Semantic context transfer across heterogeneous sources for domain adaptive video search

Yu-Gang JIANG

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

Shih-Fu CHANG

Citation

JIANG, Yu-Gang; NGO, Chong-wah; and CHANG, Shih-Fu. Semantic context transfer across heterogeneous sources for domain adaptive video search. (2009). *Proceedings of the 17th ACM international conference on Multimedia, MM'09, Beijing, China, October 19-24*. 155-164.
Available at: https://ink.library.smu.edu.sg/sis_research/6527

Semantic Context Transfer across Heterogeneous Sources for Domain Adaptive Video Search

Yu-Gang Jiang[†§], Chong-Wah Ngo[†], Shih-Fu Chang[§]
[†]Department of Computer Science, City University of Hong Kong
[§]Department of Electrical Engineering, Columbia University
{yjiang,cwngo}@cs.cityu.edu.hk; sfchang@ee.columbia.edu

ABSTRACT

Automatic video search based on semantic concept detectors has recently received significant attention. Since the number of available detectors is much smaller than the size of human vocabulary, one major challenge is to select appropriate detectors to response user queries. In this paper, we propose a novel approach that leverages heterogeneous knowledge sources for domain adaptive video search. First, instead of utilizing WordNet as most existing works, we exploit the context information associated with Flickr images to estimate query-detector similarity. The resulting measurement, named Flickr context similarity (FCS), reflects the co-occurrence statistics of words in image context rather than textual corpus. Starting from an initial detector set determined by FCS, our approach novelly transfers semantic context learned from test data domain to adaptively refine the query-detector similarity. The semantic context transfer process provides an effective means to cope with the domain shift between external knowledge source (e.g., Flickr context) and test data, which is a critical issue in video search. To the best of our knowledge, this work represents the first research aiming to tackle the challenging issue of domain change in video search. Extensive experiments on 120 textual queries over TRECVID 2005–2008 data sets demonstrate the effectiveness of semantic context transfer for domain adaptive video search. Results also show that the FCS is suitable for measuring query-detector similarity, producing better performance to various other popular measures.

Categories and Subject Descriptors

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

General Terms

Algorithm, Experimentation, Performance.

Keywords

Semantic Context Transfer, Heterogeneous Sources, Flickr Context Similarity, Domain Adaptive Video Search.

1. INTRODUCTION

Fueled by the ever-increasing amount of videos accumulated from a variety of applications, there is a need to develop automatic systems for effective and efficient content search. Different from text documents in which words are natural entities for semantic search, a video may convey mixed semantic meanings which are hard for computer to recognize, i.e., there is a well known semantic gap between computable low level features and the high level semantics.

Recent advances in multimedia research have shown encouraging progress in using a set of intermediate descriptors, namely semantic concept detectors, to bridge the semantic gap. The detectors are classifiers that automatically index the video contents with generic semantic concepts, such as *Tree* and *Water*. The indexing of these concepts allows users to access a video database by textual queries. In the search process, video clips which are most likely to contain the concepts semantically related to the query words are returned to the users. This video retrieval scenario is commonly referred to as concept-based video search.

However, due to the lack of manually labeled training samples and the limitation of computational resources, the number of available concept detectors to date remains in the scale of hundreds, which is much smaller compared to the size of human vocabulary. Therefore, one open issue underlying this search methodology is the selection of appropriate detectors for the queries, especially when direct matching of words fails. For example, given a query *find shots of something burning with flames visible*, *Explosion_fire* and *Smoke* are probably suitable detectors. Particularly, for large scale video search where the test data genre may change from time to time, the target domain data characteristics should be considered during detector selection. For instance, a detector *Military* may be highly related to a query *find shots of vehicles* in searching broadcast news video archives due to plenty of news events about wars (and thus videos showing *military vehicles*) in the Middle East, but the relationship may not hold in documentary videos. This brings a challenging question: how to adaptively select concept detectors based on the target domain data?

This paper proposes a novel approach that transfers semantic context across heterogeneous sources for domain adaptive video search. Here the semantic context can be either query-detector similarity or pairwise detector affinity, inferred from various knowledge sources. Different from most existing works in which semantic reasoning techniques based on WordNet were used for detector selection [27, 20, 19], we explore context information associated with Flickr images

MM'09, October 19–24, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-608-3/09/10 ...$10.00.

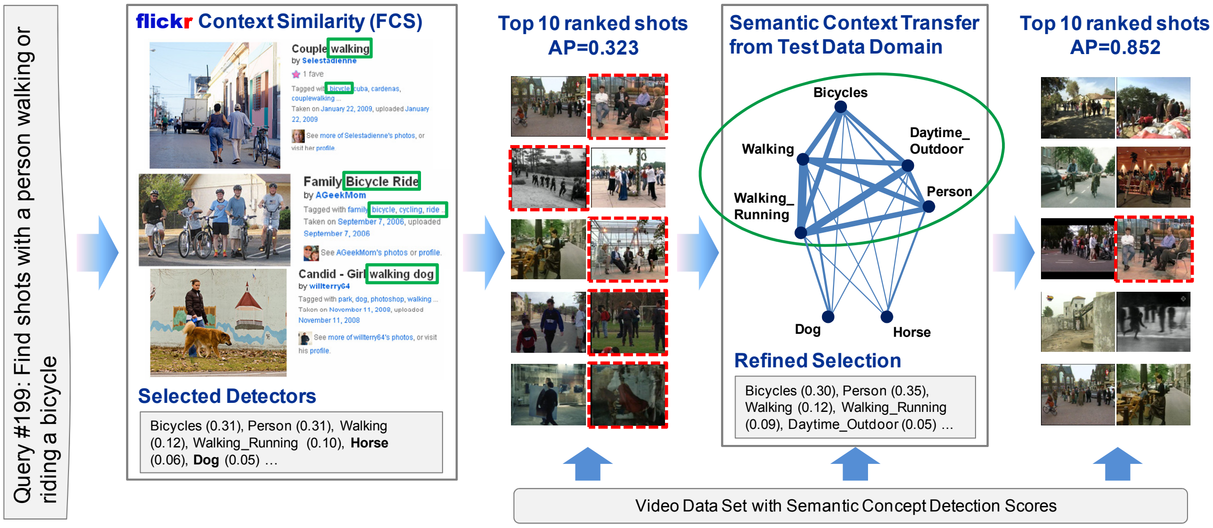

Query #199: Find shots with a person walking or riding a bicycle
flickr Context Similarity (FCS)
Couple walking
by Selestadienne
1 fave
Tagged with bicycle, cuba, cardenas, couplewalking ...
Taken on January 22, 2009, uploaded January 22, 2009
See more of Selestadienne's photos, or visit her profile.
Family Bicycle Ride
by AGeekMom
Tagged with family, bicycle, cycling, ride ...
Taken on September 7, 2006, uploaded September 7, 2006
See AGeekMom's photos or profile.
Candid - Girl walking dog
by willterry64
Tagged with park, dog, photoshop, walking ...
Taken on November 11, 2008, uploaded November 11, 2008
See more of willterry64's photos, or visit his profile.
Selected Detectors
Bicycles (0.31), Person (0.31), Walking (0.12), Walking_Running (0.10), Horse (0.06), Dog (0.05) …
Top 10 ranked shots
AP=0.323
Semantic Context Transfer from Test Data Domain
Bicycles
Walking
Daytime_ Outdoor
Walking_ Running
Person
Dog
Horse
Refined Selection
Bicycles (0.30), Person (0.35), Walking (0.12), Walking_Running (0.09), Daytime_Outdoor (0.05) …
Top 10 ranked shots
AP=0.852
Video Data Set with Semantic Concept Detection Scores

Figure 1: System architecture for domain adaptive video search, illustrated using a query from TRECVID 2007. Flickr context similarity is firstly applied to select a relevant detector set, which is then adaptively refined through transferring semantic context learnt from target data domain. The search performance in terms of average precision over the top-10 retrieved video shots is significantly improved by 164% after domain adaptation. The video shot rank lists are ordered from left to right and top to bottom (false positives are marked in red boxes).

for better query-detector similarity estimation. This measurement, named Flickr context similarity (FCS), is grounded on the co-occurrence statistics of two words in the context of images (e.g., tags, title, descriptions etc.), which implicitly reflects word co-occurrence in image context rather than textual corpus. This advantage of FCS enables a more appropriate selection of detectors for searching image and video data. For example, two words *Bridge* and *Stadium* have high semantic relatedness in WordNet, since both of them are very close to a common ancestor *construction* in the WordNet hierarchy. However, when a user issues a query *find shots of a bridge*, *Stadium* is obviously not a helpful detector since it rarely co-occurs with *bridge* in images/videos. While for the same query, FCS is able to suggest a more suitable detector *River* (cf. Section 3).

