Domain Adaptive Semantic Diffusion for Large Scale Context-Based Video Annotation

Yu-Gang Jiang[1,2], Jun Wang[1], Shih-Fu Chang[1] and Chong-Wah Ngo[2]
[1]Dept. of Electrical Engineering, Columbia University
[2]Dept. of Computer Science, City University of Hong Kong
{yjiang,cwngo}@cs.cityu.edu.hk; {jwang,sfchang}@ee.columbia.edu

Abstract

Learning to cope with domain change has been known as a challenging problem in many real-world applications. This paper proposes a novel and efficient approach, named domain adaptive semantic diffusion (DASD), to exploit semantic context while considering the domain-shift-of-context for large scale video concept annotation. Starting with a large set of concept detectors, the proposed DASD refines the initial annotation results using graph diffusion technique, which preserves the consistency and smoothness of the annotation over a semantic graph. Different from the existing graph learning methods which capture relations among data samples, the semantic graph treats concepts as nodes and the concept affinities as the weights of edges. Particularly, the DASD approach is capable of simultaneously improving the annotation results and adapting the concept affinities to new test data. The adaptation provides a means to handle domain change between training and test data, which occurs very often in video annotation task. We conduct extensive experiments to improve annotation results of 374 concepts over 340 hours of videos from TRECVID 2005-2007 data sets. Results show consistent and significant performance gain over various baselines. In addition, the proposed approach is very efficient, completing DASD over 374 concepts within just 2 milliseconds for each video shot on a regular PC.

1. Introduction

Annotating large scale video data with semantic concepts has been a popular topic in computer vision and multimedia research in recent years [4, 5, 20]. The predefined concepts may cover a wide range of topics such as those related to objects (e.g., *car*, *airplane*), scenes (e.g., *mountain*, *desert*), events (e.g., *people_marching*) etc. The annotation of these concepts enables users to specify a query using a natural language description of the semantic content of interest. For example, an incoming textual query such as *find an airplane* could be efficiently handled by returning video shots with higher likelihood to contain concept *airplane*.

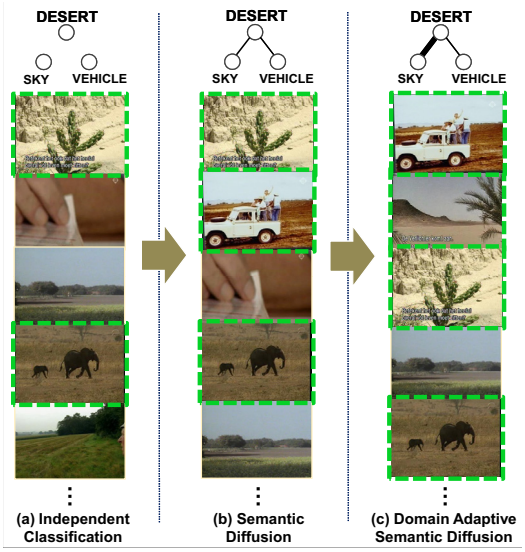


Figure 1. Illustration of context-based video annotation. (a) Top 5 video shots of concept *desert* according to the annotation scores from an existing pre-trained detector, in which the semantic context was not considered. (b) Refined shot list by semantic diffusion. The subgraph on the top shows two concepts with higher correlations to *desert*. Line width indicates graph edge weight. (c) Refined subgraph and shot list by the proposed domain adaptive semantic diffusion (DASD). The graph adaptation process in DASD is able to refine concept relationship which in turn can further help improve the annotation accuracy.

The existing studies in image and video annotation mainly aim at the assignment of single or multiple concept labels to a target data set, where the assignment is often done independently without considering the inter-concept relationship [4, 5, 7, 12]. Due to the fact that concepts do not occur in isolation (e.g., *smoke* and *explosion*), more research attentions have been paid recently for improving annotation accuracy by learning from semantic context [18, 24]. Nevertheless, the learning of contextual knowledge is often conducted in an offline manner based on train-

ing data, resulting in the classical problem of over or under fitting. For large scale video annotation which could involve simultaneous labeling of hundreds of concepts, the problem becomes worse when the unlabeled videos are from a domain different from that of the training data. This brings two challenges related to scalability: the need for adaptive learning and the demand for efficient annotation.

This paper proposes a novel and efficient approach for improving large scale video semantic annotation using graph diffusion technique. Our approach, named domain adaptive semantic diffusion (DASD), uses a graph diffusion formulation to enhance the consistency of concept annotation scores. First, we construct an undirected and weighted graph, namely semantic graph, to model the concept affinities. The graph is then applied to refine concept annotation results using a function level diffusion process. To handle the domain change problem, our approach further allows to simultaneously optimize the annotation results and adapt the geometry of the semantic graph according to the test data distribution. Figure 1 gives an idealized example of DASD. Figure 1 (a) displays the top 5 video shots from an independent concept detector *desert*. By considering the semantic context learnt offline from the manual annotations on training data, Figure 1 (b) shows better results. DASD, which is capable of adapting the already-learnt semantic context to fit the test domain data statistics, attains considerable improvement as shown in Figure 1 (c). In this example, DASD reacts to the domain-shift-of-context by adapting the affinity of *desert* and *sky*.

