1-2011

Concept-driven multi-modality fusion for video search

Xiao-Yong WEI

Yu-Gang JIANG

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Concept-Driven Multi-Modality Fusion for Video Search

Xiao-Yong Wei, *Member, IEEE*, Yu-Gang Jiang, Chong-Wah Ngo*, *Member, IEEE*

***Abstract*—As it is true for human perception that we gather information from different sources in natural and multi-modality forms, learning from multi-modalities has become an effective scheme for various information retrieval problems. In this paper, we propose a novel multi-modality fusion approach for video search, where the search modalities are derived from a diverse set of knowledge sources, such as text transcript from speech recognition, low-level visual features from video frames, and high-level semantic visual concepts from supervised learning. Since the effectiveness of each search modality greatly depends on specific user queries, prompt determination of the importance of a modality to a user query is a critical issue in multi-modality search. Our proposed approach, named concept-driven multi-modality fusion (CDMF), explores a large set of predefined semantic concepts for computing multi-modality fusion weights in a novel way. Specifically, in CDMF, we decompose the query-modality relationship into two components that are much easier to compute: query-concept relatedness and concept-modality relevancy. The former can be efficiently estimated online using semantic and visual mapping techniques, while the latter can be computed offline based on concept detection accuracy of each modality. Such a decomposition facilitates the need of *adaptive* learning of fusion weights for each user query on-the-fly, in contrast to the existing approaches which mostly adopted pre-defined query classes and/or modality weights. Experimental results on TRECVID 2005–2008 data sets validate the effectiveness of our approach, which outperforms the existing multi-modality fusion methods and achieves near-optimal performance (from *oracle fusion*) for many test queries.**

***Index Terms* — video search; multi-modality; concept-driven fusion; semantic concept.**

I. Introduction

ONE challenge of video search is the prediction of user search intention. The intention is often expressed using a short text description with several words, and/or a few visual examples of images and videos. A successful search system is therefore expected to adaptively formulate a search strategy in multi-modality forms, and eventually return a set of relevant video clips. Popularly used modalities include text search, visual search, and concept search. Text search tries to match the textual query words to video transcripts, while visual search measures the similarity between visual query examples and target videos. In concept-based search, a large number of semantic visual concept classifiers are constructed offline for indexing the video content, and efficient search is enabled by matching both textual and visual queries to the semantic concepts (e.g., [1], among others). Under this multi-modality search scenario, one key component of the strategy planning is the dynamic assignment of fusion weights to different search modalities according to a query.

In the literature, one popularly adopted search strategy is query-class dependent fusion [2–6], where the key idea is to map a user query into one of human-defined query classes. It is assumed that the optimal fusion weights of query classes are known and can be obtained through offline learning from similar queries. Therefore, the task of multi-modality fusion becomes to classify user query and then apply the learnt fusion weights for query answering. This strategy, nevertheless, encounters a number of practical and theoretical challenges. First, the collection of training queries is not a trivial process, and more importantly, the generation of ground-truth for learning optimal weights of query classes is an extremely time-consuming task. Second, it is unclear how the query classes should be defined. Previous research efforts in multimedia typically defined five to six query classes [2, 3] to model possible queries in the domain of broadcast news videos. It is difficult to evaluate how many human-defined classes are considered enough to cover the space of all possible user queries.

In this paper, we propose a novel query-adaptive fusion strategy, by mapping a multi-modality query to the large number of semantic concepts instead of a query-class, and harness the selected concepts to determine the fusion weights on-the-fly. In other words, the fusion problem is decomposed into two major stages: reasoning query-concept relatedness, and learning query-modality relevance through the selected concepts. Figure 1 illustrates the flow of our proposed approach. Given a query which contains a short text description and a few visual examples, query-to-concept mapping is firstly conducted to infer the set of semantically and visually relevant semantic concepts. In the example shown in Figure 1, concepts such as "airplane" and "sky" are reasoned through the text query, while concepts such as "flying objects" and "cloud" are inferred from the visual examples. The selected concepts are then mapped into a context graph, which is offline built to characterize the co-occurrence relations among the entire concept set. By further conducting random walk over the graph, the interaction among the concepts is modeled, and thus the relevancies of the concepts to the query can be refined. For instance, the concept "airplane flying" is discovered in

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 119709) and the National Natural Science Foundation of China (No. 61001148).

The authors are all with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong (e-mail: {xiaoyong, cwngo}@cs.cityu.edu.hk). Xiao-Yong Wei is also with the School of Computer Science, Sichuan University, Chengdu 610054, China and Yu-Gang Jiang is currently with the Department of Electrical Engineering, Columbia University, New York, NY 10027, USA (e-mail: yjiang@ee.columbia.edu).

* Corresponding author.

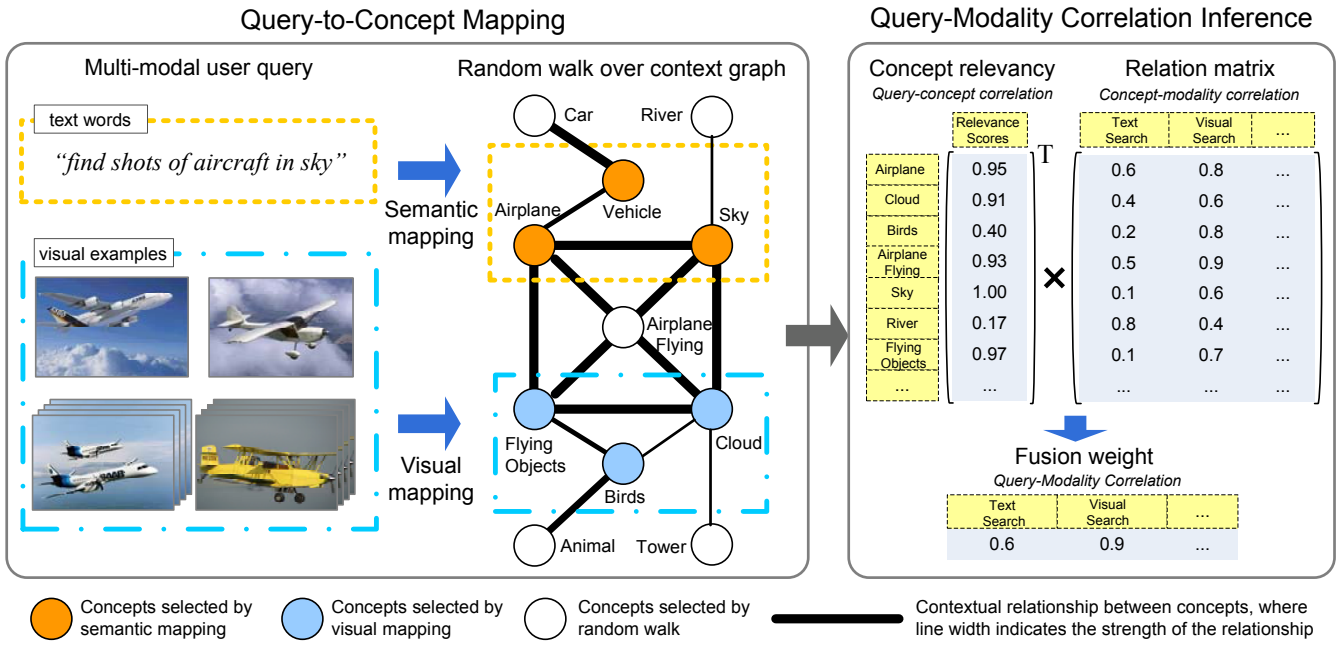


Fig. 1. Modality weight computation in CDMF. Given a user query "find shots of aircraft in sky" with a few image and video examples, relevant concepts are firstly selected with semantic and visual mapping. The selected concepts are then refined using random walk over a context graph (cf. Section IV-B). Finally, the refined concept set (query-concept correlation) is harnessed to infer query-modality relationship.

random walk and assigned with a relevance score of 0.93 to the user query, since it frequently co-occurs with many of the selected concepts during query-to-concept mapping. After concept selection, in the second stage, the concept relevancies are converted to fusion weights, through fuzzy transformation with a relation matrix. The matrix outlines the association between concepts and modalities, and can be offline learnt from the concept detection accuracy from each modality. For concept "airplane flying", the matrix indicates visual search is more reliable than text search. By considering the association between query-concept and concept-modality, concepts, collectively as a bridge to query-modality, are exploited to infer fusion weights.

