6-2012

Fast semantic diffusion for large-scale context-based image and video annotation

Yu-Gang JIANG

Qi DAI

Jun WANG

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

Citation

1

Fast Semantic Diffusion for Large-Scale Context-Based Image and Video Annotation

Yu-Gang Jiang, Qi Dai, Jun Wang, *Member, IEEE*, Chong-Wah Ngo, *Member, IEEE*, Xiangyang Xue, *Member, IEEE*, and Shih-Fu Chang, *Fellow, IEEE*

***Abstract*—Exploring context information for visual recognition has recently received significant research attention. This paper proposes a novel and highly efficient approach, which is named semantic diffusion, to utilize semantic context for large-scale image and video annotation. Starting from the initial annotation of a large number of semantic concepts (categories), obtained by either machine learning or manual tagging, the proposed approach refines the results using a graph diffusion technique, which recovers the consistency and smoothness of the annotations over a semantic graph. Different from the existing graph-based learning methods that model relations among data samples, the semantic graph captures context by treating the concepts as nodes and the concept affinities as the weights of edges. In particular, our approach is capable of simultaneously improving annotation accuracy and adapting the concept affinities to new test data. The adaptation provides a means to handle domain change between training and test data, which often occurs in practice. Extensive experiments are conducted to improve concept annotation results using Flickr images and TV program videos. Results show consistent and significant performance gain (10+% on both image and video data sets). Source codes of the proposed algorithms are available online.**

***Index Terms*—Context, image and video annotation, semantic concept, semantic diffusion (SD).**

I. Introduction

SEMANTIC annotation of large-scale image and video data has been an important research topic [2]–[7]. The annotation can act as an intermediate step for a wide range of applications such as image/video content management and retrieval. For example, once the image/video data are indexed by a large number of semantic concepts, users may efficiently locate the content of interest by using textual queries [8].

In the study of image and video annotation, most of the existing works assign single or multiple concept (class) labels to a target data set using supervised learning techniques, where the assignment is often independently done without considering the interconcept relationship [2], [3], [4], [6], [9], [10]. However, the concepts do not occur in isolation (e.g., *smoke* and *explosion*)—knowing the presence of one concept may provide helpful contextual clues for annotating other correlated concepts. Motivated by this intuition, several recent research efforts have been paid for improving annotation accuracy by exploiting context information, e.g., [11] and [12].

Existing methods for exploring context knowledge, however, were usually based on computationally expensive machine learning algorithms [11], [12], limiting their application to large data sets. In addition, since context (e.g., concept relationships) is usually learnt from a fully labeled training set, it may not be accurate when test data are from a domain different from that of the training data, which often occurs in large-scale applications. This poses two challenges related to scalability: the demand for efficient contextual annotation and the need for domain adaptive learning.

In this paper, we are interested in the problem of efficiently refining large-scale image and video annotation by exploring semantic context, i.e., the interconcept relationship. Our approach, which is named semantic diffusion (SD), uses a graph diffusion formulation to enhance the consistency of concept annotation scores. The input of our approach can be either soft prediction scores of individual supervised concept classifiers [e.g., the support vector machines (SVMs)] or binary manual labels tagged by Web users. To model the concept relationships, we first construct a weighted graph, which is named semantic graph, where nodes are the concepts, and edge weights reflect concept correlation. The graph is then applied in SD to refine concept annotation results using a function-level diffusion process to recover the consistency of the annotation scores with respect to the concept affinities.

In order to alleviate the effect of data domain changes, we further extend SD to simultaneously optimize the annotation results and adapt the geometry of the semantic graph according to the test data distribution. We name this augmented approach as domain adaptive SD (DASD). Fig. 1 gives an idealized example of our approach. Fig. 1(a) displays the top five results from an independent concept detector of *desert*. By considering the semantic context learnt offline from the manual annotations in a training data set, Fig. 1(b) shows better results. With the domain

Manuscript received September 12, 2011; revised January 23, 2012; accepted February 04, 2012. Date of publication February 15, 2012; date of current version May 11, 2012. This work was supported in part by the National 973 Program of China under Grant 2010CB327906, by the U.S. National Science Foundation under Grant CNS-0751078, and by the Research Grants Council of the Hong Kong Special Administrative Region of China under Grant CityU 119709. The associate editor coordinating the review of this paper and approving it for publication was Dr. Chun-Shien Lu.