The advantage of the proposed DASD is twofold. First, it allows the online update of semantic context for addressing the problem of domain-shift. Second, it is scalable to large data sets where only a couple of minutes is required to complete DASD over hundreds of concepts for thousands of video shots.

In the following we review related works in Section 2. We then define notations and introduce the basic theory of graph diffusion in Section 3. The graph diffusion formulation for context-based video annotation is elaborated in Section 4 and the domain adaptive semantic diffusion is proposed in Section 5. Section 6 describes experimental setup and Section 7 presents our experimental results. Finally, Section 8 concludes this paper.

2. Related Works

Vision researchers have used context information to help object recognition. In [22], Torralba et al. introduced a framework of modeling context based on the correlation between the statistics of low-level features across the entire image and the objects that it contains. Along this line, several other approaches also adopted context information from the correlation of low level features within images or semantic categories [13, 19, 23, 25]. Recently, the semantic context such as co-occurrence information was considered to enforce region-based object recognition in [2, 18]. In addition to co-occurrence, relative location was utilized in [6] in which knowledge such as "*sky* usually appear on top of *grass*" was exploited to help label image regions. These approaches in [2, 6, 18], however, are tailored for region based object and scene recognition. The semantic concepts cover a wide range of topics – some of them are depicted by the holistic representation of an entire image rather than a region (e.g., *outdoor* and *meeting*). As a result, the region-based approaches, though promising, are not applicable in many cases of video annotation.

There are also a few research efforts focusing on the utilization of semantic context for video annotation [17, 24, 10]. In [17], a multi-label learning method derived from Gibbs random field was proposed to exploit concept relationship for improving annotation. Though encouraging results were observed on a set of 39 concepts, the complexity of this method is quadratic to the number of concepts. This prevents its application to a larger number of concepts, which is necessary to provide enough semantic filters for interpreting textual queries and producing satisfactory video search results [8]. In [24], Weng et al. proposed a method to learn the inter-concept relationships and then used graphical model to improve the concept annotation results. In [10], a context-based concept fusion method using conditional random field was proposed, in which supervised classifiers were iteratively trained to refine the annotation results. Different from these existing works, in DASD we formulate context-based video annotation as a highly efficient graph diffusion process. Particularly, it involves an adaptation procedure to handle domain changes between training and test data.

3. Preliminaries

Notations. We start by defining the notations used in this paper. Let $\mathcal{C} = \{c_1, c_2, ..., c_m\}$ be a semantic lexicon of m concepts and $\mathbf{X} = \{x_i\} \in R^{n \times d}$ be a data set, where n is the number of samples and d is the sample dimensionality. From a training set $\{\mathbf{X}_{trn}, \mathbf{Y}\}$, a supervised classifier can be trained for each concept c_i, where $\mathbf{Y}$ is the ground-truth label of $\mathbf{X}_{trn}$. The classifier is then applied to a test set $\mathbf{X}_{tst}$ of n test samples and generate annotation scores $g(c_i)$, where $g : \mathcal{C} \rightarrow R^n$ denotes an annotation function and $g(c_i)$ is a $1 \times n$ score vector. Concatenating the annotation scores of all the concepts for $\mathbf{X}_{tst}$, the video annotation function can be written as $g = \{g(c_i)\}_{i=1,\cdots,m} \in R^{m \times n}$.

Our goal in this paper is to utilize the semantic context in $\mathcal{C}$ to refine the annotation score:

$$\tilde{g} = f(g, \mathbf{W}), \tag{1}$$

where $\tilde{g}$ is the refined annotation function and g denotes the initial function based on the supervised classifiers; $\mathbf{W}$

is a concept affinity matrix indicating the concept relationship in $\mathcal{C}$, which can be estimated from the training set $\{\mathbf{X}_{trn}, \mathbf{Y}\}$; $f(\cdot)$ represents the refinement function, which simultaneously updates g and adapts $\mathbf{W}$ to the test set $\mathbf{X}_{tst}$ (cf. Section 5).

In addition, the undirected and weighted semantic graph used in this paper is denoted as $\mathcal{G} = (\mathcal{C}, E, \mathbf{W})$, comprising a set of nodes (concepts) together with a set $E = \{e_{ij}\}$ of edges. $\mathbf{W}$ is the concept affinity matrix, where each entry W_{ij} indicates the weight of an edge e_{ij} between nodes c_i and c_j. Define the diagonal node degree matrix as $D_{ii} = d(c_i) = \sum_j W_{ij}$. Then the graph Laplacian is $\mathbf{\Delta} = \mathbf{D} - \mathbf{W}$ and the normalized version is $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$.

Graph Diffusion. Graph diffusion analysis has been widely used for data smoothing and multi-scale image analysis. Some earlier applications on vision tasks include edge detection and denoising in images [15] and discontinuity detection in optical flow [16]. Moreover, graph diffusion is also closely related to several transductive learning methods [26, 27], where classification functions are estimated by investigating the geometry property of data distribution.