Due to the fact that our proposed fusion strategy is driven by the semantic concepts, we name it as Concept-Driven Multi-modality Fusion (CDMF). Compared to existing fusion strategies, CDMF offers the following advantages:

- *Generalizability*: Concepts are easier to be identified and defined than query classes. Using Large-Scale Concept Ontology for Multimedia[1] (LSCOM) [23] as an example, there are 1000+ concepts being identified for broadcast video domain. Pair-wise combination of any two among the 1000+ known concepts can already answer half million queries arbitrarily generated. In Figure 1, for instance, using two concepts "airplane" and "sky", the query "find shots of aircraft in sky" can be fairly well interpreted. In contrast, predicting the query using a human-defined query class (e.g., an Object-X class) can provide only coarse estimation of search intention [2–4]. In general, the number of concepts can be much more than that of the human-defined query classes, thus offering greater capability to make the search process more flexible and generic.
- *Adaptivity*: In CDMF, the fusion weights are computed on-the-fly and adapted from query to query. This is very different from other strategies such as query-class dependent fusion where the weights, once trained, are "fixed" for each query-class. In contrast, CDMF offers a more flexible means in combining the search modalities since the fusion weights are determined by the concepts dynamically selected based on query specifications. Prelearning of "optimal" fusion weights for a query-class, on the other hand, is practically difficult, especially if the defined query class is too general such as the Object-X query class. Under this scenario, using fixed fusion weights for all queries routed to the class will limit the search performance.
- *Extensibility*: Collecting concepts and their ground-truth annotations for learning is in general more feasible than manually collecting and annotating examples of training queries. Indeed, there have been a number of large-scale or web-scale concept detectors set made publicly available. Examples include Columbia374 [7], VIREO-374 [8], MediaMill [9], and NUS-WIDE [10]. Direct utilization of these detection sets for predicting fusion weights is feasible, especially when these detector sets often come along with classifiers and their detection result. In CDMF, the correlation among concept labels is exploited for constructing a context graph, while the detection performance is used to compute concept-modality relevancy. Therefore, compared to query-class

[1]LSCOM is considered as the largest concept definition and annotation collection. It includes manually judged labels for 1000+ visual concepts and 100+ queries over 61,901 video shots of multilingual broadcast news videos.

dependent fusion, CMDF requires less additional effort. CDMF can be naturally extended to any new concept and new modality made available, respectively, by adding a new row and column to the relation matrix shown in Figure 1.

One important contribution of this paper is the proposal of using semantic concepts for multi-modality fusion. To the best of our knowledge, this is the first work on harnessing semantic concepts for computing query adaptive fusion weights. The remainder of this paper is organized as follows. In Section II, we review the existing works in multi-modality video search. Section III introduces our query-to-concept mapping method, and Section IV further refines the mapping by modeling query context. Section V describes the proposed multi-modality fusion strategy. The experimental results and performance comparison are given in Section VI. Finally, Section VII concludes this paper.

II. Related Work

Multi-modality fusion has been a long standing issue for effective video search. Most existing works linearly combined different search modalities due to its simplicity and better generalizability. Based on weight computing strategy, we roughly divide them into three main categories: heuristic fusion, query-time fusion, and supervised query-class dependent fusion.

Heuristic fusion is a widely employed scheme. The fusion weights are usually rule-based and predefined. For instance, in [11], weights are empirically pre-determined for each modality based on the types of query terms received. If only named entities are found in a query, text search will be trusted and thus given higher weights than other search modalities. Similarly, if concept names are found in the query, concept-based search will be given higher priority. For most other cases, weights are equally set for all search modalities. Despite of its simplicity, as demonstrated by [11], this scheme has shown satisfactory performance on the TRECVID[2] evaluation [12].

Different from setting predefined weights, query-time fusion dynamically decides fusion weights on-the-fly during query time. In [13, 14], weights are directly derived for each query based on the ranked lists returned from different search modalities. The idea is based on the hypothesis that the shape of a curve depicting the score distribution of a search modality, will give clue to the applicability of the modality to the query. Specifically, rapid change of retrieval scores in the top part of a search ranking list indicates the ability of a modality in distinguishing relevant search items from one another. Conversely, gradual change in initial ranking gives clue that most items are similar to each other and the modality is incapable of making clear decision. Based on this intuition, the approach in [14] derives the fusion weight of a modality by computing the ratio of MAD (mean average distance) between the top 5% retrieved items and the remaining 95% of items. This fusion scheme has also been demonstrated by [15] and shown good search performance in recent TRECVID evaluations.

Supervised query-class dependent fusion, different from the previous two schemes, estimates fusion weights using training examples [2–6]. A survey of recent advances along this direction can be found in [16]. First, a set of query classes is pre-defined to categorize the types of possible queries that can be input by users. Second, optimal fusion weights are computed respectively for each query class by learning from a set of example queries with ground-truth. During query time, a query is routed to one of classes and the predicted "optimal" weights are applied to fuse multiple search modalities. Due to the consideration of query classes, this scheme is in general more reliable than the linear fusion scheme. Nevertheless, it does suffer from several limitations. Generalization, in practice, is a problem since the number of pre-defined query classes is limited and therefore it becomes difficult to characterize all possible types of queries. Furthermore, learning of optimal weights could be challenging especially when the training queries are not representative enough for each query class.

To cope with the aforementioned issues, automatic discovery of query classes from training examples is studied in [4][6]. In [4], query classes are generated by clustering training queries based on both query text relatedness and search performance consistency of various modalities. In [5], pLQA (probabilistic latent query analysis) is proposed for class discovery – a user query is mapped into a mixture of several query classes for search. In [6], the idea of online query class generation is proposed. During query time, a query class is formed dynamically through searching similar training queries from database. Fusion weights are then determined on-the-fly by learning from the set of similar queries. Nevertheless, despite either the query classes are defined manually or automatically, these existing approaches are ineffective in dealing with rare queries which are difficult to be categorized to any class. In addition, they also rely on a large set of pre-defined training queries, which are not easy to acquire in practice.

In this paper, we address the problem of multi-modality fusion using a concept-driven paradigm. Concept detection performance with different features is utilized in our approach to infer modality weights adaptively for each query. Compared to existing modality weighting methods such as the query-class dependent fusion, our approach is more flexible (and more accurate, as will be validated in the experiments) to compute the fusion weights.

Our approach involves techniques for query-to-concept mapping and context modeling. In the literature, there have been a number of approaches for query-to-concept mapping [1, 17–20]. Most works are devoted to the semantic analysis of text queries for measuring the relatedness between queries and concepts. Mapping concepts using visual query examples has also been investigated in [1, 18]. Our concept selection method differs from these query-to-concept mapping techniques mainly in that a random walk process is further imposed to utilize concept relationship (context) for a globally more consistent selection. The context, on the other hand, has been

[2]TRECVID (TREC Video Retrieval Evaluation) is an annual video retrieval evaluation activity supported by US National Institute of Standards and Technology. Each year a new dataset, as well as ground-truth annotation for a couple of concepts and queries are provided for system evaluation of several video retrieval tasks.

utilized in several existing works for a set of related issues in multimedia search and indexing [29–34]. In [29], document-level context is exploited for leveraging the recurrent patterns among video stories to improve initial video search results from text matching. In [31, 33, 34], statistical models of inter-concept correlation are built for helping improve the quality of concept-level video indexing. Different from these existing works, context is exploited in our approach for better concept selection, not for improving video indexing performance [31, 33, 34] or reranking search results [29].

III. Query-to-Concept Mapping

A. Semantic Mapping

Semantic mapping aims to find a set of concepts that have the highest linguistic relatedness to the text queries. To measure such relatedness, we adopt our previous work in [17] by first building a semantic space (SS) and then performing concept reasoning. SS is an orthogonal linear space encapsulating the semantic relationship (mainly *is-a* relation) among text words (query terms or concepts in our case). The relationship is learnt from WordNet using an ontology-based measure named WUP [21]. In SS, a concept or word, when projected to this space, is represented as a vector. The semantic similarity (the quantitative linguistic relatedness) between two concepts C_i and C_j is measured with cosine similarity as:

$$Semantic(C_i, C_j) = \frac{S_{C_i} \cdot S_{C_j}}{|S_{C_i}||S_{C_j}|} \quad (1)$$

where S_{C_i} and S_{C_j} are the feature vectors of C_i and C_j in SS respectively. Different from conventional ontology-based measures, SS is a computable space and provides a global view of concepts in determining semantic relatedness between concepts. For details, readers can refer to [17].