To cope with the domain shift between external knowledge source (e.g., Flickr context) and test data, we propose a novel algorithm which efficiently refines the initial detector selection based on semantic context learnt from target data domain. We formulate this problem as a semantic context transfer process using manifold regularization technique. One underlying principle of our formulation is that the selected detectors should be in accordance with the target domain data characteristics. Our method is highly generic in the sense that it is capable of learning the target domain knowledge without the need of any additional manual annotation. Figure 1 uses a query from TRECVID 2007 benchmark to further illustrate the proposed domain adaptive video search framework. Given a query *find shots with a person walking or riding a bicycle*, the following concept detectors {*Bicycle*, *Person*, *Walking*, *Walking_Running*, *Horse*, *Dog*, *Traffic*} were firstly selected by FCS from a pool of detectors defined in LSCOM [18]. Though we see that most of the selected detectors are suitable, a few of them are not consistent with the overall meaning of this query, such as *Horse* and *Dog* (chosen by query words *riding* and *walking* respectively). Through transferring semantic context automatically learnt from the test data, our approach ensures the semantic consistency of the selected detectors. As shown in Figure 1, it successfully removed the concepts *Horse* and *Dog*, while simultaneously added a new detector *Daytime_Outdoor* into the refined set because it frequently co-occurs with most selected concepts according to the test domain semantic context. This adaptation process significantly improved the search performance by 164% in terms of average precision over the top-10 ranked video shots.

The major contributions of this paper are summarized as follows.

1. We propose a novel algorithm that transfers semantic context across heterogeneous sources for domain adaptive video search. Our approach is highly efficient, enabling online detector selection for domain adaptive search of large scale video databases.

2. Through mining the context information associated with Flickr images, a word semantic similarity measurement, FCS, is developed, which is suitable for estimating query-detector similarity for concept-based video search.

In the following we review existing works in Section 2. We then describe the definition of the Flickr context similarity in Section 3. Section 4 elaborates our formulation of semantic context transfer for domain adaptive video search. The experimental results on video search and comparisons with the state of the arts are presented in Section 5. Finally, Section 6 concludes this paper.

2. RELATED WORK

Traditional video search systems usually extracted low-level features for direct matching with user query [26]. Such approaches often face difficulties in interpreting semantic queries due to the existence of the semantic gap. More recently, concept-based video search has been proposed by pooling a set of pre-trained semantic detectors to bridge the semantic gap. The semantic concepts cover a wide range of topics, such as objects (e.g., *Car* and *Bicycle*), scene (e.g., *Mountain* and *Desert*), events (e.g., *Meeting* and *Entertainment*) etc. The concept detectors can act as useful semantic filters for video search [27, 20, 29]. Such a video search framework involves two major efforts – the offline concept detection and the online selection of detectors for efficient search. Generic concept detection technique has been investigated by numerous studies in recent years [28, 12]. In order to identify a suitable set of concepts for detection, collaborative efforts have been pooled to assess the usefulness, observability, and feasibility of concepts [18], resulting a large scale concept ontology for multimedia (LSCOM) which includes a lexicon of more than 2000 concepts and annotations of 449 concepts. With LSCOM, two detector sets, Columbia374 [36] and VIREO-374 [12], were released, including low-level features, 374 concept detectors (classifier models), and detection scores on TRECVID 2005–2008 data sets. The 374 concepts are a subset of LSCOM with more than 10 annotated positive samples. In addition, another detector set commonly used is MediaMill-101 [28], containing 101 concept detectors.

Based on the detector sets, concept-based video search is executed through selecting appropriate detectors to interpret query semantics. The selection can be performed either through text matching between query words and concept names [27, 20], or based on the detection scores of the detectors to query image/video examples [4, 27, 29]. We only focus on the review of the text-based selection, since practically it will be quite difficult for users to acquire examples for their queries. We broadly divide existing works for text-based query-detector mapping into two categories based on the adopted knowledge source: 1) general purpose ontology [27, 20, 19, 30]; 2) large scale Web corpus [20, 8]. The former contains limited expert knowledge, while the latter has better coverage of contents, but it is also noisy.

Ontology-based mapping is grounded on general purpose vocabularies such as WordNet [6]. Through utilizing information from WordNet, e.g., word frequencies and hierarchical structure, a number of ontology reasoning techniques have been developed for estimating linguistic relatedness of words. Given a textual query, the detectors can be online selected based on their relatedness to the query words. Specifically, RES [24] which utilizes information content to measure word relatedness is adopted in [27, 20]. In addition, Lesk semantic relatedness [17] was used in [19] for detector selection. Other popular ontology-based mapping techniques include Wu&Palmer (WUP) [34], and Jiang&Conrath (JCN) [10]. With the ontology reasoning techniques, a recent work in [30] constructed a vector space, named ontology-enriched semantic space (OSS), by considering the pairwise relatedness of the concepts. In OSS, both query words and concept detectors are represented as vectors, and the relatedness measurement inferred from OSS has the merit of global consistency.

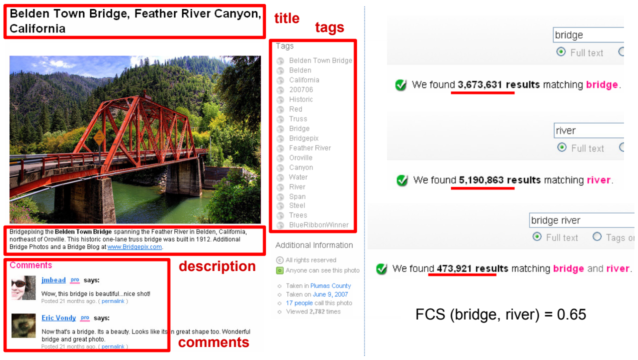
Compared to the rich information available on the Web, the knowledge in WordNet is derived from much smaller and outdated corpora (e.g., the information content is estimated from the Brown dictionary). The major shortcomings of such corpora are the low coverage of popular query words and potentially biased estimation of word/concept frequency, which stimulated researches on exploring the largest database available on the earth. In [20], Neo et al expanded the query words using internet news articles for better interpretation of the query semantics. The expanded query words are then used for detector selection, either by direct text matching or the ontology-based semantic reasoning techniques. A more recent work in [8] endeavored to estimate information content of words based on two web-based corpora: 1) samples of web pages which were downloaded using terms in WordNet as queries; 2) all the web pages indexed by Google (concept frequency is efficiently estimated by Google page hits). With the web-based information content, concept selection was done using JCN [10] for video search.