Generally speaking, function estimation through graph diffusion rests on the assumption that a studied function $g(\cdot)$ is expected to be smooth with respect to the manifold geometry of the discrete data [21]. Specifically, with a definition of the energy function $\mathcal{E}(g)$, the smoothed function $\tilde{g}$ can be derived using the gradient descent approach. As described in [3], the gradient of a function g on graph $\mathcal{G}$ is defined as:

$$(\nabla_g)(c_i, c_j) = W_{ij}\left(\frac{g(c_i)}{\sqrt{d(c_i)}} - \frac{g(c_j)}{\sqrt{d(c_j)}}\right). \quad (2)$$

Therefore, given the initial function value g, $\tilde{g}$ can be estimated by iteratively repeating the following one step diffusion process along the gradient directions:

$$g_t = g_{t-1} \pm \alpha \nabla_g, \quad (3)$$

where the coefficient $0 < \alpha \ll 1$ is step size. The symbol $+$ or $-$ depends on the objective of maximizing or minimizing the particular energy function $\mathcal{E}$.

4. Efficient Diffusion of Semantic Context

In this section, we formulate context-based video annotation as an efficient graph diffusion process. We start by presenting the construction of semantic graph and then describe its utilization for refining the annotation function g.

4.1. Semantic Graph

The semantic graph $\mathcal{G}$ is characterized by the relationship between concepts, i.e., the affinity matrix $\mathbf{W}$. We estimate the concept relationship using the training set $\mathbf{X}_{trn}$ and its corresponding label matrix $\mathbf{Y}$, where $y_{ij} = 1$ denotes the presence of concept c_i in the sample x_j, otherwise $y_{ij} = 0$. The concept relationship is computed using Pearson product moment correlation as

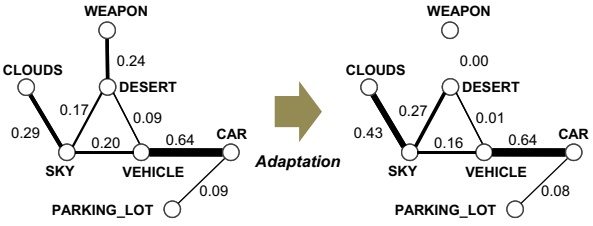
$$PM(c_i, c_j) = \frac{\sum_{k=1}^{|\mathbf{X}_{trn}|}(y_{ik} - \mu_i)(y_{jk} - \mu_j)}{(|\mathbf{X}_{trn}| - 1)\sigma_i\sigma_j}, \quad (4)$$

where μ_i and σ_i are the sample mean and standard deviation, respectively, of observing c_i in the training set $\mathbf{X}_{trn}$.

Since the correlation calculated by the above equation can be either negative or positive, we construct two positive-weighted semantic graphs accordingly:

- $\mathcal{G}^+ = (\mathcal{C}, E^+, \mathbf{W}^+)$ considers positive correlation of the concepts, i.e. an edge $e_{ij} \in E^+$ is established when $PM(c_i, c_j) > 0$ and $W^+_{ij} = PM(c_i, c_j)$;
- $\mathcal{G}^- = (\mathcal{C}, E^-, \mathbf{W}^-)$ considers negative correlation of the concepts, in which an edge $e_{ij} \in E^-$ is established when $PM(c_i, c_j) < 0$ and $W^-_{ij} = -PM(c_i, c_j)$.

We will discuss the semantic diffusion over $\mathcal{G}^-$ in the end of Section 5. Otherwise, in most part of the paper, we will focus only on $\mathcal{G}^+$ and use $\mathcal{G}^+$ and $\mathcal{G}$ interchangeably without specific declaration. Figure 2 (a) visualizes a fraction of the semantic graph $\mathcal{G}^+$.



(a) Broadcast News Videos (b) Documentary Videos

Figure 2. A fraction of the semantic graph before and after domain adaptation. Thick edges represent strong correlation between concepts, quantified by the values nearby the edges. (a) The initial concept relationship computed using the manual annotations on TRECVID 2005 development set; (b) The updated concept relationship for TRECVID 2007 test set after performing domain adaptation (more explanations in Section 7).

4.2. Semantic Diffusion

Recall that the function value $g(c_i) \in \mathcal{R}^n$ on a semantic node c_i denotes the concept annotation scores on test set $\mathbf{X}_{tst}$, and $n = |\mathbf{X}_{tst}|$ is the number of test samples. Intuitively, the function values $g(c_i)$ and $g(c_j)$ should be consistent with the affinity between concepts c_i and c_j, i.e. W_{ij}. In other words, strongly correlated concepts should have similar concept annotation scores. Motivated by this semantic consistency, here we formulate our problem as a graph diffusion process and define a cost function on the semantic graph as:

$$\mathcal{E}(g) = \frac{1}{2}\sum_{i,j=1}^{m} W_{ij}||\frac{g(c_i)}{\sqrt{d(c_i)}} - \frac{g(c_j)}{\sqrt{d(c_j)}}||^2. \quad (5)$$

Apparently, this cost function evaluates the smoothness of the function g over semantic graph $\mathcal{G}$. Therefore, minimizing $\mathcal{E}$ makes the annotation results more consistent with the concept relationships, which are captured by $\mathcal{G}$.