Denote $\mathcal{Q} = \{q_1, q_2, \ldots\}$ as a text query, and $\mathcal{V} = \{C_1, C_2, ...\}$ as a concept set. Through projecting the query terms in $\mathcal{Q}$ into the SS, the relatedness between a text term $q_i \in \mathcal{Q}$ and a concept $C_i \in \mathcal{V}$ can be computed by Eqn (1). By considering all the query terms and selecting one concept with the highest relatedness for each term, we have

$$\mathcal{S} = \bigcup_{q_i \in \mathcal{Q}} \underset{C_j \in \mathcal{V}}{\operatorname{argmax}} \{Semantic(q_i, C_j)\} \quad (2)$$

where $\mathcal{S}$ is the set of semantic concepts selected for query answering, and $Semantic(q_i, C_j)$, simply denoted as $Semantic(C_j)$ in later discussions, is adopted as the semantic similarity of concept C_j to query $\mathcal{Q}$. In Eqn (2), the mapping from words to concepts is performed on the basis of one-to-one. In other words, the number of chosen concepts is at most equal to the amount of words in $\mathcal{Q}$, i.e., $|\mathcal{S}| \leq |Q|$. In addition, we only consider nouns and gerunds of a query, assuming that noun mostly indicates the name of place, thing or person, and gerund describes an action or event (e.g., *walking* and *running*).

B. Visual Mapping

Different from text-based semantic relatedness, visual mapping considers the selection of concepts by investigating visual query examples. The intuition is that, by surveying the presence (absence) of concepts in the positive (negative) query examples, concepts are picked based on their prevalence and discriminativeness. We adopt concept detectors to predict the presence of concepts to visual queries. Given a query, a concept is considered prevalent if its detector outputs consistently high detection scores on most of the visual examples (i.e., positive examples to the query). In addition, we randomly draw samples from testing set as pseudo-negative examples, assuming that the majority of samples in the testing set are not relevant to the query. A concept is thus considered discriminative if its detection scores exhibits distinguishable distributions between positive and negative examples. Therefore, a concept relevant to a query should be prevalent and discriminative.

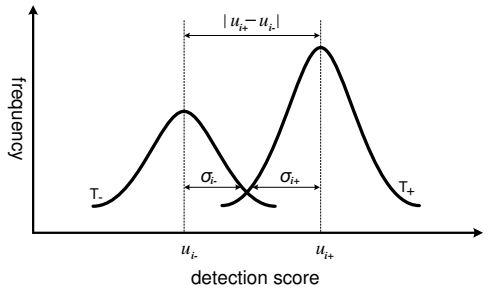


Fig. 2. Illustration of signed Fisher ratio.

Denote $\{u_{i+}, \sigma_{i+}\}$ as the mean and standard deviation of prediction scores from a concept detector on the positive examples. Similarly, $\{u_{i-}, \sigma_{i-}\}$ are those on the negative training examples. We propose a signed Fisher ratio (SFR) to measure the relevance of C_i to query as

$$Visual(C_i) = sign(u_{i+} - u_{i-}) \cdot \frac{(u_{i+} - u_{i-})^2}{\sigma_{i+}^2 + \sigma_{i-}^2} \quad (3)$$

where the sign function contrasts the prevalence of C_i in positive and negative samples. Positive value will be assigned, indicating the usefulness of C_i to visual query, if C_i receives higher prediction scores on average in positive than in negative samples. The discriminativeness of a concept is further determined by the second part of the equation, which is the original formula of Fisher ratio, for measuring class separability. With reference to Figure 2, SFR computes the relevance of a concept by investigating the distribution of prediction scores in positive (T_+) and negative (T_-) samples. The farther apart the centers of the two Gaussian distributions are, the larger the relevance of a concept is to a query. The wider the distribution spreads out, the less trustful a concept is for its fluctuating performance. By SFR, all concepts are eventually ranked based on their relevance. We consider the top-k most relevant concepts where the value of k is empirically set equal to the number of concepts selected by semantic mapping. This aims to prevent the case that the selected concepts by visual mapping overwhelm those selected from semantic mapping.

The idea of SFR is similar to DBCS (Distribution-based Concept Selection) recently proposed in [18]. One major difference is that DBCS does not take into account the second moment (or deviation) of distribution. This may result in unreliable prediction due to the use of the randomly picked

pseudo-negative samples. Consequently, DBCS is sensitive to the selection of negative samples, and less stable if compared to SFR.

IV. Modeling Query Context

The concepts inferred through semantic and visual mapping carry isolated pieces of signal, either semantically or visually, and there is basically no interaction between them. Nevertheless, concepts do not exist in isolation. For instance, given a query "*find shots of car and pedestrian*", the concepts by semantic mapping may include *car* and *crowd*, while the concepts by visual mapping may be composed of *building*, *road* or even *bridge* and *flag*. How to utilize these concepts by assigning appropriate weights to weigh their importance, in general, can depend on the interaction among them. Concepts such as *car* and *road*, *crowd* and *building* are contextually related, and thus should play major roles in query answering. For concepts *bridge* and *flag*, which may be wrongly selected due to prediction error, could be assigned less weight if knowing that their correlations with the other concepts are weak. Furthermore, other concepts such as *street* and *traffic* may be further selected if their co-occurrence relationship with the initially selected concepts (e.g., *car*, *road*) is known a priori.

In this section, context graph is built by offline learning the context relationship among concepts. For a given query, the two sets of concepts selected by semantic and visual mapping are represented as nodes in the graph. By assigning initial weights to the corresponding nodes, random walk is then performed to iteratively propagate their weights across other concepts (nodes) in the graph. The final set of concepts, along with their refined weights after random walk, is utilized for multi-modality fusion. For convenience, we term the selected concepts by text queries as sMap concepts, and those selected by visual query examples as vMap concepts.

A. Building Context Graph

Context graph, denoted as $\mathcal{G}$, is an undirected graph with concepts as nodes. To construct $\mathcal{G}$, a context space is built to model the contextual relationship among concepts. The space is leant by firstly measuring the correlation (co-occurrence) in ground-truth labels of the concepts through Pearson product moment (PM). The pair-wise concept correlations are then refined into globally consistent contextual similarity scores based on our recent work [22]. In the context space, a concept can be represented by a vector, where each entry is the PM value to a reference concept (a basis of the space). Similar to semantic space, context space is a linear space. Under this space, contextual similarity between two concepts C_i and C_j can be measured with

$$Context(C_i, C_j) = \frac{V_{C_i} \cdot V_{C_j}}{|V_{C_i}||V_{C_j}|} \tag{4}$$

where V_{C_i} and V_{C_j} are the concept vectors of C_i and C_j in the context space respectively. Different from PM which measures correlation locally based on two concepts only, the context space provides a global view of concept correlation, by also considering the contextual similarity of concepts with respect to basis vectors forming the space. Reader can refer to [22] for more details.

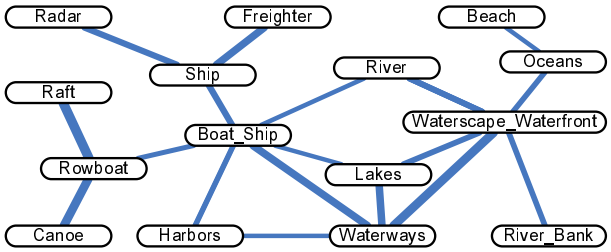


Fig. 3. Water-related concepts: a partial view of the context graph. Edge width indicates the strength of contextual relationship. Only edges with contextual similarities larger than 0.3 are shown.