Other works for estimating word relatedness using information from the Web include normalized Google distance (NGD) [5] and Flickr distance [33], which have not been tested in the context of video search. Similar to [8], NGD also utilized the page hits returned by Google to estimate word relatedness. In view that all these popular measurements are based on textual documents and thus may not reflect word co-occurrence relationship in images/videos, Flickr distance was proposed by measuring image similarity based on visual features. This method, though promising in revealing visual co-occurrence, is computationally expensive to estimate pairwise relatedness of all the popular query words that a user may use. In this paper, as described in the following section, we adopt context information associated with Flickr images for measuring the word relatedness, which is as efficient as NGD from Google web search and also reflects the visual co-occurrence of words (cf. Section 5.2).

While the selection of detectors has been investigated in various works, the issue of domain changes in video search has not yet been fully investigated. In existing approaches [27, 8, 19, 30], the selected detectors are directly applied to response a query without considering data characteristics of target domain. Since the selection is done based on either ontology or Web sources, domain shift occurs in most of the cases in video search. In this paper, we consider this challenging issue through adapting detector selection based on the semantic context learnt from target domain. As a fact to recognize the importance of coping with data domain changes, there are a variety of domain transfer learning approaches developed in machine learning community [2, 23] and various application areas, such as text classification [35], natural language processing [9], and most recently, semantic concept detection [37, 11]. Different from these works that are all tailored for classification tasks, our approach aims to adapt the query-detector similarity, not classification models, for domain adaptive video search.

3. FLICKR CONTEXT SIMILARITY

The growing practice of online photo sharing has resulted in a huge amount of consumer photos accessible online. In addition to the abundant photo content, another attractive aspect of such photo sharing activity is the context information generated by users to depict the photos. As shown in Figure 2 (a), the rich context information includes title, tags, description and comments, which have been utilized for various applications, such as iconic image generation [16], tag disambiguation [32], and location-based photo organization [1]. In this section, we explore such context information for word similarity measurement, aiming to reflect their



Belden Town Bridge, Feather River Canyon, California

Bridgepixing the **Belden Town Bridge** spanning the Feather River in Belden, California, northeast of Oroville. This historic one-lane truss bridge was built in 1912. Additional Bridge Photos and a Bridge Blog at www.Bridgepix.com.

jmbead pro says:
Wow, this bridge is beautiful...nice shot!
Posted 21 months ago. (permalink)

Eric Vondy pro says:
Now that's a bridge. Its a beauty. Looks like its in great shape too. Wonderful bridge and great photo.
Posted 21 months ago. (permalink)

✓ We found **3,673,631 results** matching **bridge**.

✓ We found **5,190,863 results** matching **river**.

✓ We found **473,921 results** matching **bridge** and **river**.

FCS (bridge, river) = 0.65

(a) (b)

Figure 2: (a) Rich context information associated with a Flickr image. (b) The total number of images returned using keyword-based search in Flickr image context.

co-occurrence statistics in visual data rather than the text corpora used in [27, 20, 19, 8, 5].

Given two words, we compute their relatedness based on the number of Flickr images associated with them. With the number of hits returned by Flickr, we apply NGD derived from Kolmogorov complexity theory to estimate word distance [5]:

$$\mathrm{NGD}(x,y)=\frac{\max\{\log h(x),\log h(y)\}-\log h(x,y)}{\log N-\min\{\log h(x),\log h(y)\}},\qquad(1)$$

where $h(x)$ denotes the number of images associated with word x in their context, and $h(x,y)$ denotes the number of images associated with both words x and y; N is the total number of images on Flickr, which is roughly estimated as 3.5 billion by the time we did the experiments. The NGD is then converted to Flickr context similarity (FCS) using a Gaussian kernel, defined as

$$\mathrm{FCS}(x,y)=e^{-\mathrm{NGD}(x,y)/\rho},\qquad(2)$$

where the parameter ρ is empirically set as the average pairwise NGD among a randomly pooled set of words. Similar way of setting ρ has been shown to be effective for kernel based classification tasks [38]. An example of calculating FCS is shown in Figure 2 (b).

The major advantage of using full context information instead of tags alone is the better coverage of words. Figure 3 shows the frequency of 374 LSCOM concepts in various sources including Google web search, Flickr image context/tags, and the LSCOM manual annotations on TRECVID 2005 development set (43,873 shots). Obviously Google web search has the best coverage: the most rare concept (*Dredge_Powershovel_Dragline*) still appears in 2,120 web pages. Also, it can be clearly seen that the concept coverage of Flickr context is much better than that of Flickr tags. Only 2 concepts have zero frequency in context (*Non-US_National_Flags* and *Dredge_Powershovel_Dragline*), while in the tags, 53 concepts were not found. Although the coverage of Flickr context is not as good as Google web search, as will be shown in the experiments, it has the merit of reflecting the visual co-occurrence of words.

It is worthwhile to point out that the web-based sources are indeed noisy. For example, the precision of Flickr tags was found to be around 50% in [15]. The noise issue also exists in many web pages indexed by Google. A web page may contain multiple paragraphs of texts discussing different topics, resulting in misleading estimation of word co-occurrence. However, as was noted in [5], such noise can be partially made up by the huge data base size. This can be explained intuitively by the fact that two unrelated words may occasionally co-occur because of the noise, but probably not always. In other words, when the data base size increases, the number of co-occurrence of two related words will mostly increase at a much faster rate than that between two unrelated words. While we believe that techniques such as tag disambiguation [32] and image content based verification (Flickr distance [33]) are promising for alleviating the issue of noise, practically FCS is a much easier and cheaper way to measure the visual co-occurrence of all the words in human vocabulary.

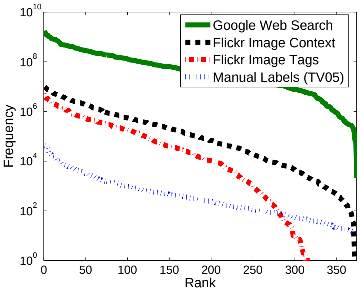


Figure 3: The frequency of 374 LSCOM semantic concepts in various sources. Note that the Y-axis is plotted in log scale.

4. SEMANTIC CONTEXT TRANSFER

This section describes our semantic context transfer algorithm. We start by defining a few notations. Let $\mathcal{C}=\{c_1,c_2,\cdots,c_m\}$ denote a semantic lexicon of m concepts and $\{\mathcal{X}_{trn},\mathcal{Y}_{trn}\}$ be a training data set, where $\mathcal{Y}_{trn}$ is the ground-truth label of $\mathcal{X}_{trn}$. Based on the training set, a classifier/detector is developed for each concept c_i using any supervised learning algorithm, such as SVMs. Another piece of useful information that can be learnt from the training set is inter-concept relationship, which can be easily computed based on the correlation of ground-truth labels. Formally, these are expressed as

$$\{\mathcal{X}_{trn},\mathcal{Y}_{trn}\}\rightarrow\{W_{trn},\mathcal{D}\},\qquad(3)$$

where $\mathcal{D}$ denotes a concept detection function for the m concepts and $W_{trn}\in\mathbb{R}^{m\times m}$ indicates the pairwise concept affinity. A large value w_{ij} in W_{trn} means two concepts c_i and c_j frequently co-occur (e.g., *car* and *road*). The detection function is then applied to a target data set $\mathcal{X}_{tgt}$ containing n test samples and produce detection score:

$$\mathcal{F}_{tgt}=\mathcal{D}(\mathcal{X}_{tgt}),\qquad(4)$$

where $\mathcal{F}_{tgt}=\{f(c_i)\}_{i=1,\cdots,m}\in\mathbb{R}^{m\times n}$.

In the search process, given a textual query q, external knowledge source such as WordNet ontology or Flickr context is used to measure query-detector similarity. This results in a vector $w_q=\{s(q,c_i)\}_{i=1,\cdots,m}$, which weighs the importance of the m detectors to the query q. The term $s(q,c_i)$, representing the similarity of c_i to q, is computed by accumulating the similarity of c_i to each query word in q.