We apply gradient descending to gradually reduce the value of the cost function. Rewrite the Equation 5 into matrix formulation as:

$$\mathcal{E}(g) = \frac{1}{2}\mathrm{tr}(g^T \mathbf{L} g), \tag{6}$$

where $\mathbf{L} \in R^{m \times m}$ is the graph Laplacian of the semantic graph $\mathcal{G}$. From Equation 2, the gradient of $\mathcal{E}$ with respect to g on the semantic graph is:

$$\nabla_g \mathcal{E} = \mathbf{L} g. \tag{7}$$

Thus we can derive the following iterative diffusion process

$$\begin{aligned} g_t &= g_{t-1} - \alpha \nabla_{g_{t-1}} \mathcal{E} = g_{t-1} - \alpha \mathbf{L} g_{t-1} \\ &= (\mathbf{I} - \alpha \mathbf{L}) g_{t-1} = (\mathbf{I} - \alpha \mathbf{L})^2 g_{t-2} \\ &= \cdots = (\mathbf{I} - \alpha \mathbf{L})^t g_0. \end{aligned} \tag{8}$$

Through exponentiating the graph Laplacian with step size α, we can get

$$\begin{aligned} g_t &= \left(\mathbf{I} - t(\alpha \mathbf{L}) + \frac{t^2}{2!}(\alpha \mathbf{L})^2 - \frac{t^3}{3!}(\alpha \mathbf{L})^3 + \cdots \right) g_0 \\ &\approx \left(\mathbf{I} - t(\alpha \mathbf{L}) + \frac{1}{2}(\alpha \mathbf{L} t)^2 - \frac{1}{6}(\alpha \mathbf{L} t)^3 \right) g_0. \end{aligned} \tag{9}$$

Omitting the high order terms $\mathcal{O}((\alpha \mathbf{L} t)^3)$, the above cubic form approximates the exponential diffusion procedure. Instead of iterative diffusion on the initial function value g_0, Equations 8 and 9 give the one step close form diffusion operation through applying the diffusion kernel $\mathcal{K}_t$ on g_0, which is defined as:

$$\mathcal{K}_t = (\mathbf{I} - \alpha \mathbf{L})^t. \tag{10}$$

Notice that the cost function in Equation 5 has two main fundamental differences from the existing graph-based semi-supervised learning (SSL) techniques such as [27, 26]. First, the semantic graph is formed with concepts as nodes, and consistency is defined in the concept space. This is in contrast to graph-based SSL where a node is a data sample and smoothness is thus measured in the feature space. Second, with given label information, graph-based SSL methods drive label propagation through minimizing the cost function with either elastic regularization or strict constraint, e.g., the harmonic function formulation in [27]. Our cost function aims at recovering the consistency of annotation scores with respect to the semantic graph. It is minimized using gradient descending and this leads to a close form solution for efficient semantic diffusion. While in graph-based SSL methods, the optimization procedure commonly involves an expensive matrix inverse operation. For large scale applications, compact representation and efficient optimization are always critical and our formulation takes into account both factors.

5. Domain Adaptive Semantic Diffusion

A challenge tightly correlates with scalability issue is the problem of over and under fitting. Specifically, the learnt semantic graph is not expected to completely capture the context relationship of unseen video data. This can happen particularly if there is lack of sufficient training samples or the unseen data has been shifted to another domain different from the training samples. For the latter case, this can imply a real-word problem that the semantic graph learnt from one domain (e.g., broadcast news videos) cannot always properly reflect the semantic context in another domain (e.g., documentary videos). For example, the concept *weapon* always co-occurs with *desert* in news videos due to plenty of events about Iraq war. While such context relationship can be captured in $\mathcal{G}$, misleading annotation will be generated if applying to documentary videos where such relationship is seldom observed. In this section, we address this problem by proposing the DASD algorithm. DASD captures the test domain knowledge by learning from the responses of g over the new and previously unseen data. The initial semantic graph learnt from training samples is online adapted to fit the new knowledge mined from test data.

DASD combines both semantic diffusion and graph adaptation by minimizing the cost function via alternatively updating g and the concept affinity. Notice that the symmetric affinity matrix $\mathbf{W}$ indirectly imposes on the diffusion procedure in the form of normalized graph Laplacian $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} = \mathbf{I} - \tilde{\mathbf{W}}$. Recall the cost function in Equation 6 and express it as a function of $\tilde{\mathbf{W}}$:

$$\begin{aligned} \mathcal{E}(g, \tilde{\mathbf{W}}) &= \frac{1}{2}\mathrm{tr}(g^T \mathbf{L} g) = \frac{1}{2}\mathrm{tr}\left(g^T \left[\mathbf{I} - \mathbf{D}^{-\frac{1}{2}} \mathbf{W} \mathbf{D}^{-\frac{1}{2}} \right] g \right) \\ &= \frac{1}{2}\mathrm{tr}(g^T g) - \frac{1}{2}\mathrm{tr}\left(g^T \mathbf{D}^{-\frac{1}{2}} \mathbf{W} \mathbf{D}^{-\frac{1}{2}} g \right) \\ &= \frac{1}{2}\mathrm{tr}(g^T g) - \frac{1}{2}\mathrm{tr}\left(g^T \tilde{\mathbf{W}} g \right), \end{aligned} \tag{11}$$