Using Eqn (4), concepts are organized to form the context graph $\mathcal{G}$, through establishing edges for any two concepts with contextual similarities larger than zero. Edge weights are set equal to corresponding similarities for modeling the contextual closeness among concepts. Since the measure in Eqn (4) is symmetric, $\mathcal{G}$ is an undirected graph. Figure 3 shows a partial view of $\mathcal{G}$ constructed from LSCOM concepts [23]. It shows how water related concepts in LSCOM are connected in $\mathcal{G}$. For instance, the concept *boat_ship* has direct links with concepts *lakes*, *harbors*, *waterways*, and indirect links to concepts such as *raft* and *ocean*.

B. Random Walk

Given a context graph $\mathcal{G}$ of n nodes (or concepts), the random walk process [24] is modeled as

$$\mathcal{W} = \{\mathcal{G}, P, x_{(\pi)}\} \tag{5}$$

where $P = [p_{ij}]_{n \times n}$ is the transition matrix, and x_π is a column vector encapsulating the stationary probabilities of the concepts at given state π. The transition probability p_{ij} between two concepts C_i and C_j indicates the probability of reaching C_j from C_i. We set p_{ij} as:

$$p_{ij} = \frac{Context(C_i, C_j)}{\sum_{C_t \in \mathcal{V} \setminus C_j} Context(C_t, C_j)} \tag{6}$$

which is the context similarity between C_i and C_j, normalized by the sum of similarities from all concepts which are incident to C_j.

The stationary probabilities in $\mathcal{W}$ are initialized based on the set of sMap and vMap concepts selected by a given query. Let C_j be a concept selected, the initial weight of C_j is set according to Eqn (1) and Eqn (3) as following

$$x_{(0)}(j) = \max\{Semantic(C_j), Visual(C_j)\} \tag{7}$$

For concepts neither semantically nor visually selected, their weights are initialized to zero. During random walk, the stationary probability of a concept C_j is iteratively updated. At time instance k, the update can be expressed as

$$x_{(k)}(j) = \alpha \sum_{i \neq j} x_{(k-1)}(i) p_{ij} + (1 - \alpha) x_{(0)}(j) \tag{8}$$

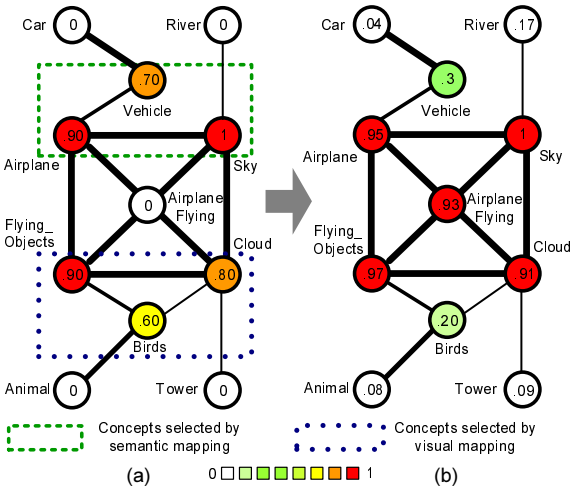


Fig. 4. Illustration of random walk over the context graph for query "*find aircraft in sky*": (a) at the initial state; (b) at the state of convergence. Edge width indicates the strength of concept relationship in the context space.

where $\alpha \in [0, 1]$ is a parameter to control the speed of convergence. Eqn (8) has two components: one part is information fusion from neighboring concepts, and the other is the initial probability. Therefore, α is also a parameter that balances the contribution of the two parts. Eqn (8) can be executed iteratively, until meeting the convergence condition of $|x_{(k+1)} - x_{(k)}| \rightarrow 0$.

The stationary probabilities of concepts eventually form a vector $\mathcal{P} = [p_1, \ldots, p_n]$, with p_i indicating the relevance of a concept C_i to the query. Note that the non-zero entries in $\mathcal{P}$ include not only sMap and vMap concepts, but also other concepts which are reached during the random walk. Figure 4 illustrates an example for the query "*find aircraft in sky*". Two groups of concepts are selected by semantic and visual mapping respectively. Their similarities with respect to the query are mapped to the graph nodes as initial stationary probability on the context graph, as shown in Figure 4(a). After random walk, five more concepts are inferred, each carrying different probabilities. Among them, the concept *airplane_flying*, which is strongly and consistently connected to four other concepts, receives higher probability after a few iterations. Erroneous concepts such as *car* and *river*, even though may be contextually related to the query, are also inferred but with much lower probabilities. In addition, the original weights of concepts such as *airplane* and *cloud* are boosted due to their strong contextual interaction with other relevant concepts. On the contrary, the relevancy of concepts such as *vehicle* and *birds* are reduced eventually due to their weak interaction with the other concepts. In brief, grounding on the context relationship provided by context graph and the initial scores of concepts by query mapping, random walk is effective in amending the relevance of concepts based on information exchange and propagation.

V. Concept-Driven Multi-Modality Fusion

While the selected concepts carry semantics of a query, not all the concept detectors are reliable for video search. On the other hand, different search modalities (expert) may work well for different types of queries. In the following we first detail the inference of modality weights for concepts, and then present the formulation of concept driven fusion.

Let $\mathcal{V} = [C_1, C_2, \ldots, C_n]$ as the order set with n concepts, and $\mathcal{M} = [M_1, M_2, \ldots, M_m]$ as the order set with m modalities. The fuzzy transformation $T_{\mathcal{R}}$ from $\mathcal{V}$ to $\mathcal{M}$ can be described as

$$T_{\mathcal{R}} : \mathscr{F}(\mathcal{V}) \longrightarrow \mathscr{F}(\mathcal{M}) \tag{9}$$

where $\mathscr{F}$ denotes a fuzzy set. A fuzzy set $\mathscr{F}(\mathcal{S})$ can be simply explained as a membership vector defined on a classical set $\mathcal{S}$, where each dimension of the vector indicates a degree of membership (valued in the interval of $[0, 1]$) to corresponding element in $\mathcal{S}$. Therefore, $\mathscr{F}(\mathcal{V})$ indicates the relationship of query $\mathcal{Q}$ to $\mathcal{V}$, while $\mathscr{F}(\mathcal{M})$ indicates the usefulness of each modality in $\mathcal{M}$ for $\mathcal{Q}$. To model the fuzzy relationship, a relation matrix $\mathcal{R} \in \mathscr{F}(\mathcal{V} \times \mathcal{M})$, which describes the relationship between concepts and modalities, is defined as

$$\mathcal{R} = \begin{bmatrix} \mathbf{r}_{11} & \mathbf{r}_{12} & \mathbf{r}_{13} & \ldots & \mathbf{r}_{1m} \\ & & & \ldots & \\ & & & \ldots & \\ \mathbf{r}_{n1} & \mathbf{r}_{n2} & \mathbf{r}_{n3} & \ldots & \mathbf{r}_{nm} \end{bmatrix} \tag{10}$$

where $\mathbf{r}_{ij}$ specifies the retrievability of concept i using modality j. The relation matrix $\mathcal{R}$ can be estimated with training or subjective pairwise evaluation. We adopt the first scheme and the details will be discussed in the experiment section (Section VI-C).

Combining with the set of concepts inferred from semantic and visual mapping, Eqn (9) becomes,

$$T_{\mathcal{R}}(\mathcal{P}) = \mathcal{P} \circ \mathcal{R} \in \mathscr{F}(\mathcal{M}) \tag{11}$$

where $\circ$ is a fuzzy composition operation. Given a query $\mathcal{Q}$, the transformation $\mathscr{F}(\mathcal{M})$ can be represented as a vector $\mathcal{W} = [w_1, \ldots, w_m]$ with w_i specifying the weight (importance) of the i^{th} modality to $\mathcal{Q}$. There are many implementations for the composition operation. We adopt matrix multiplication for simplicity. Therefore, Eqn (11) can be rewritten as:

$$\mathcal{W} = [w_j]_{1\times m} = [\sum_{i=1}^{|\mathcal{V}|} p_i \cdot \mathbf{r}_{ij}]_{1\times m} \tag{12}$$

With Eqn (12), the idea of concept-driven multi-modality fusion becomes intuitive. The importance of a modality M_j is jointly determined by the relevancy p_i of concept C_i and r_{ij} which indicates the retrievability of concept C_i using M_j. The significance of M_j towards a query $\mathcal{Q}$, reflected through w_j of Eqn (12), is accumulated from all the selected concepts. Eventually, the vector $\mathcal{W}$ is directly employed for multi-modality fusion. While equations (11) and (12) are based on linear fusion, the main novelty lies in how the weights are derived based on the selection of suitable concepts for each query and the computation of the concept-modality relation matrix.

