. Two of main challenges in this framework are: 1) the development of large-scale concept detectors, with appropriate set and size, for semantic detection and search, 2) the selection of appropriate concept detectors for answering user queries.

For the first issue, we adopt an existing concept lexicon definition and consider the problem of constructing effective feature representation for concept detection. Starting from the bag-of-visual-words representation method, we performed a comprehensive study of several factors that will affect the performance of concept detection [1]. We have

Figure 1: Framework of concept-based video search. Showcased our technique in the annual NIST TREC video retrieval evaluation (TRECVID) event [2], and achieved top performance in concept detection. We have also extended our approach to detect 374 semantic concepts and released the detectors to the research community (http://vireo.cs.cityu.edu.hk/research/vireo374/).

The second challenging issue is to find an optimal combination of concept detectors for a query. Since the size of the concept lexicon is much smaller than the English vocabulary used by human, this mapping problem is usually much more complicated than the example given above in which we could find a concept with its name appeared in the query. For instance, given a query of “find video clips with weapon” and suppose “weapon” does not belong to our concept lexicon, we need to develop advanced techniques to find the concept(s) that is semantically relevant to the query, say “military”. To tackle this issue, we proposed two novel spaces, namely semantic space (SS) and observability space (OS) [5, 4]. These two spaces are complementary to each other in the way that SS models the semantic relationship among concepts while OS models the co-occurrence relationship. An integrative utilization of SS and OS makes the concept selection more complete and the concept fusion more impartial, and furthermore, enables exploring more aspects of semantic concepts, such as reliability and diversity of detectors.

2. BUILDING LARGE-SCALE CONCEPT DETECTORS
In order to identify the right set of detectors to develop, collaborative effort from various research organizations have been pooled to assess the utility, observability, and flexibility of concepts. One typical example is the release of LSCOM (http://www.lscom.org/) which includes 894 concepts. The concepts cover a wide range of topics, including objects (e.g.,...
our approach to a large set of 374 LSCOM concepts. The choices 1-7 (see Figure 2 for more details).

Figure 2: Comparison of our concept detection performance with all 161 type-A submissions in NIST TRECVID 2008. Our submissions are shown in red and the rest are from other research teams. Note that for the two submissions with performance around median, no any new training data are used.

Detecting these concepts is highly challenging. Popularly used global features only describe the overall distribution of color and texture information of a video frame, which are not adequate for concept detection due to serious background clutter and occlusion that exist in unconstrained video databases. Instead we focus on the utilization of local invariant features (keypoints) such as SIFT (Scale Invariant Feature Transform), which has been shown to be effective for many computer vision tasks. Based on the local keypoints, a video frame can be represented as a “bag-of-visual-words (BoW)” which is analogous to the bag-of-words representation of textual documents in IR.

The performance of BoW in semantic concept detection is subject to various representation choices. Thus we conducted a comprehensive study on these choices, aiming to provide a guideline of effectively using BoW for concept detection [1]. Specifically, we evaluated the effect of: 1) visual vocabulary size; 2) word weighting scheme; 3) stop-word removal; 4) feature selection; 5) spatial information; 6) visual bi-gram; and 7) visual word proximity and linguistics. We shown that with a careful design of representation choices, the performance of BoW for concept detection can easily double that without these considerations [1].

In the annual NIST TRECVID activity, our technique based on BoW has been able to offer very competitive performance – our overall performance in term of mean average precision ranked 3rd in year 2007 by considering all the representation choices 1-4, and 1st in year 2008 by considering the repre-

tion choices 1-7. We conducted a comprehensive study on these choices, aiming to provide a guideline of effectively using BoW for concept detection [1]. Specifically, we evaluated the effect of: 1) visual vocabulary size; 2) word weighting scheme; 3) stop-word removal; 4) feature selection; 5) spatial information; 6) visual bi-gram; and 7) visual word proximity and linguistics. We have shown that with a careful design of representation choices, the performance of BoW for concept detection can easily double that without these considerations [1].

In the annual NIST TRECVID activity, our technique based on BoW has been able to offer very competitive performance – our overall performance in term of mean average precision ranked 3rd in year 2007 by considering the representation choices 1-4, and 1st in year 2008 by considering all the choices 1-7 (see Figure 2 for more details).

To stimulate innovation of new techniques and reduce the replication effort of our concept detection method, we extend our approach to a large set of 374 LSCOM concepts, namely VIREO-374. We released the features, classifier models, and detection scores on several popular datasets, which are available at [http://vireo.cs.cityu.edu.hk/research/vireo374/]. VIREO-374 contains a large number of concept detectors and has been downloaded by nearly 50 research teams. Figure 3 shows the places where the recent visitors came from.

3. Concept-Based Video Search

With the concept detectors, another issue is to select the most semantically relevant concepts for a query. In this section

1The 374 concepts are those with more than 10 positive training samples in the LSCOM ground-truth annotation.

Figure 3: Visitor Map of VIREO-374 website.

3.1 Concept Modeling

3.1.1 Orthogonal and Ontology-Enriched SS

Instead of relying on traditional way of ontology reasoning, we construct an orthogonal semantic space (SS) for uniform comparison of concept similarity. Given a vocabulary set \( V = \{C_1, C_2, \ldots, C_n\} \) of \( n \) concepts, we first construct a \( n \times n \) concept similarity matrix \( R \) where each entry \( r_{ij} \) represents the traditional ontological similarity of a concept pair \((C_i, C_j)\). With the ontology-enriched matrix \( R \), basis of SS can be estimated by solving

\[
C^T C = R
\]  

Solving the equation by performing spectral decomposition to \( R \), the set of basis vectors \( C \) is estimated. The SS spanned by basis vectors in \( C \) is orthogonal while enriched by ontology knowledge learnt from \( R \). For unknown concept \( C_u \notin V \), the concept vector \( \hat{C}_u \) is predicted as

\[
\hat{C}_u = (C^T C)^{-1} \hat{R}_u
\]  

where \( \hat{R}_u \) is a vector obtained by computing the traditional ontological similarities of \( \hat{C}_u \) to the concepts in \( V \).