Y.-G. Jiang, Q. Dai, and X. Xue are with the School of Computer Science, Fudan University, Shanghai 200433, China (e-mail: ygj@fudan.edu.cn; xyxue@fudan.edu.cn).

J. Wang is with IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: wangjun@us.ibm.com).

C.-W. Ngo is with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong (e-mail: cwngo@cs.cityu.edu.hk).

S.-F. Chang is with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: sfchang@ee.columbia.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2012.2188038

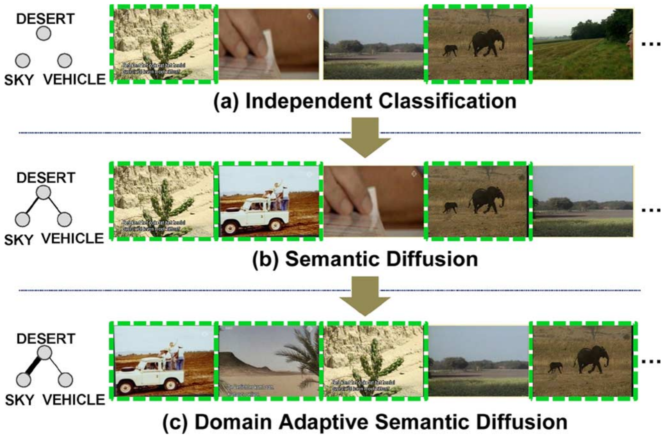


Fig. 1. Illustration of the proposed SD approach. (a) Top five results of concept *desert* according to the annotation scores from an existing individual classifier, where the semantic context was not considered. (b) Refined results by SD. The subgraph on the left shows two concepts with higher correlations to *desert*. Line width indicates graph edge weight. (c) Refined subgraph and top results by DASD. The graph adaptation process in DASD is able to refine concept relationship, which, in turn, can further improve the annotation accuracy.

adaptation extension, considerable improvement could be further obtained [see Fig. 1(c)]. The domain adaptation process is capable of adapting the already-learnt semantic context to fit the test domain data characteristics. In this example, it reacts the domain-shift-of-context by adapting the affinity of *desert* and *sky*.

In addition to offering better performance compared with existing alternatives for context-based annotation, our proposed approach holds the following advantages. First, the concept relationships are novelly encapsulated into a semantic graph, based on which both SD and DASD, derived from graph diffusion techniques, are extremely efficient. It requires less than 1 ms to perform diffusion over tens of concepts for each image or video shot. Second, our approach is able to perform batch-mode operation for a large data set that cannot fit into memory and, therefore, is directly applicable to large-scale applications. Third, it allows an online update of semantic context for alleviating the challenging problem of data domain changes. Additionally, we also investigate the effect of contextual annotation quality on SD result by simulating contextual annotations at various performance levels. This study is helpful for learning the upper limit of contexts in visual annotation and provides insights to answer an open question of “how good are contextual detectors good enough?”.

The remainder of this paper is organized as follows. We review related works in Section II. We then define notations and briefly introduce the graph diffusion technique in Section III. Our proposed formulation for context-based annotation is elaborated in Section IV. Section V describes the experimental setup, and Section VI presents our experimental results on image and video annotation. Finally, we conclude in Section VII.

II. Related Works

Recent research has yielded many techniques for image and video annotation, where the main effort has been paid to developing effective feature representation [13]–[16] and model learning techniques [5], [9], [10], [17]–[19]. In the following, we concentrate our review on the use of context information for improved annotation, which is the focus of this paper.