Compared to query-class dependent fusion, concept-driven fusion offers two advantages. First, the relation matrix $\mathcal{R}$ in Eqn (10) can be learned through reusing training samples

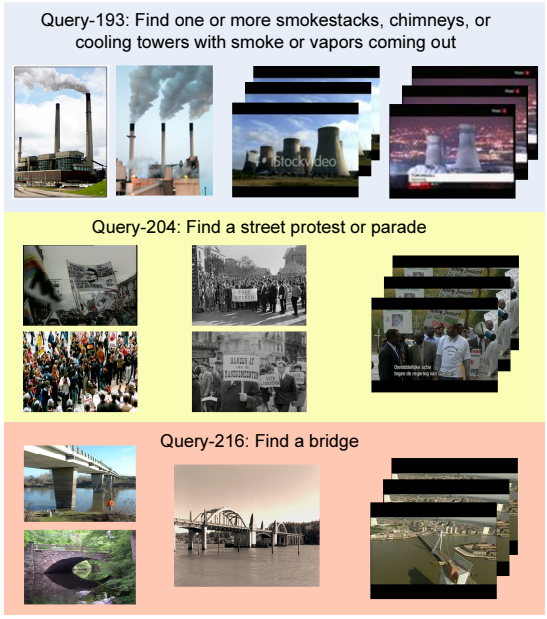


Fig. 5. Examples of textual and visual queries.

of the concepts (details in Section VI-C). This is much less time-consuming than collecting and annotating a new set of queries for training. Second, our matrix $\mathcal{R}$ can be trained incrementally. When new search modality is available, only an additional column of $\mathcal{R}$ needs to be added, without the update of other entries in $\mathcal{R}$. This is in contrast to query-class dependent fusion, where the weights of all modalities need to be re-computed from scratch for each update.

VI. Experiment

A. Dataset and Evaluation

We conduct experiments using four datasets: TV05, TV06, TV07 and TV08, from TRECVID annual evaluations in years 2005 to 2008 respectively [12]. TV05 and TV06 are composed of broadcast news videos in English, Chinese and Arabic. There are 85 hours (45,765 shots) and 80 hours (79,484 shots) of testing videos in TV05 and TV06 respectively. TV07 and TV08 are Dutch videos from the Netherlands Institute for Sound and Vision, containing mainly documentary videos of 50 hours (18,142 shots) and 100 hours (33,726 shots) respectively in the testing sets. Each dataset of TV05–TV07 comes along with 24 search queries, while TV08 contains 48 queries. The queries are given in multiple forms of texts, images and/or video clips. Figure 5 shows a few example queries[3]. Text queries are generally short with an average of 3.6 meaningful words (e.g., nouns and gerunds). Each text query is also accompanied with 3.3 image examples and 2.1 video clips on average. We extract keyframes from each query video clip and treat them equivalently as the query images. In total, we have 15.4 images on average for each query.

In the experiments, the retrieved items (video shots) are ranked according to their computed scores to the queries.

[3]Query IDs are defined by TRECVID, which are available at http://www-nlpir.nist.gov/projects/tv2009/old.topics.features.txt.

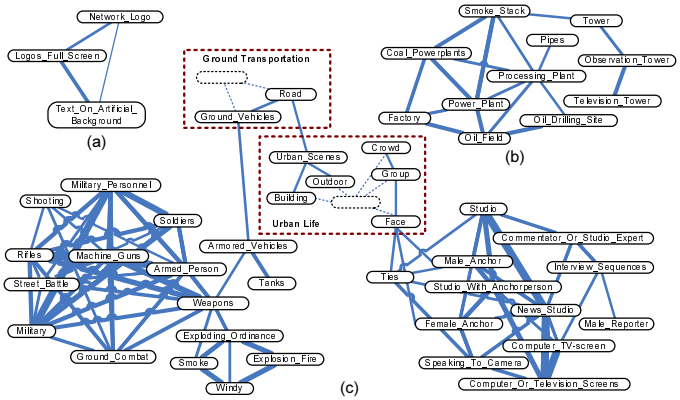


Fig. 6. A partial view of the context graph $\mathcal{G}$. Edge width indicates concept affinity. Only edges with contextual similarities larger than 0.3 are drawn.

Following the TRECVID evaluation, we use average precision (AP) to evaluate the results of ranked lists on TV05-TV07 and inferred average precision (InfAP) on TV08. AP approximates the area under precision-recall curve, while InfAP estimates the traditional AP when the testing data sets are partially labeled [25]. To aggregate the performance over multiple concepts, we use mean AP for TV05-TV07 and mean InfAP (MAP) for TV08.

B. Construction of Context Graph

As introduced in Section IV-A, prior to the construction of the context graph $\mathcal{G}$, a context space is learnt by using LSCOM (Large-Scale Concept Ontology for Multimedia) concepts [23]. LSCOM includes 834 concepts and their annotations on development dataset of TRECVID 2005. A total of 374 concepts, each has more than 10 positive training examples, are selected for learning the context space [22]. With this space, the context graph $\mathcal{G}$ is formed by connecting concepts which are contextually related based on Eqn (4). Under this formulation, all the 374 concepts are connected as a single graph, and they are reachable to each other by traversing the edges in $\mathcal{G}$.

While all concepts are directly or indirectly connected, by manual browsing we find that there are as high as 196 "clusters" of concepts in $\mathcal{G}$. Concepts in each cluster are basically tightly connected. Among these clusters, there are 192 small-size clusters each including only 1 to 4 concepts, 3 middle-size clusters each including 5 to 15 concepts, and a large-size cluster containing a total of 93 concepts. Figure 6 shows three examples of the clusters in $\mathcal{G}$. Concepts in a small cluster normally belong to rare or specific concepts (e.g., *network_logo* and *text_on_artificial_background*), as shown in Figure 6(a). Concepts in a median cluster are more general. For instance, in Figure 6(b), industry-related concepts such as *power_plant* and *oil_drilling_site* are closely connected to each other. The water-related cluster shown in Figure 3 is also another example. The largest cluster found in $\mathcal{G}$ is shown in Figure 6(c). This cluster includes mainly person-related concepts, and can be roughly divided into four sub-clusters: military affairs, ground transportation, news studio, and urban

life related concepts. This cluster locates at the center of $\mathcal{G}$ and directly connects to all other clusters.

The graph $\mathcal{G}$ models the context relationship among concepts reasonably well, and mostly consistent with human perception. Each cluster connects object, scene and/or action related concepts, which frequently co-occur. For instance, referring to the sub-clusters on the left hand side of Figure 6(c), the included concepts are *soldiers*, *street_battle*, *shooting* and *smoke* under the context of war. At the border of the cluster, concepts such as *armored_vechicles* are further connected to ground transportation related concepts. Further away by traversing the edges, clusters of concepts about news studio are found, as shown on the right hand side of Figure 6(c).

C. Learning Concept-Modality Association

The proposed approach involves the learning of the relation matrix $\mathcal{R}$ in Eqn (10). We estimate each entity $\mathbf{r}_{ij}$ in $\mathcal{R}$ based on the retrievability of concept C_i using corresponding search modality $\mathcal{M}_j$. In our experiments, we consider three modalities: text-only, visual-only, and concept-only search. Detailed descriptions of the three modalities will be given in the next section. To compute the retrievability of a modality for a concept C_i, we treat the concept as a *simulated query* by using concept name as *text query* and ten randomly chosen positive samples as *visual query examples*. We then evaluate search performance of this simulated query using the three modalities over a validation set in TV05 (a subset of its development set). The performance essentially indicates the retrievability of C_i using different modalities, and therefore can be used in matrix $\mathcal{R}$. For example, the performances for concept C_{51} (*airport*) are 0.0042, 0.4524, and 0.7743 respectively from text, visual, and concept search. Larger value reflects higher search reliability of corresponding modality. This score vector [0.0042, 0.4524, 0.7743] then becomes a row $\mathbf{r}_i$ in the matrix $\mathcal{R}$. After computing simulated search performances of all the 374 concepts, we have $\mathcal{R} \in \mathbb{R}^{374\times 3}$ fully computed, which will be used throughout this paper.