where $\tilde{\mathbf{W}}$ is the normalized affinity matrix. Although our goal is to adapt the concept affinity matrix $\mathbf{W}$, in the diffusion process $\tilde{\mathbf{W}}$ directly affects the annotation results. Hence, we compute the partial differential of $\mathcal{E}$ with respect to $\tilde{\mathbf{W}}$ instead of $\mathbf{W}$ as:

$$\frac{\partial \mathcal{E}}{\partial \tilde{\mathbf{W}}} = -g g^T. \tag{12}$$

Similar to the Section 4.2, we use the gradient method to iteratively modify the normalized affinity matrix $\tilde{\mathbf{W}}$:

$$\tilde{\mathbf{W}}_t = \tilde{\mathbf{W}}_{t-1} - \beta \frac{\partial \mathcal{E}}{\partial \tilde{\mathbf{W}}_{t-1}} = \tilde{\mathbf{W}}_{t-1} + \beta g_{t-1} g_{t-1}^T, \tag{13}$$

Figure 3. Example keyframes of several concepts evaluated in TRECVID 2005–2007. The left shows examples from broadcast news videos in 2005 and 2006, while the right shows examples from documentary videos in 2007. Note that the appearance of the same concept from the two data domains may be visually very different.

where β ($0 < \beta \ll 1$) is the step size for gradient search.

The adaptation process of $\tilde{\mathbf{W}}$, executed in Equation 13 with the aim of minimizing $\mathcal{E}$, can be explained intuitively as follows. Recall that $g \in R^{m \times n}$, where m is the number of concepts and n is the number of test samples. The dot product $gg^T \in R^{m \times m}$ in the above equation implies the pairwise concept affinities estimated by the annotation scores in the test domain (each row of g is normalized to unit length). Thus the above equation implicitly and gradually incorporates the new domain knowledge into $\tilde{\mathbf{W}}$.

To combine the graph adaptation process and the semantic diffusion described in Section 4.2, we update the normalized graph Laplacian as:

$$\begin{aligned} \mathbf{L}_t &= \mathbf{I} - \tilde{\mathbf{W}}_\mathbf{t} = \mathbf{I} - \tilde{\mathbf{W}}_{t-1} - \beta g_{t-1} g_{t-1}^T \\ &= \mathbf{L}_{t-1} - \beta g_{t-1} g_{t-1}^T. \end{aligned} \quad (14)$$

Now we can derive the following iterative alternating optimization procedure:

$$\begin{aligned} g_t &= (\mathbf{I} - \alpha \mathbf{L}_{t-1}) g_{t-1}, \\ \mathbf{L}_t &= \mathbf{L}_{t-1} - \beta g_{t-1} g_{t-1}^T. \end{aligned} \quad (15)$$

The above two equations form the DASD process, which jointly imposes the semantic diffusion of annotation scores and the adaptation of semantic graph structure.

All the above derivation is based on the positive graph $\mathcal{G}^+$. The diffusion and adaptation on $\mathcal{G}^-$ can be done in a similar manner as $\mathcal{G}^+$. Because in $\mathcal{G}^-$ the edge weights hint the dissimilarity between the nodes (concepts), here the graph diffusion is casted in the way of maximizing the cost function as:

$$(g^*, (\tilde{\mathbf{W}}^-)^*) = \arg \max_{g, \tilde{\mathbf{W}}^-} \mathcal{E}. \quad (16)$$

Repeat the similar derivations as before, we can obtain the following DASD process on $\mathcal{G}^-$:

$$\begin{aligned} g_t &= (\mathbf{I} + \alpha \mathbf{L}_{t-1}^-) g_{t-1}, \\ \mathbf{L}_t^- &= \mathbf{L}_{t-1}^- + \beta g_{t-1} g_{t-1}^T. \end{aligned} \quad (17)$$

TRECVID-	Data domain	Development set	Test set
2005	Broadcast News	80h (43,873)	80h (45,765)
2006	Broadcast News	–	80h (79,484)
2007	Documentary	50h (21,532)	50h (22,084)

Table 1. Descriptions of TRECVID 2005–2007 data sets. The total number of video shots in each data set is shown in the parenthesis. Note that the 160h data from TRECVID 2005 was used as development data for TRECVID 2006.

6. Experimental Setup

We conduct experiments using the TRECVID 2005–2007 video data sets. The data sets were used in the annual TRECVID benchmark evaluation by NIST [20]. In total, there are 340 hours of video data. The videos are partitioned into shots and one or more representative keyframes are extracted from each shot. As shown in Table 1, the 2005 and 2006 videos are broadcast news from different TV programs in English, Chinese and Arabic, while the 2007 data set consists mainly of documentary videos in Dutch. These data sets are suitable for evaluating the performance in handling domain changes. The contents of the videos are also highly diversified, making large scale video annotation a challenging task.

We use a total of 374 concepts defined in LSCOM [14] for constructing the semantic graph. These concepts are defined by LSCOM according to criteria such as concept utility, observability and the feasibility of developing classifiers for them using current technologies. Based on LSCOM, NIST selected 10-20 concepts each year and provided ground-truth for performance evaluation. Figure 3 shows example keyframes for several concepts evaluated by NIST. Note that this is a multi-labeling task, meaning that each shot can be labeled with more than one concept.