With w_q and the concept detection score $\mathcal{F}_{tgt}$, the relevance score of the samples in $\mathcal{X}_{tgt}$ to q is computed as

$$f(q) = \frac{\sum_i^m s(q, c_i) f(c_i)}{\sum_i^m s(q, c_i)}, \tag{5}$$

where $f(q) \in \mathbb{R}^{1 \times n}$ is utilized to sort the samples to response query q. In practice, it is not necessary to consider all the m available detectors for each query word. A common practice is to use a sparse w_q by simply selecting the top-k relevant detectors for each query word, and then pool the selected detectors from all words for evaluating $f(q)$.

Equation 5 directly applies the similarity learnt from external sources, i.e., $s(q, c_i)$, to fuse the detectors trained individually from a training set. While the external knowledge is being leveraged, an important missing part is that $s(q, c_i)$ and $f(c_i)$ do not consider the data characteristics of target domain. The $\mathcal{X}_{tgt}$ could be in a domain more specific than the external knowledge, while is also different from the data distribution of the training set $\mathcal{X}_{trn}$. This section addresses this issue by presenting a novel two-step semantic context transfer algorithm. Specifically, the algorithm aims to transfer the semantic context inferred from target domain to adapt $f(c_i)$ and $s(q, c_i)$. The former adaptation is offline conducted by the time $\mathcal{X}_{tgt}$ arrives, while the latter is computed on-the-fly when queries are issued.

Offline Semantic Context Transfer. Given the initial detection score $\mathcal{F}_{tgt}$ and the concept affinity matrix W_{trn}, which is derived from the train set $\mathcal{X}_{trn}$, offline semantic context transfer adapts the concept affinity W_{trn} according to target domain data characteristics. The adapted concept affinity, W_{tgt}, further refines the detection score:

$$\{\mathcal{F}_{tgt}, W_{trn}\} \rightarrow \left\{\hat{\mathcal{F}}_{tgt}, W_{tgt}\right\}, \tag{6}$$

where $\hat{\mathcal{F}}_{tgt}$ is the refined detection score. This step is essentially a process of context-based concept fusion, which was initially proposed in [13], in which we named it as domain adaptive semantic diffusion.

Online Semantic Context Transfer. The vector w_q for a query q is estimated from external knowledge source, which obviously cannot accurately characterize the query-detector similarity in a new domain. The online semantic context transfer aims to simultaneously adapt w_q and update $f(q)$ based on target domain data characteristics, defined as

$$\left\{\hat{\mathcal{F}}_{tgt}, W_{tgt}, f(q), w_q\right\} \rightarrow \left\{\hat{f}(q), \hat{w}_q\right\}, \tag{7}$$

where $\hat{w}_q$ and $\hat{f}(q)$ contains the updated query-detector similarity and refined query relevance score respectively. This process is online executed for the queries given on-the-fly.

In the following we briefly introduce our formulation for offline transfer, based on which we derive the online transfer algorithm, which is the main focus of this paper.

4.1 Offline Detector Refinement

We first formulate the offline semantic context transfer for the refinement of concept detector scores. Considering the fact that the data distribution may change between $\mathcal{X}_{trn}$ and $\mathcal{X}_{tgt}$, in order to handle this issue, semantic context transfer should be investigated to infer a better concept affinity W_{tgt}. To achieve this, we define a risk function as

$$\{\hat{\mathcal{F}}_{tgt}, W_{tgt}\} = \arg\min_{\mathcal{F}, W} J(\mathcal{F}_{tgt}, W_{trn}), \tag{8}$$

where $\hat{\mathcal{F}}_{tgt}$ is the refined concept detection scores and W_{tgt} is the adapted concept affinity. Specifically, the risk function contains two components: intra-domain consistency constraint and inter-domain shift regularizer, defined as

$$\begin{aligned} J(\mathcal{F}_{tgt}, W_{trn}) &= \frac{\lambda}{2} \sum_{i,j=1}^{m} w_{ij}^{tgt} ||f(c_i) - f(c_j)||^2 \qquad (9) \\ &+ \frac{1}{2} \sum_{i,j=1}^{m} ||w_{ij}^{tgt} - w_{ij}^{trn}||^2, \end{aligned}$$

where $f(c_i)$ is the prediction score for concept c_i over test samples in target domain; w_{ij}^{trn} and w_{ij}^{tgt} represent the affinity of concept c_i and c_j in training and target test data respectively; λ captures the trade-off between the two terms.

This risk function can be intuitively explained as follows. First, the intra-domain consistency constraint ensures similar concept detection scores if two concepts are strongly correlated to each other, i.e. w_{ij} is large. In other words, minimizing J makes the detection scores consistent with the concept affinity. Second, the inter-domain shift regularizer means the adapted concept affinity W_{tgt} should not deviate too much from the initial one W_{trn}. Similar assumption is also adopted in classifier transfer learning approaches such as [35]. Therefore reducing the value of J enables the simultaneous refinement of both the detection score and the concept affinity.

To minimize J, we rewrite it into matrix form as

$$\begin{aligned} J(\mathcal{F}_{tgt}, W_{trn}) &= \frac{\lambda}{2} tr\{\mathcal{F}_{tgt}^{\top}(I - W_{tgt})\mathcal{F}_{tgt}\} \qquad (10) \\ &+ \frac{1}{2} tr\{(W_{tgt} - W_{trn})^{\top}(W_{tgt} - W_{trn})\}. \end{aligned}$$

Deriving the partial differential of J with respect to W_{tgt} and zero it as

$$\begin{aligned} \frac{\partial J}{\partial W_{tgt}} = 0 \quad &\Rightarrow \quad W_{tgt} - W_{trn} - \frac{\lambda}{2}\mathcal{F}_{tgt}\mathcal{F}_{tgt}^{\top} = 0 \\ &\Rightarrow \quad W_{tgt} = W_{trn} + \frac{\lambda}{2}\mathcal{F}_{tgt}\mathcal{F}_{tgt}^{\top}. \qquad (11) \end{aligned}$$

To derive the optimal detection score $\hat{\mathcal{F}}_{tgt}$, we apply stochastic gradient decent to recover the intra-domain consistency. With the concept affinity W_{tgt} in target domain, $\mathcal{F}_{tgt}$ can be updated as

$$\hat{\mathcal{F}}_{tgt} = \mathcal{F}_{tgt} - \eta \nabla_{\mathcal{F}_{tgt}} J, \tag{12}$$

where $\nabla_{\mathcal{F}_{tgt}} J = \lambda(I - W_{tgt})\mathcal{F}_{tgt}$ is the partial differential of J with respect to $\mathcal{F}_{tgt}$. The parameter η is commonly referred to as learning rate.

Note that in Equation 11 the concept affinity W_{tgt} is optimized based on the initial score $\mathcal{F}_{tgt}$. Practically Equations 11 and 12 can be applied iteratively to gradually adapt the concept affinity matrix and then accordingly refine the detection scores (cf. Algorithm 1).