D. Single Modality Search

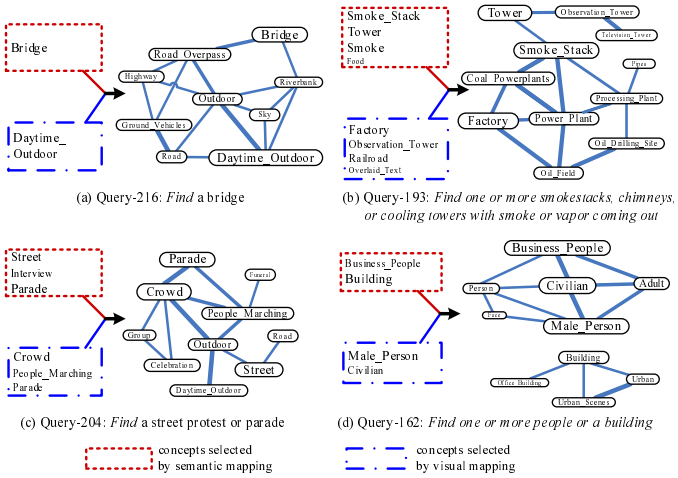
This section compares search performance of text-only, visual-only, and concept-only searches. In particular, we contrast the use of sMap and vMap concepts for search. The aim is to verify the set of selected concepts, after undergone context modeling, is effective in predicting search intention. This will also in turn justify the validity of these concepts for multi-modality fusion. The experimental setup of each single modality search is as following:

- *Text-only search* is performed by matching the text queries against the speech transcripts of the video shots. The text search is implemented using the popularly adopted Lemur system from CMU [26].
- *Visual-only search* considers the visual examples of the queries. These examples include images and/or short video clips. We adopt the supervised learning approach in [27], by learning ten SVMs for each query. Visual examples are used as the positive training samples for all the SVMs. Another ten sets of pseudo-negative examples are randomly sampled from the dataset and used separately for each SVM. The visual features for learning SVMs are concept scores output by 374 concept detectors from VIREO-374 [8]. In other words, each sample is represented by a 374-d feature vector, where each element is a probability indicating the confidence of detecting the corresponding concept in the sample. Finally, with the prediction outputs from the ten SVMs, we adopt average fusion to combine the results for ranking the video shots, and the ranked shots are used as visual-only search result.
- *Concept-only search* is performed by mapping text queries and/or visual queries against the concepts detected in video shots. Detection outputs of the selected concepts are then linearly fused as concept-only search result. The fusion weight is determined from the query-to-concept mapping, i.e., Eqn (1) for semantic mapping, Eqn (3) for visual mapping. Similar to visual-only search, VIREO-374 is employed for concept detection. To evaluate the contributions of each component described in this paper, here we conduct four runs of concept-only search under different settings: 1) `sMap` only uses text queries for concept selection; 2) `vMap` which uses visual query examples only; 3) `svMap` takes a union of the concepts selected by both sMap and vMap, and the result is indeed an average fusion of those by `sMap` and `vMap`; and 4) `svMap-RW` further applies random walk to refine the concept set of `svMap`, and the final stationary probabilities on concepts are used as fusion weights. Throughout the experiments, the parameter α in Eqn (8) is empirically set to 0.8.

TABLE I
SINGLE MODALITY SEARCH PERFORMANCE.

TV-	Text	Visual	Concept			
			sMap	vMap	svMap	svMap-RW
05	.059	.010	.123	.065	.098	.128
06	.026	.016	.044	.021	.047	.048
07	.004	.038	.029	.020	.030	.039
08	.008	.035	.039	.020	.035	.046

Table I shows the performances of the single-modality runs on the four TRECVID datasets. Among the three search modalities, concept-only search exhibits the best performance. This confirms the effectiveness of using semantic concept detectors for video search, which was also observed in several previous works [11, 15]. For the concept-based runs, we have the following major observations. First, selecting concepts based on text query alone (`sMap`) tends to be better than that using visual query examples (`vMap`). This confirms the power of textual queries in reflecting the query semantics, as was also observed by several previous works [19, 32]. Second, directly merging concepts selected by both `sMap` and `vMap` (i.e., `svMap`) does not show clear performance gain, which indicates that the concepts selected by `vMap` are indeed noisy. However, this does not mean `vMap` is not useful at all. In fact many concepts selected in `vMap` is complementary to that from `sMap` – among the 120 test queries, only 9 of them have overlap in concepts selected by both methods. The



(a) Query-216: *Find* a bridge

(b) Query-193: *Find one or more smokestacks, chimneys, or cooling towers with smoke or vapor coming out*

(c) Query-204: *Find* a street protest or parade

(d) Query-162: *Find one or more people or a building*

Fig. 7. Examples of query-to-concept mapping. The font size of a concept name indicates its relevance score to the query. The edge width indicates the strength of contextual relationship among concepts. Only edges with contextual similarities larger than 0.3 are drawn.

usefulness of vMap is evidenced by further applying random walk to refine the selection. We observed that the performance of svMap−RW is higher than that of sMap−RW. This is due to the fact that, in the random walk process, the importance of relevant concepts is always boosted, while insignificant (noisy) concepts, which reside in different clusters of the context graph from the majority concepts, are either excluded or given much lower weights.

We further analyze different runs of concept-only search using four example queries, as shown in Figure 7. In 7(a), semantic and visual mapping infer two sets of different concepts respectively. By random walk, more concepts are mined from context graph for fusion. In 7(b), semantic mapping selects concepts such as *smoke_stack*, *tower* and *smoke* which are semantically related to text queries, while visual mapping chooses concepts such as *factory* and *observation_tower* which are visually similar to the "chimneys" and "cooling towers" appeared in the visual queries. By random walk, potentially more useful concepts, but selected by neither semantic nor visual mapping, such as *coal_powerplants* are further mined. On the other hand, wrongly selected and insignificance concepts such as *food* and *overlaid_text* are "excluded" by assigning lower weights (stationary probabilities) after random walk. Similarly for the query in 7(c), the concept *interview* is successfully excluded after random walk. Therefore, random walk indeed plays an important role to alleviate the ambiguity due to semantic and visual mapping. For instance, due to out-of-vocabulary problem, the best match for the query term *vapor* in 7(b) is the concept *food* by semantic mapping. The concept *overlaid_text* is always selected by visual mapping due to the existence of text caption in visual queries. Random walk is effective in pruning or assigning lower weights to these concepts through modeling the contextual relationship of selected concepts. Nevertheless, it is also worth noting that the result of random walk is also somehow governed by the connectivity in context graph. For example, in 7(d), though both building and person related concepts are selected, the person-related concepts are eventually assigned with higher weights after random walk. This is simply because the person-related concepts reside in the largest clusters of context graph (see Figure 6(c)), and therefore results in higher weights through mutual concept interaction. In our experiments, there are around ten queries encountering this problem. In general, this issue, which also involves the type and size of concept vocabulary suitable for building a context graph that does not bias any particular type of concepts, is not trivial and deserves further studies.