In our experiments, for TRECVID 2005, we adopt the development set as our target database and report performance of 39 concepts for the ease of comparison with other existing works such as [10]. The development set is partitioned into training, validation and test sets. For TRECVID 2006 and 2007, we report performance of the official evaluated concepts on each year's test set.

Baseline Detector and Graph Construction. To test the proposed method over a large concept pool, we adopt some generic semantic concept detectors that can be easily deployed. For this, we use the publicly available VIREO-374[1] [11] as baseline, which includes SVM models of 374 LSCOM concepts. These models have been shown in TRECVID evaluations to achieve top performance. Late fusion is used to combine classifier scores from multiple single-feature SVMs, which are trained using three individual features: grid-based color moments, wavelet texture, and bag-of-visual-words. Details for extracting such features from video keyframes can be found in [11]. However, it is worth noting that the proposed context-based video annotation framework is applicable to general concept detectors using different features, including motion. Later in the paper we will show the effectiveness of the proposed technique over different baseline classification models.

The semantic graph $\mathcal{G}$ is constructed based on the ground-truth labels on the 2005 development set, where the edge weights are calculated by Equation 4. In practice, k-NN is commonly used to generate sparse graphs [9]. In our experiments, we empirically keep 6 strongest edges for each semantic node and break the remaining connections. During graph adaptation, in order to keep the graph sparse, we round small gradients in the partial differential in Equation 12 to zero and only keep the largest gradient for each node.

Evaluation Criteria. Following the TRECVID evaluation, for each semantic concept, we use average precision (AP) to evaluate the performance on TRECVID 2005, and use inferred AP for TRECVID 2006 and 2007. AP approximates the area under the precision-recall curve. The inferred AP is an approximation of the AP. It is designed for partially labeled data sets (TRECVID 2006 and 2007 test sets) in order to reduce manual labeling effort [20]. To aggregate the performance over multiple semantic concepts, mean AP or mean inferred AP (MAP) is adopted.

7. Results and Comparison

This section describes our experimental results and gives comparative studies with the state of the arts. In most of the experiments we only use the semantic graph $\mathcal{G}^+$. $\mathcal{G}^-$ is only used in one experiment to study the effect of negative correlation. There are some parameters in the proposed method, such as the step sizes α, β, and the number of iterations (diffusion time t). We empirically determine their suitable values, and will show the insensitivity of final performance to particular settings.

Table 2 shows the results over the TRECVID 2005-2007 data sets, achieved by the VIREO-374 baseline, the semantic diffusion (SD, as described in Section 4.2), and the DASD. When SD is used, the performance gain (relative improvement) on TRECVID 2005-2007 data sets ranges from 11.8% to 15.6%. The results confirm the effectiveness of formulating graph diffusion for improving video annotation accuracy. Figure 4 shows per-concept performance of the 20 evaluated concepts in TRECVID 2006. Our approach consistently improves all the concepts. Among a total of 79 concepts from TRECVID 2005 to 2007 as reported in Table 2, almost all concepts show improvement, except five concepts which suffer from slight performance degradation.

[1]Download site: http://vireo.cs.cityu.edu.hk/research/vireo374/

TRECVID-	2005	2006	2007
# of evaluated concepts	39	20	20
Baseline (MAP)	0.166	0.154	0.099
SD	11.8%	15.6%	12.1%
DASD	11.9%	17.5%	16.2%

Table 2. Overall performance gain (relative improvement) on TRECVID 2005–2007 data sets. SD: semantic diffusion. DASD: domain adaptive semantic diffusion.

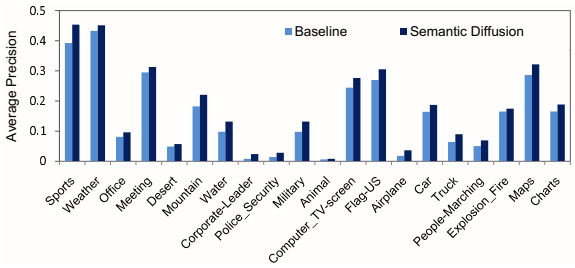


Figure 4. Per-concept performance before and after semantic diffusion on TRECVID 2006 test set. Consistent improvements are observed for all of the 20 semantic concepts.

Due to change of video domains from news to documentary, the semantic graph $\mathcal{G}$ constructed from TRECVID 2005 obviously does not fit TRECVID 2007, which requires the adaptation of the semantic graph. As shown in Table 2, the DASD further boosts the performance on TRECVID 2006 and 2007. It is easy to explain that there is basically no improvement on TRECVID 2005 since the semantic graph is constructed on the same data set. Note that although both TRECVID 2005 and 2006 data sets are broadcast news videos, they were captured in different years so that the video content may change a lot. We consider the graph adaptation process as an important merit of our approach: it can automatically refine the semantic geometry to fit the test data, which will in turn help improve the video annotation performance. Figure 2 shows a fraction of the semantic graph, from which we have a few observations – the adaptation process enhances the affinity of *sky* and *clouds*, and breaks the edge between *desert* and *weapon*. The concept *weapon* frequently co-occurs with *desert* scene in TRECVID 2005 broadcast news videos because there are many events about Iraq war, while in the documentary videos, this is seldom observed.