Contextual cues have been utilized for object and scene classification. In [20], Torralba introduced a framework of modeling context based on the correlations between the statistics of low-level features across an 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 [21]–[24]. Recently, semantic context such as cooccurrence information was considered to enforce region-based object recognition in [11], [25], and [26], where a graphical model was used to achieve the goal. In addition to cooccurrence, spatial context was utilized in [27] and [28], where knowledge such as “*sky* usually appears on top of *grass*” was exploited to help label image regions. In [29], contextual clues were modeled in a hierarchical region tree for scene understanding. These approaches in [11] and [25]–[29], however, were tailored for region-based object and scene recognition, where the region segmentation process is usually slow. In real-life examples, the concepts could cover a wide range of semantic 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, although promising, are not applicable in many cases of semantic annotation.

There are also several works focusing on learning context models from entire images implicitly [10], [13] or explicitly [12], [30]–[34]. The GIST feature by Oliva and Torralba [13] models the holistic energy distribution of a scene, and the spatial pyramid representation by Lazebnik *et al.* [10] also generates a holistic image description. These representations implicitly capture the semantic context to some extent. The authors of [7], [12], [30], [31], and [33]–[35] explicitly modeled the relationship of semantic categories and further improved upon the individual classifiers based on the holistic image descriptors. The proposed work of this paper belongs to this category. In [30], Lin *et al.* proposed a probabilistic model to recognize people, event, and location by exploring contextual knowledge in the three groups of categories. In [35], Song *et al.* proposed an approach to explore context from object detection/localization for general object recognition tasks in images. This approach, although very interesting, is too slow for large-scale applications since detecting objects in images is always a time-consuming procedure. In [31], Rasiwasia and Vasconcelos constructed a semantic space to model semantic context using mixtures of Dirichlet distributions. Based on their contextual models, an image can be represented by a vector of posterior concept probabilities. In [12], a multilabel learning method derived from Gibbs random field was proposed to exploit concept relationship for improving video annotation. Although encouraging results were observed on a set of 39 semantic concepts, the complexity of this method is quadratic to the number of concepts. A recent work by Lu *et al.* [7] used a contextual kernel method, which has a cubic complexity. The high complexity prevents both works to be applied to a larger number of concepts, which is necessary in practice, in order to provide enough semantic filters for interpreting textual queries and producing satisfactory search results [8]. In [32], Tang *et al.* explored concept relationship context with graph-based semisupervised learning (SSL) for improved concept annotation. In [33], Weng and Chuang proposed a method to learn the interconcept relationships and then used a graphical model to improve the concept annotation results. In [34], a context-based concept fusion method using conditional random field was proposed, in which supervised classifiers were iteratively trained to refine the annotation results.

In this paper, we address the problem of context-based refinement of individual concept annotation results. Different from all the existing works, we formulate context-based semantic annotation as a simple and highly efficient graph diffusion process. Our formulation has a linear complexity to the number of testing samples (images or video clips). Particularly, it is also capable of handling potential domain changes of semantic context between training and test data, which has not been investigated in the prior works. This paper extends upon a previous conference publication [36]. The extensions include new experiments on Flickr images, analysis on the power of context for visual annotation, and amplified discussions and explanations throughout the paper.

III. Preliminaries

A. Notations

Let us first define several notations. Let $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$ be a semantic lexicon of m concepts and $\mathbf{X}_{\text{tst}} = \{x_i\} \in R^{n \times d}$ be a test data set, where n is the number of samples, and d is the dimensionality of sample feature. From a training set $\{\mathbf{X}_{\text{trn}}, \mathbf{Y}\}$, a supervised classifier can be trained for each concept c_i, where $\mathbf{y}_i$, a column vector of $\mathbf{Y}$, is ground-truth labels of $\mathbf{X}_{\text{trn}}$ for concept c_i. The classifier is then applied to the test set $\mathbf{X}_{\text{tst}}$ and generates annotation scores $g(c_i)$, where $g(\cdot)$ denotes an annotation function over the test samples, and $g(c_i)$ is a $1 \times n$ score vector. Concatenating the annotation scores of all the concepts for $\mathbf{X}_{\text{tst}}$, the annotation function can be written as $g(\mathcal{C}) = \{g(c_i)\}_{i=1,\ldots,m} \in R^{m \times n}$.