4.2 Online Adaptation of Query-Detector Similarity

Now we consider the problem of online updating the query-detector similarity w_q based on target domain data characteristics. Recall that $w_q = \{s(q, c_i)\}_{i=1,\cdots,m} \in \mathbb{R}^{m \times 1}$, where $s(q, c_i)$ represents the relevance score of concept c_i to query q, estimated from external knowledge source such as Flickr context. Motivated by the online manifold regularization technique in feature space [7], we propose the following online semantic context transfer algorithm. Specifically, a new node is added into the concept space to represent q. We define the following new terms first:

$$\mathcal{F}^{\star}_{tgt} = \begin{bmatrix} \hat{\mathcal{F}}_{tgt} \\ f(q) \end{bmatrix} \qquad W^{\star}_{tgt} = \begin{bmatrix} W_{tgt} & \hat{w}_q \\ \hat{w}_q^{\top} & 0 \end{bmatrix} \tag{13}$$

$$I^{\star} = \begin{bmatrix} I & 0 \\ 0 & 1 \end{bmatrix} \qquad W^{\star}_{trn} = \begin{bmatrix} W_{tgt} & w_q \\ w_q^{\top} & 0 \end{bmatrix}$$

where $\hat{w}_q$ is the adapted query-detector similarity vector and $f(q) \in \mathbb{R}^{1\times n}$ is the initial relevance score to q, computed by Equation 5. Note that the refined detection score $\hat{\mathcal{F}}_{tgt}$ and the adapted concept affinity matrix W_{tgt} are used as inputs of the online semantic context transfer. The new matrices $W^{\star}_{trn}$ and $W^{\star}_{tgt}$ are also symmetric. We now rewrite the risk function in Equation 10 into the following online form:

$$\begin{aligned} & J^{\star}(\hat{\mathcal{F}}_{tgt}, W_{tgt}, f(q), w_q) \\ & = \frac{\lambda}{2} tr\{{\mathcal{F}^{\star}_{tgt}}^{\top}(I^{\star} - W^{\star}_{tgt})\mathcal{F^{\star}}_{tgt}\} \\ & \quad + \frac{1}{2} tr\{(W^{\star}_{tgt} - W^{\star}_{trn})^{\top}(W^{\star}_{tgt} - W^{\star}_{trn})\} \\ & = \frac{\lambda}{2}\Phi(\mathcal{F}^{\star}_{tgt}, W^{\star}_{tgt}) + \frac{1}{2}\Omega(W^{\star}_{tgt}, W^{\star}_{trn}), \end{aligned} \tag{14}$$

where Φ and Ω represent the online version of intra-domain consistency constraint and inter-domain shift regularizer respectively. Apparently, by treating query q as a new node in the concept space, minimizing $J^{\star}$ with respect to $\hat{w}_q$ facilitates the adaptation of the query-detector relationship. Also, the adapted query-detector relationship can be applied to refine the query relevance score $f(q)$. These dual processes are analogous to the adaptation of concept affinity and the refinement of concept detection score during offline transfer. To minimize Equation 14, we first expand Φ as follows:

$$\begin{aligned} & \Phi(\mathcal{F}^{\star}_{tgt}, W^{\star}_{tgt}) \\ & = tr\left\{\left[\hat{\mathcal{F}}_{tgt}^{\top}\ f(q)^{\top}\right]\left(I^{\star} - \begin{bmatrix} W_{tgt} & \hat{w}_q \\ \hat{w}_q^{\top} & 0 \end{bmatrix}\right)\begin{bmatrix} \hat{\mathcal{F}}_{tgt} \\ f(q) \end{bmatrix}\right\} \\ & = tr\{\hat{\mathcal{F}}_{tgt}^{\top}(I - W_{tgt})\hat{\mathcal{F}}_{tgt} - f(q)^{\top}\hat{w}_q^{\top}\hat{\mathcal{F}}_{tgt} \\ & \quad - \hat{\mathcal{F}}_{tgt}^{\top}\hat{w}_q f(q) + f(q)^{\top} f(q)\}. \end{aligned} \tag{15}$$

Likewise, we also expand the online version of inter-domain shift regularizer as

$$\begin{aligned} & \Omega(W^{\star}_{tgt}, W^{\star}_{trn}) \\ & = tr\left\{\left(\begin{bmatrix} W_{tgt} & \hat{w}_q \\ \hat{w}_q^{\top} & 0 \end{bmatrix} - \begin{bmatrix} W_{tgt} & w_q \\ w_q^{\top} & 0 \end{bmatrix}\right)^2\right\} \\ & = tr\left\{\begin{bmatrix} 0 & \hat{w}_q - w_q \\ \hat{w}_q^{\top} - w_q^{\top} & 0 \end{bmatrix}^2\right\} \\ & = tr\left\{(\hat{w}_q - w_q)(\hat{w}_q - w_q)^{\top}\right\} \\ & \quad + (\hat{w}_q - w_q)^{\top}(\hat{w}_q - w_q). \end{aligned} \tag{16}$$

With Equations 15 and 16, now we can easily derive the partial differential of the risk function $J^{\star}$ with respect to $\hat{w}_q$:

$$\begin{aligned} \frac{\partial J^{\star}}{\partial \hat{w}_q} & = \frac{\lambda \cdot \partial\Phi}{2\partial\hat{w}_q} + \frac{\partial\Omega}{2\partial\hat{w}_q} \\ & = -\lambda\hat{\mathcal{F}}_{tgt} f(q)^{\top} + 2(\hat{w}_q - w_q). \end{aligned} \tag{17}$$

Algorithm 1 : Semantic Context Transfer

```
Offline Transfer:
Input: the initial detection score $\mathcal{F}_{tgt}$;
        the initial concept affinity matrix $W_{trn}$.
Initialization:
        $\hat{\mathcal{F}}^0_{tgt} = \mathcal{F}_{tgt}$; $W^0_{tgt} = W_{trn}$.
Loop: $t = 0, \cdots, T_1$
        $W^{t+1}_{tgt} = W^t_{tgt} + \frac{\lambda}{2}\hat{\mathcal{F}}^t_{tgt}{\hat{\mathcal{F}^t}}_{tgt}^{\top}$ (Equation 11);
        $\hat{\mathcal{F}}^{t+1}_{tgt} = \hat{\mathcal{F}}^t_{tgt} - \eta\lambda(I - W^{t+1}_{tgt})\hat{\mathcal{F}}^t_{tgt}$ (Equation 12).
Output: the refined detection score $\hat{\mathcal{F}}_{tgt}$;
          the adapted concept affinity matrix $W_{tgt}$.

Online Transfer:
Input: the refined detection score $\hat{\mathcal{F}}_{tgt}$;
        the adapted concept affinity matrix $W_{tgt}$;
        the initial query-detector similarity $w_q$ of a new query $q$.
Initialization:
        $\hat{w}^0_q = w_q$; $\hat{f}(q)^0 = \frac{\hat{\mathcal{F}}_{tgt} w_q}{\vec{1} w_q}$.
Loop: $t = 0, \cdots, T_2$
        $\hat{w}^{t+1}_q = \hat{w}^t_q + \frac{\lambda}{2}\hat{\mathcal{F}}_{tgt}{\hat{f}(q)^t}^{\top}$ (Equation 20);
        $\hat{f}(q)^{t+1} = \frac{\hat{\mathcal{F}}_{tgt}\hat{w}^{t+1}_q}{\vec{1}\hat{w}^{t+1}_q}$.
Output: the refined query relevance score $\hat{f}(q)$.
```

Zeroing the above partial differential, the optimal query-detector similarity $\hat{w}_q$ is computed as

$$\begin{aligned} \nabla_{\hat{w}_q} J^{\star} = 0 \quad & \Rightarrow \quad -\lambda\hat{\mathcal{F}}_{tgt} f(q)^{\top} + 2(\hat{w}_q - w_q) = 0 \\ & \Rightarrow \quad \hat{w}_q = w_q + \frac{\lambda}{2}\hat{\mathcal{F}}_{tgt} f(q)^{\top}. \end{aligned} \tag{18}$$

With the adapted query-detector similarity $\hat{w}_q$, the query relevance score can be updated accordingly:

$$\hat{f}(q) \;=\; \frac{\hat{\mathcal{F}}_{tgt}\hat{w}_q}{\vec{1}\hat{w}_q}, \tag{19}$$

where $\vec{1} = \{1, \cdots, 1\} \in \mathbb{R}^{1\times m}$ is a row vector. Similar to the offline transfer process, $J^{\star}$ can be gradually minimized through iteratively updating w_q and $f(q)$ as follows

$$\begin{aligned} \hat{w}^{t+1}_q & = \hat{w}^t_q + \frac{\lambda}{2}\hat{\mathcal{F}}_{tgt}{\hat{f}(q)^t}^{\top}, \\ \hat{f}(q)^{t+1} & = \frac{\hat{\mathcal{F}}_{tgt}\hat{w}^{t+1}_q}{\vec{1}\hat{w}^{t+1}_q}. \end{aligned} \tag{20}$$

The above equations achieve the simultaneous online refinement of query-detector similarity and query relevance score. Note that in our implementation, to keep w_q sparse, we round elements in $\hat{\mathcal{F}}_{tgt} f(q)^{\top}$ with small absolute values to zero at each iteration.