Effect of Parameters. Figure 5 shows the MAP over TRECVID 2006 test set using a range of step sizes and dif-

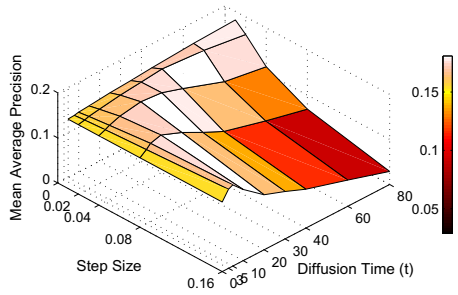


Figure 5. The MAP performance on TRECVID 2006 test set under various parameter settings.

fusion time values. As can be seen, there is a tradeoff between the values of t (number of iterations, namely, diffusion time) and the gradient decent step sizes (α and β). The finer step sizes are used, the more iterations are needed to reach the best diffusion performance. Interestingly, our empirical results indicate that different step sizes, when combined with corresponding best diffusion times, result in consistent performances.

We also evaluated the same parameter setting (e.g., α and $\beta = 0.04$) over different data sets, TRECVID 2005–2007, and verified that the optimal diffusion time $t = 20$ consistently achieves the best or close to best performances. These findings confirm the performance stability of the proposed method over parameter settings. The same parameters (α, $\beta = 0.04$; t=20) are used throughout the experiments in this paper.

Effect of Negative Correlation. In order to study the effect of negative correlations, we conduct another experiment on TRECVID 2007 test set. The performance gain is merely 1.3% when using the negative graph $\mathcal{G}^-$ alone (Equation 17). When both $\mathcal{G}^+$ and $\mathcal{G}^-$ are used (alternatively apply Equations 15 and 17), the MAP performance is 0.114, which is about the same with that using $\mathcal{G}^+$ alone (0.115). Based on these results we speculate that although negative correlations captured by $\mathcal{G}^-$ may slightly improve the performance, practically using $\mathcal{G}^+$ alone is preferred since it represents a good tradeoff between performance and speed.

Effect of Concept Affinity Estimation Method. In the above experiments the concept affinities are computed based on the ground-truth labels. An alternative way is to estimate the concept affinities according to the baseline annotation scores on each year's test data. Let $\mathcal{T}$ be the test data set and g_k^i be the baseline annotation scores of concept c_i in test shot k. Similar to Equation 4, the weight w_{ij} of the edge (c_i, c_j) can be calculated as $w_{ij} = \frac{\sum_{k=1}^{|\mathcal{T}|}(g_k^i - \mu_i)(g_k^j - \mu_j)}{(|\mathcal{T}|-1)\sigma_i\sigma_j}$, where μ_i and σ_i are the mean and standard deviation of the annotation scores of c_i, respectively, in $\mathcal{T}$. With the new edge weights, we only consider positive correlations and construct a new graph $\mathcal{G}_\mathcal{T}$.

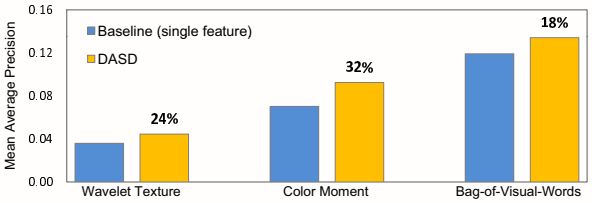


Figure 6. Performance of DASD using various baselines on TRECVID 2006 test set. The text under blue and orange bars indicates the feature used in the corresponding SVM detector.

	Jiang et al. [10]	Aytar et al. [1]	Weng et al. [24]	DASD
2005	2.2%	4.0%	N/A	11.9%
2006	N/A	N/A	16.7%	17.5%

Table 3. Performance comparison of DASD with several existing works.

We rerun the semantic diffusion on TRECVID 2007 test set using $\mathcal{G}_\mathcal{T}$. The overall performance gain is 9.3%, which is lower compared to 12.1% when using $\mathcal{G}$ derived from the training set. Nevertheless, this process is more economic than the construction of $\mathcal{G}$ using fully labeled training set. Manual labels are difficult to obtain in practice, especially when the number of concepts is in the order of thousands. Thus, constructing graph based on initial annotation scores is a promising way when a fully labeled training set such as TRECVID 2005 is unavailable.

Effect of Baseline Performance. We now evaluate the sensitivity of our approach to the performance of baseline detectors. As mentioned in Section 6, instead of using fused output of the three SVMs as baseline, here we use the prediction output of each single (and relatively weak) SVM as initial value for function $g(\cdot)$. The proposed DASD is then applied to these three weak baselines respectively. The results on TRECVID 2006 test set are shown in Figure 6. Apparently, the MAPs of all these weak detectors are consistently and steadily improved with quite high performance gain. Particularly, it improves the baseline of wavelet texture (MAP is just 0.036) by 24% (the left orange bar). From this experiment, we can conclude that the proposed DASD is able to achieve consistently better performance over various baseline detectors, even for some fairly weak ones.