E. Multi-modality Search

In this section, we conduct multi-modality search experiments. We compare the performance of our approach, namely CDMF, to heuristic fusion [11], linear oracle fusion, query-time fusion [14], and query-class dependent oracle fusion. All the approaches are tested using the three single modalities (text, visual and svMap−RW) introduced in the previous subsection. The settings of these approaches are as follows:

- Heuristic fusion (HF): heuristic linear fusion with predefined weights for modalities is the simplest approach for modality combination. We adopt the scheme in [11] which empirically determines the weight of each modality depending on nature of query terms. If only named entities are found in a query, the weights are set as 0.6, 0.3 and 0.1 for text, visual and concept respectively. If only the names of concept detectors are found, the weights are set as 0.1, 0.3, 0.6 accordingly. If both named entities and concept names are found, the weights become 0.35, 0.3 and 0.35. For any other cases, the weights for three modalities are 0.33, 0.33 and 0.34. While the scheme seems empirical, [11] has reported satisfactory search performance on TV08 dataset.
- Linear oracle fusion (LF^*): To test the upper limit of search performance using linear fusion, we perform oracle fusion by optimizing the fusion weights of each individual query. The optimal weights are obtained using Grid Search, and the best AP is then reported for each query.
- Query time fusion (QT): Instead of employing the predefined fusion weights as in HF, QT dynamically decides the weights on-the-fly during query time. We employ the approach in [14] which derives weights directly from the ranked lists of different modalities. As introduced in the related works, their idea is based on the hypothesis that the shape of the curve of the score distribution can imply the applicability of a modality to search. We adopt the same implementation as in [14] to derive the weight of a modality by computing the ratio of MAD (mean average distance) between the top 5% retrieved items and the remaining 95% items. The fusion scheme has been demonstrated in [14][15] and shows excellent performance in TRECVID 2007 and 2008 evaluations.
- Query-class dependent oracle fusion (QC^*): Different from the previous three approaches, QC^* predefines a set of classes and maps a query to one of the classes for fusion. The fusion weights, which are learned from

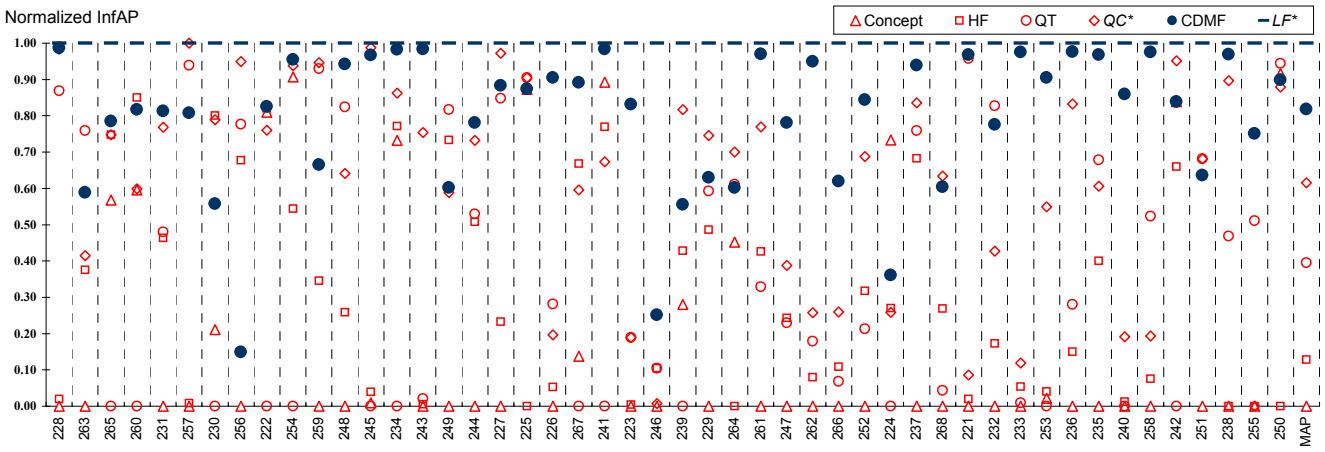


Fig. 8. Performance comparison of different multi-modality fusion strategies on TV08 dataset. The InfAP (y-axis) is scaled with *max-min* normalization. The query IDs are sorted in descending order according to the original (un-normalized) InfAP of LF^*.

TABLE II
PERFORMANCE COMPARISON OF CDMF AND OTHER STATE-OF-THE-ART MULTI-MODALITY FUSION STRATEGIES.

TV-	HF	LF^*	QT	QC^*	CDMF
05	.145	.179	.123	.167	.165
06	.056	.076	.055	.069	.071
07	.046	.065	.039	.056	.058
08	.048	.084	.059	.070	.074

* *oracle fusion*

TABLE III
SIGNIFICANCE TEST AT 0.05 LEVEL ($X \gg Y$ INDICATES THAT X IS SIGNIFICANTLY BETTER THAN Y).

TV-	Fusion Methods
05	CDMF, $QC^* \gg$ HF $\gg$ Concept $\gg$ QT
06	CDMF, $QC^* \gg$ HF, QT $\gg$ Concept
07	CDMF, $QC^* \gg$ HF $\gg$ QT, Concept
08	CDMF $\gg$ $QC^* \gg$ QT $\gg$ HF, Concept

training queries, are determined based on the characteristics of a class. In the experiment, we classify the set of queries into four classes: named entities (NE), person-thing (PT), event (E) and place (P), based on the query categorization provided by TRECVID [12]. Among the 120 queries being tested, 12 queries are assigned to class NE, 107 to PT, 53 to E and 23 to P. For each class, we use Grid Search again to find optimal weights which maximize the MAP of queries belong to this class for oracle fusion. If a query belongs to multiple classes, we use the average of corresponding optimal weights. This basically provides a clue to the upper limit performance of the query-class dependent fusion method.

Table II compares the performances of the four approaches. Except query-time fusion, all approaches show better search performance than single modality search. Compared to heuristic fusion (HF) and query time fusion (QT), CDMF exhibits performance improvements ranging from 13.79% to 54.16%. Even when compared to query-class dependent oracle fusions (QC^*) and linear oracle fusion (LF^*), CDMF slightly outperforms QC^* on TV06-TV08 while approaching the MAP of LF^* on TV06 and TV07 datasets. Note that LF^* can be treated as the upper limit of all the existing multi-modality video search methods (including QT and QC^*) where the modalities are linearly combined.

To further confirm the performance improvement, we conduct significance test on CDMF, QC^*, QT, HF and also the uni-modality concept-only search, using randomization test [28] suggested by TRECVID. The results at 0.05 significance level are shown in Table III. Overall, CDMF is significantly better than QT, HF and concept-only search. There is no performance difference between CDMF and QC^* on TV05 and TV07 datasets, but significance difference is found on TV06 and TV08 datasets.

Figure 8 shows the detailed performance of different approaches over the 48 queries on TV08. For the ease of illustration, the InfAP (y-axis) is normalized so that LF^* which shows the upper limit performance is always at a value of 1. We also sort the queries based on the oracle performance from LF*. From left to right, the queries become more difficult to answer. Among the 48 queries, compared to the methods other than LF^*, CDMF achieves the best performance for 29 queries, followed by QC^* for 12 queries and QT for 6 queries. There are also 12 queries where the performance of CDMF is almost the same to LF^*. In addition, we also found that CDMF performs consistently over queries at different difficulty levels, achieving near-oracle performance for both "easy" and "hard" (Figure 8 from left to right) queries.

Compared to HF, CDMF has the capability of dynamically assigning appropriate weights to the search modalities for fusion. For example, in Query-243 "*find one or more people, each looking into a microscope*", the selected concepts include *person*, *standing*, and *microscope*. In concept-driven fusion, three weights of 0.17, 0.19 and 0.64 are respectively assigned to concept, visual and text modalities. The assignment of the fusion weights can be intuitively explained as follows. Because the selected concepts either have very few training examples (e.g., *microscope*) or significantly vary from sample to sample in appearance (e.g., *standing*), their detectors are not very robust. Therefore, the concept-only modality is unreliable for this query. Also, the limited number of visual query examples

cannot adequately characterize the query intention. On the other hand, the text-only modality works fairly well for this query since the query term "microscope" is very descriptive and suitable for text search. In contrast, HF assigns weights of 0.60, 0.30, and 0.10 for this query, because the query term "microscope" directly matches to a semantic concept. Since detector reliability is not considered, this will not guarantee satisfactory search performance.

Query-class dependent fusion (QC^*), even under oracle setting, is found to be difficult to obtain optimal fusion weights that are appropriate for all query members. In our experiments, the fact that many queries are categorized as Person-Thing (PT) has made the searching of optimal weights which suits all the queries in PT class almost impossible. For instance, while concept modality is found to be effective for a majority of queries in PT which search for general object (e.g., person or boat), the modality is incapable for queries looking for specific objects under certain event (e.g., Query-255 "*one person getting out of or getting into a vehicle*"). To further confirm our observations, we conduct a clustering experiment to group the queries according to their optimal fusion weights (from LF^*). The aim is to see whether these queries form any natural clusters where uniform cluster(class)-level fusion weights can generate near-optimal performance. As expected, no clear pattern has been observed. This again confirms the need of using query-adaptive fusion methods such as CDMF. Finally, Query time fusion (QT) seems to be more appropriate for queries which favor text matching. In general, the rapid change of scores at initial rank list is more easily observed in text modality than other modalities. Thus, the weights determined by QT are somewhat biased.