4.3 Algorithm Summary and Discussion

We summarize both offline and online semantic context transfer processes in Algorithm 1. To see how the test domain knowledge is used, we intuitively explain the algorithm as follows. Recall that $\mathcal{F}_{tgt} = \{f(c_i)\} \in \mathbb{R}^{m\times n}$. The product $\mathcal{F}_{tgt}\mathcal{F}_{tgt}^{\top}$ implies the pairwise concept affinity computed by the detection scores in the target domain (the score vector of each concept has been normalized to unit length). Therefore, the offline transfer process implicitly incorporates the target domain knowledge in $\hat{\mathcal{F}}_{tgt}$. Similarly, the knowledge is also ingested into $\hat{w}_q$ through applying term $\hat{\mathcal{F}}_{tgt}{\hat{f}(q)}^{\top}$, which hints the query-detector similarity based on their score distribution in the target domain.

It is interesting to note that our formulation of semantic context transfer can be connected to the graph-based semi-supervised learning algorithms with manifold regularization [39, 40], which aim to propagate labels based on the geometry structure of data sample manifold. Mathematically, our intra-domain consistency constraint is similar to the smoothness constraint in [39, 40]. However, in our formulation we treat concepts (and also queries in the online version), not data samples, as nodes in the manifold structure. In other words, our formulation recovers the consistency of the detection score (query relevance score) with respect to the concept affinity (query-detector similarity). Also, different from [39, 40] where the manifold structure is fixed, we use the domain shift of manifold structure as a regularization term and alternatively modify the structure to well fit the target domain data characteristics.

5. EXPERIMENTS

5.1 Data Sets and Evaluation

We conduct video search experiments using the TRECVID 2005-2008 data sets (abbr. TV05/06/07/08), which were used in the annual benchmark evaluation organized by NIST [25]. In total, there are 440 hours of video data and 120 officially evaluated queries. The data sets are accompanied with a standard reference of shot boundaries, which form the basic unit for evaluation. Detailed descriptions of each year's data are listed in Table 1. As shown in the table, TV05 and TV06 are broadcast news videos from US, Arabic, and Chinese sources, while TV07 and TV08 are mostly documentary videos from the Netherlands Institute for Sound and Vision. Table 2 shows a few example textual queries, which are usually very short and abbreviated with a few words. Throughout the experiments, we consider only nouns and gerunds in the queries for detector selection, assuming that nouns indicate the name of place, thing or person (e.g., *computer*), and gerunds describes an action/event (e.g., *walking*). In the benchmark evaluations, most queries are also associated by a few query image/video examples, while in the experiments, we only use the textual queries. Compared to querying with the image/video examples, this represents a more general and practical scenario of video search.

For the semantic concept detectors, we use VIREO-374 [12] for TV05–07. VIREO-374 is composed of detectors for 374 LSCOM semantic concepts and their detection scores on TV05–07 test sets. The detectors in VIREO-374 are trained using the TV05 development set. Each detector is associated with three SVM classifiers trained using different image features (color moment, wavelet texture and bag-of-visual-words) extracted from video frames. The outputs of the three classifiers are combined as the final detection score using average fusion. For TV08, we use the recently released CU-VIREO374 [14]. Based on the detection scores, we directly work on each year's test set to evaluate the effectiveness of our domain adaptive video search approach.

Table 1: Descriptions of TV05–08 data sets. The total number of video shots in each data set is shown in the parenthesis. The 160h (100h) data from TV05 (07) is used as development data for TV06 (08).

TV	Data domain	Devel. set	Test set	# queries
05	Broadcast news	80h(43,873)	80h(45,765)	24
06	Broadcast news	–	80h(79,484)	24
07	Documentary	50h(21,532)	50h(22,084)	24
08	Documentary	–	100h(35,766)	48

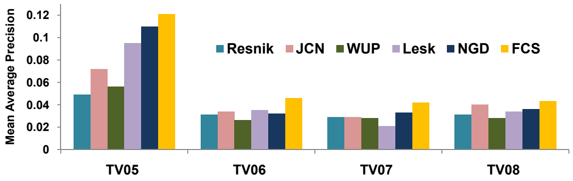


Figure 4: Performance comparison of different query-detector similarity measurements on TV05-08 test sets.

For each query, the retrieved video shots are ranked according to their scores to the selected concept detectors. The search performance is evaluated using average precision (AP), defined as $AP = \frac{1}{\min(R,k)} \sum_{j=1}^{k} \frac{R_j}{j} \times I_j$, where R and R_j are the total number of true positives in the whole test set and the top-j shots respectively; $I_j = 1$ if the jth shot is relevant and 0 otherwise. To aggregate the performance over multiple queries, mean average precision (MAP) is used. In the experiments, unless otherwise stated, we set $k = 1000$ following TRECVID standard.

In the following we first compare FCS with a variety of existing word similarity measurements. We then evaluate the effectiveness of our semantic context transfer algorithm for domain adaptive video search.

5.2 Query-Detector Similarity Measure

To verify the merit of using Flickr context for estimating query-detector similarity, we compare FCS with five other measures, including normalized Google distance (NGD) [5] using Yahoo web search as knowledge source, and WordNet based measurements including RES [24], JCN [10], WUP [34], and Lesk [17]. RES and JCN uses information content estimated based on the Brown Corpus for reasoning the query-detector relationship, while Lesk and WUP use glosses and path-length/depth in WordNet hierarchy respectively. For each query word, we select top-3 most related detectors. The search results returned by the detectors are then linearly fused (Equation 5). Depending on the measure used, the weight of a detector is set equal to its similarity to the corresponding query.

Figure 4 shows the detailed experimental results in terms of MAP over different test sets. We see that the performance of web knowledge based measurements, especially the FCS, is apparently better than the WordNet-based ones. The improvement is particularly obvious for TV05 and TV07 where there are plenty of query words for which detectors with exactly the same names cannot be found. The web provides up-to-date information and better coverage of words, which is indeed very helpful for such cases. For example, for query term *Condoleeza_Rice*[1] which does not appear in WordNet, detectors *Colin_Powell* and *Donald_Rumsfeld* are suggested by NGD and FCS respectively, since both of them frequently co-occur with *Condoleeza_Rice*. On the other hand, FCS also constantly outperforms NGD with a large margin (performance improvement ranges from 7% to 41%). This indeed confirms the advantage of using Flickr context for estimating query-detector similarity – as we discussed in Section 3, it is able to reflect the word co-occurrence in visual content (images) rather than text corpus.

[1] A name entity detection tool is applied so that words from one name will not be treated as separate query terms.

Table 2: Detector selection using various query-detector similarity measurements. The detectors are selected based on the query words shown in bold.