Comparison with the State of the Art. We compare our approach to three existing works in [1, 10, 24] for context-based video annotation. As shown in Table 3, on TRECVID 2005, our approach is able to improve 11.9%. With similar experimental settings, Jiang et al. reported performance gains of 2.2% over all 39 concepts and 6.8% over 26 selected concepts [10], and Aytar et al. improved 4% [1]. The most recent work on utilizing semantic context for video annotation is [24], in which a performance gain of 16.7% over the same VIREO-374 baseline was reported on TRECVID 2006 test set. However, techniques in [24] did not show the domain adaptation ability and the parameters were op-

TRECVID-	2005	2006	2007
SD	59s	84s	12s
DASD	89s	165s	28s

Table 4. Run time of SD and DASD on TRECVID 2005–2007 data sets. The experiments are conducted on a Intel Core 2 Duo 2.2GHz PC with 2G RAM.

timized separately for each of the concepts. In the contrast, our approach adapts the concept affinity and uses uniform parameter setting for all the concepts. We demonstrated a performance gain of 17.5%, which to the best of our knowledge is the highest reported improvement on exploiting semantic context for video annotation.

Run Time. Our approach is extremely efficient. Constructing the semantic graph $\mathcal{G}$ using TRECVID 2005 development set takes 284 seconds. The complexity of the DASD algorithm is $O(mn)$, where m is the number of concepts and n is the number of video shots. Table 4 listed the detailed run time on each data set. We see that on the TRECVID 2006 data set which contains 79,484 video shots, the DASD algorithm finishes in just 165 seconds. In other words, running the DASD over the 374 concepts for each video shot only takes 2 milliseconds. This is much faster than the existing works in [10, 17, 24], in which tens or hundreds of hours are required due to the expensive training process involved in their approaches.

8. Conclusion

We have presented a novel and efficient approach, named DASD, to exploit semantic context for improving large scale video annotation accuracy. The semantic context is modeled in an undirected and weighted concept graph, which is then used to recover the consistency and smoothness of video annotation results via a function level graph diffusion process. The extensive experiments on 374 semantic concepts over 340 hours of video data show that the semantic context is powerful for enhancing video annotation accuracy and the proposed DASD algorithm consistently and significantly improves the performance over the vast majority of the evaluated concepts. The proposed DASD algorithm is able to adapt the concept affinity to test data from a different domain. The experimental results confirm that this adaptive approach can alleviate the domain-shift-of-context problem and show further improvement of the video annotation accuracy. In addition, we demonstrate that the semantic graph can be bootstrapped using initial annotation results from individual classifiers, which is very helpful when an exhaustively labeled training set is not available to construct the semantic graph.

Acknowledgement

This material is based upon work supported in part by a DARPA grant under Contract No. HR0011-08-C-0135, a NSF grant CNS-0751078, and a grant from City University of Hong Kong (Project No. 7002241).

References

[1] Y. Aytar, O. B. Orhan, and M. Shah. Improving semantic concept detection and retrieval using contextual estimates. In *ICME*, 2007.
[2] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *ECCV*, 2004.
[3] F. Chung. *Spectral Graph Theory*. American Mathmatical Society, 1997.
[4] M. Everingham, A. Zisserman, C. Williams, and L. V. Gool. The pascal visual object classes challenge 2006 (voc 2006) results. In *University of Oxford Technical Report*, 2006.
[5] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *CVPR Workshop on Generative-Model Based Vision*, 2004.
[6] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, 2008.
[7] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 2005.
[8] A. Hauptmann, R. Yan, and W.-H. Lin. How many high level concepts will fill the semantic gap in video retrieval? In *CIVR*, 2007.
[9] T. Jebara, J. Wang, and S.-F. Chang. Graph construction and b-matching for semi-supervised learning. In *ICML*, 2009.
[10] W. Jiang, S. F. Chang, and A. Loui. Context-based concept fusion with boosted conditional random fields. In *ICASSP*, 2007.
[11] Y. G. Jiang, C. W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *CIVR*, 2007.
[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
[13] K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the tree: a graphical model relating features, objects and the scenes. In *NIPS*, 2003.
[14] 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.
[15] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. In *PAMI*, 1990.
[16] M. Proesmans, L. J. VanGool, E. Pauwels, and A. Oosterlinck. Determination of optical flow and its discontinuities using nonlinear diffusion. In *ECCV*, 1994.
[17] G. J. Qi, X. S. Hua, Y. Rui, J. Tang, T. Mei, and H. J. Zhang. Correlative multi-label video annotation. In *ACM Multimedia*, 2007.
[18] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007.
[19] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling appearance, shape and context. In *IJCV*, 2007.
[20] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *ACM MIR*, 2006.
[21] A. D. Szlam, M. Maggioni, and R. R. Coifman. Regularization on graphs with function-adapted diffusion processes. *Journal of Machine Learning Research*, 9:1711–1739, 2008.
[22] A. Torralba. Contextual priming for object detection. In *IJCV*, 2003.
[23] A. Torralba, K. Murphy, W. Freeman, and M. Rubin. Context-based vision system for place and object recognition. In *CVPR*, 2003.
[24] M.-F. Weng and Y.-Y. Chuang. Multi-cue fusion for semantic video indexing. In *ACM Multimedia*, 2008.
[25] L. Wolf and S. Bileschi. A critical view of context. In *IJCV*, 2006.
[26] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In *NIPS*, 2004.
[27] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 2003.