VII. Conclusion

We have presented an approach named CDMF that dynamically leverages multiple modalities for video search, where the modality weights are computed based on a novel concept-modality relation matrix. We show that modality weights can be accurately computed for each query on-the-fly, which largely extends existing popular techniques that map a query to one of a few categories with pre-computed weights.

Experimental results suggest that the semantic concepts not only can be used in the concept-based search modality, but also could be explored for determining the weights of the search modalities. With our proposed approach, more suitable query adaptive weights can be computed without requiring additional training queries as in many existing methods.

Furthermore, our single modality search experiment reveals another application perspective of utilizing the context information (concept relationship), where a random walk process is imposed over a context graph to produce better concept selection for concept-based search. We have shown that this process is helpful for suppressing irrelevant concepts, while bringing into more relevant ones at the same time.

Although our approach does not produce excellent performance for all the evaluated queries, the results are encouraging for most of them. Compared to single modality search, the results clearly show that video search using multi-modalities is more effective, and suggest going beyond the popular query-class-based fusion of the multiple search modalities.

References

[1] 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.

[2] R. Yan, J. Yang, and A. Hauptmann, "Learning query-class dependent weights in automatic video retrieval," in *ACM International Conference on Multimedia*, 2004.

[3] T.-S. Chua, S.-Y. Neo, H.-K. Goh, M. Zhao, Y. Xiao, G. Wang, S. Gao, K. Chen, Q. Sun, and T. Qi, "TRECVID 2005 by NUS PRIS," in *NIST TRECVID Workshop*, 2005.

[4] L. S. Kennedy, A. P. Natsev, and S. Chang, "Automatic discovery of query-class-dependent models for multimodal search," in *ACM International Conference on Multimedia*, 2005.

[5] R. Yan and A. G. Hauptmann, "Probabilistic latent query analysis for combining multiple retrieval sources," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006.

[6] L. Xie, A. Natsev, and J. Tesic, "Dynamic multimodal fusion in video search," in *IEEE International Conference On Multimedia and Expo*, 2007.

[7] A. Yanagawa, S. F. Chang, L. Kennedy, and W. Hsu, "Columbia university's baseline detectors for 374 lscom semantic visual concepts," Columbia University, Technical Report, 2007.

[8] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. G. Hauptmann, "Representations of keypoint-based semantic concept detection: a comprehensive study," *IEEE Transaction on Multimedia*, vol. 12, no. 1, pp. 42–53, 2010.

[9] 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 International Conference on Multimedia*, 2006.

[10] T.-S. Chua, J.-H. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from national university of singapore," in *International Conference on Image and Video Retrieval*, 2009.

[11] C. G. M. Snoek, K. E. A. van de Sande, O. de Rooij, and et. al., "The MediaMill TRECVID 2008 semantic video search engine," in *NIST TRECVID Workshop*, 2008.

[12] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVID," in *ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2006.

[13] P. Wilkins, P. Ferguson, and A. F. Smeaton, "Using score distributions for query-time fusion in multimediaretrieval," in *ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2006.

[14] P. Wilkins, T. Adamek, G. J. Jones, and et. al., "TRECVID 2007 experiments at Dublin City University," in *NIST TRECVID Workshop*, 2007.

[15] J. Cao, Y.-D. Zhang, B.-I. Feng, and et. al., "MCG-ICT-

CAS TRECVID 2008 automatic video retrieval system," in *NIST TRECVID Workshop*, 2008.

[16] L. Kennedy, S.-F. Chang, and A. Natsev, "Query-adaptive fusion for multimodal search," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 567–588, 2008.

[17] X.-Y. Wei, C.-W. Ngo, and Y.-G. Jiang, "Selection of concept detectors for video search by ontology-enriched semantic spaces," *IEEE Trans. on Multimedia*, vol. 10, no. 6, pp. 1085–1096, 2008.

[18] J. Cao, H.-F. Jing, C.-W. Ngo, and Y.-D. Zhang, "Distribution-based concept selection for concept-based video retrieval," in *ACM International Conference on Multimedia*, 2009.

[19] 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 *International Conf. on Image and Video Retrieval*, 2006.

[20] Y.-G. Jiang, C.-W. Ngo, and S.-F. Chang, "Semantic context transfer across heterogeneous sources for domain adaptive video search," in *ACM International Conference on Multimedia*, 2009.

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

[22] X.-Y. Wei, Y.-G. Jiang, and C.-W. Ngo, "Exploring inter-concept relationship with context space for semantic video indexing," in *ACM International Conference on Image and Video Retrieval*, 2009.

[23] 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.

[24] A. N. Langville and C. D. Meyer, "A survey of eigenvector methods for web information retrieval," *SIAM Review*, vol. 47, no. 1, pp. 135–161, 2005.

[25] J. A. Aslam and E. Yilmaz, "Inferring document relevance via average precision," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, USA, August 2006, pp. 601–602.

[26] CMU, "The lemur toolkit for language modeling and information retrieval," 2006.

[27] A. P. Natsev, M. R. Naphade, and J. Tesic, "Learning the semantics of multimedia queries and concepts from a small number of examples," in *ACM International Conference on Multimedia*, 2005.

[28] 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.

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

[30] X.-R. Li, D. Wang, J.-M. Li, and B. Zhang, "Video search in concept subspace: a text-like paradigm," in *ACM International Conference on Image and Video Retrieval*, 2007.

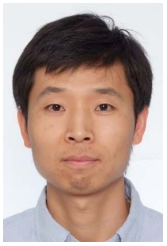
[31] M. R. Naphade and T. S. Huang, "A probabilistic framework for semantic video indexing, filtering, and retrieval," *IEEE MultiMedia*, vol. 3, no. 1, pp. 141–151, 2001.

[32] A. P. Natsev, A. Haubold, and et. al., "Semantic concept-based query expansion and re-ranking for multimedia retrieval," in *ACM International Conference on Multimedia*, 2007.

[33] G.-J. Qi, X.-S. Hua, and et. al., "Correlative multilabel video annotation with temporal kernels," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 5, no. 1, pp. 1–27, 2008.

[34] N. Rasiwasia, and N. Vasconcelos, "Image retrieval using query by contextual example," in *ACM International Conference on Multimedia Information Retrieval*, 2008.

Xiao-Yong Wei (M'10) is an Associate Professor with the School of Computer Science, Sichuan University, China. His research interests include multimedia computing, data mining, computational linguistics, and software engineering. He received his Ph.D. degree in Computer Science from City University of Hong Kong in 2009, his M.Eng. degree in Computer Science from Yunnan University in 2006, and his B.Eng. degree in Construction Engineering from Kunming University of Science and Technology in 2000. He was a Senior Research Associate in Dept. of Computer Science and Dept. of Chinese, Linguistics and Translation of City University of Hong Kong in 2009 and 2010 respectively. He had worked as a Manager of Software Dept. at Para Telecom Ltd., China, from 2000 to 2003.



Yu-Gang Jiang is a Postdoctoral Research Scientist in the Department of Electrical Engineering, Columbia University. His research interests include multimedia information retrieval, visual content analysis, and computer vision. He has authored more than 20 publications in these fields. He received his Ph.D. degree in Computer Science from City University of Hong Kong in 2009, and M.Eng. and B.Eng. degrees from Beijing Normal University and Jilin University, China, respectively. During his Ph.D. dissertation research, he also spent a year at Columbia University as a visiting scholar.

Chong-Wah Ngo (M'02) received his Ph.D in Computer Science from the Hong Kong University of Science & Technology in 2000. He received his MSc and BSc, both in Computer Engineering, from Nanyang Technological University of Singapore. He is currently an Associate Professor in City University of Hong Kong. He was with Beckman Institute of University of Illinois in Urbana-Champaign as post-doctoral researcher, and with Microsoft Research Asia as visiting researcher. His research interests include video computing and multimedia information retrieval.