ID	Query	WordNet (WUP)	Google Web Search (NGD)	Flickr Context (FCS)
171	A **goal** in a soccer match	Striking	Sports	Soccer
188	Something burning with **flames** visible	Sky	Soldiers	Smoke
196	**Scenes** with snow	Landscape	Person	Urban_Scenes
205	A **train** in motion	Vehicle	Car	Railroad

Table 2 gives a few example queries to further compare different measurements. Due to space limitation, we only list the most suitable detector for one selected word from each query. In the query ID-188, for query word *flames*, detector *Soldiers* is selected by NGD and *Smoke* is chosen by FCS. While the selection of *Soldiers* (e.g., in war scenes) and *Smoke* is potentially helpful for searching videos about *flames*, such semantic relationship was not captured by WordNet. Similar observation also holds for queries ID-171/196/205. More interestingly, compared to NGD, we see that FCS is capable of selecting more suitable detectors. For example, in query ID-205, *Car* is selected by NGD to response query term *train* because there are many web pages containing content related to both kinds of vehicles. However, obviously *Car* is less likely to be helpful for retrieving *Train* since they rarely co-occur in image/video data. While for FCS, a more suitable detector, *Railroad*, is selected. These observations again confirm the advantage of using FCS for detector selection in video search. However, note that although FCS shows promising results in many cases, the selection is done based on knowledge from the web without considering the data characteristics in the target data domain. In the next subsection, we have extensive experiments to see how the semantic context transfer algorithm works for coping with domain change.

5.3 Effect of Semantic Context Transfer

The performances of semantic context transfer (SCT) over different data sets are shown in Table 3. To better analyze the performance, we list the MAP over top k = 10, 30, 100 and 1000 ranked shots[2]. Clearly, SCT shows noticeable performance gains for most of the experimental settings. When $k = 1000$, except on TV07 that the performance is about the same, the improvement on the other three test sets ranges from 8% to 16%. This is not completely a surprise because among all the test sets, TV07 has the smallest number of true positives (196 per query on average). From our analysis on the distribution of true positives of the rank lists, for TV07, more relevant shots are observed within the top-100, while for the other test sets, the relevant shots tend to spread throughout the result list. When considering less top-k retrieved shots, the improvement of MAP becomes more obvious. For instance, on TV07, the improvement over the top-10 ranked list is as high as 23%. This is significant since practically for most search applications, top-10 is a reasonable number of results that a user might browse.

To study the effect of offline transfer, we conduct another experiment on TV08, in which we only apply Equations 11 and 12 to update the detection score, omitting the online transfer process. The performance in terms of MAP-1000 is 0.046, which is obviously lower than that when online transfer is jointly applied (0.050). As mentioned in Section 4, it is understandable because in video search the query-detector similarity is highly important. Also, improving the accuracy of each single concept detector does not guarantee better fusion performance. The performance gain of using improved detector for video search is similar to that reported in an earlier work [31].

Table 3: Search performance on TV05–08. MAP-k means MAP over top k ranked shots[2].

TV	Method	MAP-10	MAP-30	MAP-100	MAP-1000
05	FCS	0.229	0.196	0.160	0.118
	FCS+SCT	**0.245**	**0.203**	**0.170**	**0.127**
06	FCS	0.121	0.086	0.062	0.045
	FCS+SCT	**0.147**	**0.098**	**0.066**	**0.050**
07	FCS	0.064	0.048	0.041	**0.042**
	FCS+SCT	**0.079**	**0.050**	**0.043**	**0.042**
08	FCS	-	-	-	0.043
	FCS+SCT	-	-	-	**0.050**

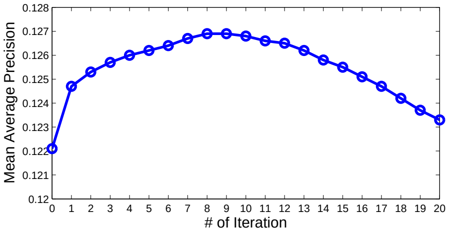


Figure 5: MAP performance on TV05 by varying the number of iterations for online semantic context transfer.

In order to verify whether the performance improvement is due to chance, we conduct significant test based on the per-query APs (k=1000). We adopt the randomization test suggested by TRECVID[3], where the target number of iterations is set to 10,000. At the 0.05 level of significance, FCS+SCT is significantly better than FCS, while FCS is also significantly better than other word similarity measurements.

5.3.1 Parameter Sensitivity

There are mainly four parameters in the semantic context transfer algorithm, including λ, η, and the number of offline and online iterations T_1 and T_2 respectively. Throughout the experiments, λ, η and T_1 are uniformly set as 0.1, 0.05 and 20 respectively, following our findings in [13].

For the online transfer iteration number T_2, we empirically determine its suitable value. We evaluate the sensitivity of search performance to T_2 on TV05. As shown in Figure 5, the performance increases significantly at the beginning and then remains fairly stable for a couple of iterations. Also, we see that the best or close to best accuracies are achieved when T_2 is around 8. The same T_2 is applied in all the experiments and consistent performance gains are observed

[2] For TV08, TRECVID only provided incomplete ground-truth labels and used inferred AP [3] for result evaluation. Since the evaluation tool from TRECVID does not support the calculation of inferred AP when $k \neq 1000$, we only report result of $k = 1000$.

[3] http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/

Table 4: Significance test based on query types. $x \gg y$ means x is significantly better than y.

Query Type	Method
Event	FCS+SCT≫FCS≫Lesk; FCS≫NGD
PT	FCS+SCT≫FCS≫Lesk; FCS≫NGD
Place	FCS+SCT≫FCS≫NGD≫Lesk
NE	FCS+SCT≫Lesk; FCS≫Lesk; NGD≫Lesk

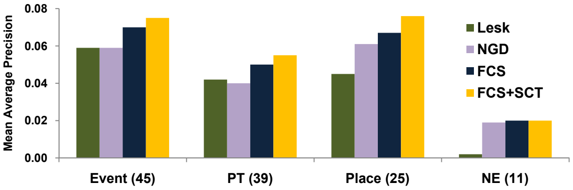


Figure 6: Performance of different query types. The number of queries for each type is shown in the parenthesis. PT and NE stand for person-things and name-entity respectively.

over the TV06–08 test sets (cf. Table 3), which confirm the performance stability of the proposed algorithm over parameter settings.

5.3.2 Speed Efficiency

Speed is a critical requirement for online video search. Our online semantic context transfer algorithm is extremely efficient. The complexity of the algorithm is $O(mn)$, where m is the number of available detectors and n is the number of test shots. More specifically, the total run time of the 24 queries on TV06 (79,484 video shots) is 30.2 seconds on a regular PC (Intel Core2 Duo 2.2GHz CPU and 2GB RAM). In other words, performing online transfer for one query only takes 1.26 seconds. It can be even faster if executed on a more powerful machine with parallel computing capability. Clearly, this satisfies the need for online search.

5.4 Performance based on Query Types

To further study the effectiveness of each similarity measure and the semantic context transfer algorithm, we now discuss search performance based on query types. We roughly group the 120 queries into four categories: event, person+things (PT), place, and name entity (NE). The grouping is based on the query classification suggested by TRECVID [22]. Because most queries are related to PT, we exclude a query from PT once it belongs to one of the other three categories.

Figure 6 shows the performances of MAP-1000. The web-based measures (NGD and FCS) are particularly good for name entity queries. This is intuitive as most of name entities are not defined in WordNet. Overall, the best performances of all the query classes come from FCS+SCT. This observation shows the advantage of stability of our approach over various query types. On the other hand, FCS also outperforms NGD and Lesk, which are computed based on the web pages and the WordNet vocabulary respectively. To further verify the consistency of performance improvement for different query types, we also conduct randomization test at 0.05 level of significance. Results are summarized in Table 4. Except NE, FCS+SCT is significantly better than FCS and FCS is more effective than NGD and Lesk. For NE, no improvement is observed from SCT because the detectors selected by NGD/FCS are already very appropriate.

Table 5: Performance comparison on TV05–08 test sets.