With the vector representations, the semantic relatedness between two concepts can be directly measured with cosine similarity, which is not simply based on traditional ontological similarity, but also the similarities of these two concepts with respect to the bases computed based on the matrix \( R \). In other words, comparing similarity of any two concepts is globally, instead of locally, measured in SS.

3.1.2 Observability Space

Complementary to SS, observability space (OS) gives cues to concepts of how they co-occur in video domain. To build OS, we adopt similar procedure as SS to learn concept observability. The main difference is that the matrix \( R \) is computed by learning from the pair-wise concept co-occurrence, instead of semantic relatedness. We employ Pearson product-moment (PM) correlation to compute concept observability. Applying PM to the concept pairs in vocabulary set \( V \), we form an observability matrix \( O \), which is similar to how \( R \) is formed in SS. \( O \) will be further decomposed to compute a new transformed space of observability. OS is orthogonal and offers a globally consistent space for observing the co-occurrence among concepts.

3.2 Reasoning and Searching

3.2.1 Concept Selection

Given a query and a pool of detectors, we select two groups of concepts from semantic and observability spaces respec-
The first group of concepts, denoted as $A$, refers to anchor concepts which describe the semantics of a query. The second group, denoted as $B$, is composed of bridge concepts which describe the co-occurrence relationship between the anchor concepts in $A$.

Let a user query be $Q = \{q_1, q_2, \ldots\}$ where $q_i$ is a query item carrying semantic meaning which could be represented as a vector via Eq. (2). With a detector set, the most relevant detector to $q_i$ is retrieved as an anchor concept of $Q$. Anchor concepts could be projected to OS to observe the frequency of co-occurrence among each other. We first perform clustering to group anchor concepts which are close in OS. This forms several subspaces of OS, in which each corresponds to a cluster of anchor concepts. We identify bridge concepts based on the rule that given any two anchor concepts in a cluster, the observability score of a bridge concept to any of the anchor concepts should be larger than the two anchor concepts themselves. Bridge concepts selected by this scheme are more specific to the context of the query.

For instance, car_on_road could be found in the subspace spanned by anchor concepts car and road. Apparently, by jointly utilizing anchor and bridge concepts, we can represent the queries more precisely.

### 3.2.2 Multi-level Concept Fusion

In addition to the $A$ and $B$ detector sets, we also introduce the positive ($P^+$) and negative ($N^-$) detector sets. The fusion of concept sets, i.e., $\{A, B, P^+, N^-, \}$, is conducted in a progressive manner with one set of detectors being fused one at a time, addressing different facets of detectors.

#### Reliability-based Fusion: Within OS, for each concept $C$, a set of positive correlated concept $P^+$ is formed by selecting concepts with higher observability scores to $C_i$. Similarly, a negative correlated concept set $N^-$ is generated containing the concepts with lower observability scores. Let $D(C)$ be a list of detector scores of concept $C$ on a test data set, $P^+$ and $N^-$ are used to refine the $D(C)$ as follows:

$$D(C) = D(C) + \sum_{C_i \in (P^+ \cup N^-)} \lambda(C, C_i) D(C_i)$$

where $\lambda(C, C_i)$ is the observability of $C$ and $C_i$ and the improved detector $D(C)$ is the refined detector score list of $C$. Note that the values of $\lambda(C, C_i)$ is positive if $C_i \in N^-$.

#### Observability-based Fusion: We linearly fuse the bridge concepts together with anchor concepts to enhance the observability of concept pairs in $A$.

$$D(A) = D(A) + \frac{1}{|N(A)|} \sum_{B_i \in N(A)} \lambda(B_i, A) \times D(B_i)$$

where $N(A)$ is the set of bridge concepts whose nearest anchor concept is $A$.

#### Semantic-based Fusion: The fusion of anchor concept detectors is performed directly in SS by employing $D(A)$ detectors. To answer the query, all observability-enhanced detector score lists of anchor concepts are linearly fused, weighted by the semantic relatedness of corresponding anchor concepts to current query.

#### Diversity-based Fusion: Weighting detectors individually in SS has the disadvantage that the diversity of detectors is not addressed during fusion. For instance, given the set of anchor concepts $A=\{person, face, police, newspaper\}$, the search results will be biased by the first three concepts re-

![TRECVID 2008 (Automatic Search)](image)

Figure 4: Comparison of our automatic search performance with all 56 type-A submissions in NIST TRECVID 2008. Our submission is shown in red and the rest are from other research teams.

### 3.3 Key Results

We has experimented the proposed approach on TRECVID 2005-2007 datasets and queries. The results presented in [3][5] confirm SS’s superiority over traditional ontology reasoning, while the results in [4] confirm the advantages of considering observability, reliability and diversity, in addition to the semantics of detectors to queries. Figure 4 compares the performance of our search method to other research teams such as IBM, Mircoosoft, CMU and so on. Even only utilizing visual information (other teams use text search to boost the results), our performance ranks at the 3rd places.

### 4. CONCLUSION AND DEMO

This paper summarizes our studies on concept-based video search. Based on the developed detector set VIREO-374, concept selection and fusion are conducted by reasoning on two novel spaces: SS and OS. Encouraging empirical performance, for both concept detection and automatic video search, are shown based on TRECVID 2008 benchmark. In the demo, we will showcase our research prototypes: VIREO-374 and concept-based video search engine.

### 5. REFERENCES