TV	Haubold et al. [8]	Snoek et al. [27]	Neo et al. [20]	Wei et al. [31, 21]	FCS+ SCT
05	0.028	0.049	0.113	**0.127**	**0.127**
06	0.020	N/A	N/A	0.049	**0.050**
07	0.018	N/A	N/A	0.039	**0.042**
08	N/A	N/A	N/A	0.042	**0.050**

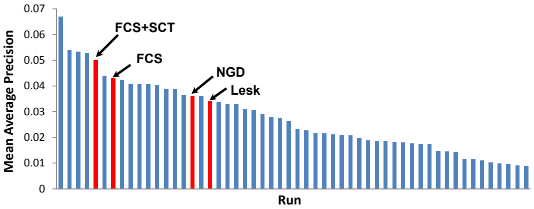


Figure 7: MAP comparison with the top-50 (out of 82) official submissions of the automatic video search task in TRECVID 2008.

5.5 Comparison to the State of the Arts

In this section, we compare our results to several recent works on concept-based video search [8, 20, 27, 31]. Web-based information content is used in [8, 20] for detector selection, and WordNet ontology based similarity measurement is adopted in [27]. In [31], a multi-level fusion framework is developed considering the semantics, observability, reliability and diversity for detector selection. Note that several results reported in these works used different detector set. The aim of the comparison is to show how the whole framework proposed in this paper performs compared with the state of the arts. Table 5 lists the performance of each approach over TV05–08 test sets. Our domain adaptive video search framework (FCS+SCT) performs best for all the four years' test sets. Note that several useful factors such as the diversity of the selected detectors [31] have not been considered in our current framework and therefore can be adopted for further improvement. Figure 7 further compares our results with the official submissions in TV08. Among all the 82 submissions, the proposed FCS+SCT using textual query alone ranks fifth, while all the top four runs adopted both textual query and image/video examples, e.g., the best performing system [29] contains three modalities: text matching, concept-based search, and image/video example matching.

6. CONCLUSIONS

We have presented an approach that transfers semantic context across heterogeneous sources for domain adaptive video search. Given a textual query, we utilize Flickr context for initial concept detector selection, and then transfer semantic context learnt from target data domain to improve concept detector accuracy and refine query-detector similarity. The extensive experiments confirm the advantage of FCS (Flickr context similarity) in revealing visual co-occurrence of words and the effectiveness of our semantic context transfer algorithm for domain adaptive search. Significant and consistent improvements are reported over the challenging TV05–08 video search benchmarks.

Practically a video search system may be applied to data from any domain. Our algorithm learns semantic context from target domain based on the outputs of the pre-trained semantic detectors, without requiring any manual annotation on the new test data. The unsupervised learning of domain change is considered as an important merit of our proposed work. Additionally, we also demonstrated that the domain adaptation process can be performed online.

Currently our approach treats words extracted from a query with equal importance. In other words, while the relevancy of detectors are ranked according to a query word, the significance of a query word to overall search performance is not considered. Thus one possible improvement to current work is the refinement of search result by also ranking the importance of query words according to context or by natural language processing. In addition, instead of uniformly choosing three detectors for each query word as in our current work, adaptive selection of detectors is another interesting future direction that may lead to further improvement.

Acknowledgement

The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 119709), a DARPA grant under Contract No. HR0011-08-C-0135, and a NSF grant CNS-0751078. We are grateful to Jun Wang for his help in the formulation part and Xiao-Yong Wei for providing baseline results using WordNet.

7. REFERENCES

[1] S. Ahern, M. Naaman, R. Nair, and J. Yang. World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. In *JCDL*, 2007.

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS*, 2006.

[3] J. A. Aslam, V. Pavlu, and E. Yilmaz. Statistical method for system evaluation using incomplete judgments. In *ACM SIGIR*, 2006.

[4] M. Campbell and et al. IBM research trecvid-2006 video retrieval system. In *TRECVID Workshop*, 2006.

[5] R. L. Cilibrasi and P. M. Vitanyi. The google similarity distance. *IEEE Trans. on Knowledge and Data Engineering*, 19:370–383, 2007.

[6] C. Fellbaum and Ed. *WordNet: an electronic lexical database*. The MIT Press, 1998.

[7] A. B. Goldberg, M. Li, and X. Zhu. Online manifold regularization: A new learning setting and empirical study. In *ECML PKDD*, 2008.

[8] A. Haubold and A. Natsev. Web-based information content and its application to concept-based video retrieval. In *CIVR*, 2008.

[9] J. Jiang and C. Zhai. Instance weighting for domain adaptation in NLP. In *ACL*, 2007.

[10] J. J. Jiang. Semantic similarity based on corpus statistics and lexical taxonomy. In *ROCLING*, 1997.

[11] W. Jiang, E. Zavesky, S.-F. Chang, and A. Loui. Cross-Domain Learning Methods for High-Level Visual Concept Classification. In *ICIP*, 2008.

[12] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *CIVR*, 2007.

[13] Y.-G. Jiang, J. Wang, S.-F. Chang, and C.-W. Ngo. Domain adaptive semantic diffusion for large scale context-based video annotation. In *ICCV*, 2009.

[14] Y.-G. Jiang, A. Yanagawa, S.-F. Chang, and C.-W. Ngo. CU-VIREO374: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection. In *Columbia University ADVENT Technical Report #223-2008-1*, August 2008.

[15] L. Kennedy, S.-F. Chang, and I. Kozintsev. To search or to label?: Predicting the performance of search-based automatic image classifiers. In *ACM MIR*, 2006.

[16] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How flickr helps us make sense of the world: Context and content in community contributed media collections. In *ACM Multimedia*, 2007.

[17] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine code from an ice cream cone. In *SIGDOC*, pages 24–26, 1986.

[18] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large scale concept ontology for multimedia. *IEEE Multimedia*, 2006.

[19] A. Natsev, A. Haubold, J. Tesic, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *ACM Multimedia*, 2007.

[20] 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 *CIVR*, 2006.

[21] C.-W. Ngo, Y.-G. Jiang, X. Wei, W. Zhao, F. Wang, X. Wu, and H.-K. Tan. Beyond semantic search: What you observe may not be what you think. In *TRECVID Workshop*, 2008.

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

[23] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: Transfer learning from unlabeled data. In *ICML*, 2007.

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

[25] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *ACM MIR*, 2006.

[26] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content based image retrieval at the end of the early years. *IEEE Trans. on PAMI*, 22:1349–1380, 2000.

[27] 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*, 9(5):975–986, 2007.

[28] 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 Multimedia*, 2006.

[29] S. Tang, J.-T. Li, M. Li, and et al. Trecvid 2008 participation by mcg-ict-cas. In *TRECVID Workshop*, 2008.

[30] X.-Y. Wei and C.-W. Ngo. Ontology-enriched semantic space for video search. In *ACM Multimedia*, 2007.

[31] X.-Y. Wei and C.-W. Ngo. Fusing semantics, observability, reliability and diversity of concept detectors for video search. In *ACM Multimedia*, 2008.

[32] K. Weinberger, M. Slaney, and R. van Zwol. Resolving tag ambiguity. In *ACM Multimedia*, 2008.

[33] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li. Flickr distance. In *ACM Multimedia*, 2008.

[34] Z. Wu and M. Palmer. Verb semantic and lexical selection. In *ACL*, 1994.

[35] G. Xue, W. Dai, Q. Yang, and Y. Yu. Topic-bridged PLSA for cross-domain text classification. In *SIGIR*, 2008.

[36] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia university's baseline detectors for 374 lscom semantic visual concepts. In *Columbia University ADVENT Technical Report #222-2006-8*, March 2007.

[37] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *ACM Multimedia*, 2007.

[38] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. In *IJCV*, 2007.

[39] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In *NIPS*, 2004.

[40] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 2003.