Denote $\mathbf{W}$ as an $m \times m$ concept affinity matrix indicating the concept relationships in $\mathcal{C}$. Our goal in this paper is to utilize the semantic context $\mathbf{W}$ to refine the annotation score, i.e.,

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

where $\tilde{g}(\mathcal{C})$ is the refined annotation function, and $g(\mathcal{C})$ denotes the initial function based on the supervised concept classifiers; $f(\cdot)$ represents the refinement function, which simultaneously updates $g(\mathcal{C})$ and adapts $\mathbf{W}$ to the test set $\mathbf{X}_{\text{tst}}$ (if domain adaptation is considered).

The 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. $\mathbf{W}$, as mentioned above, is the concept affinity matrix, where each entry W_{ij} indicates the weight of an edge e_{ij} from nodes c_i and c_j. Define the diagonal node degree matrix of the graph as $D_{ii} = d(c_i) = \sum_j W_{ij}$. Then, the graph Laplacian is $\mathbf{\Delta} = \mathbf{D} - \mathbf{W}$, and its normalized version is $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$.

B. Graph Diffusion

Graph diffusion is a widely used technique in data smoothing. Earlier applications of image and video analysis include edge detection and denoising in images [37], discontinuity detection in optical flow [38], etc. Graph diffusion is also closely related to several semisupervised machine learning algorithms [39], [40], where classification functions are determined based on geometry distribution of data samples in feature space.

Function estimation by graph diffusion rests on the assumption that a studied function $g(\cdot)$ is expected to be *smooth* with respect to the manifold geometry of discrete data [41]. With a definition of a cost function $\mathcal{E}(g)$, a smoothed form of the function $g(\cdot)$ can be derived using techniques such as gradient descent, as will be described in our formulation in the next section.

IV. Fast Semantic Diffusion

Now, we formulate context-based image/video annotation as an efficient graph diffusion process. We start by presenting the construction of the semantic graph $\mathcal{G}$ and then describe our formulations of SD.

A. Semantic Graph Construction

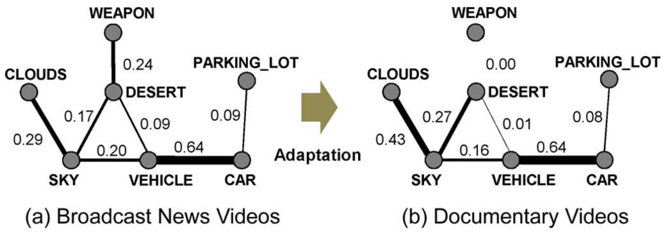
Graph is widely used for abstract data representation. Conventional graph-based methods, such as the popular graph-based SSL algorithms [39], [40], treat samples as nodes, and the graph edges are weighted by the sample similarity. The function to be estimated is usually associated with labels of the samples. Different from the graph modeling techniques, in our semantic graph, each concept c_i in a semantic space is treated as a node, and an edge e_{ij} denotes the relationship between concepts c_i and c_j.

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 $\mathbf{X}_{\text{trn}}$ and its corresponding label matrix $\mathbf{Y}$, where $y_{ij} = 1$ denotes the presence of concept c_i in the sample x_j; otherwise, $y_{ij} = 0$. Based on the label matrix, one simple way to estimate the concept relationship is to use Pearson product-moment correlation, i.e.,

$$PM(c_i, c_j) = \frac{\sum_{k=1}^{|\mathbf{X}_{\text{trn}}|}(y_{ik} - \mu_i)(y_{jk} - \mu_j)}{(|\mathbf{X}_{\text{trn}}| - 1)\,\sigma_i\sigma_j} \tag{2}$$

where μ_i and σ_i are the sample mean and standard deviation (std), respectively, in observing c_i in the training set $\mathbf{X}_{\text{trn}}$. The PM correlation calculated by the above equation can be either negative or positive. Since it is not clear which kind of correlation is more helpful, we construct two undirected positive-weighted graphs separately.

- $\mathcal{G}^+ = (\mathcal{C}, E^+, \mathbf{W}^+)$ 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, i.e., an edge $e_{ij} \in E^-$ is established when $PM(c_i, c_j) < 0$ and $W_{ij}^- = -PM(c_i, c_j)$.



(a) Broadcast News Videos

(b) Documentary Videos

Fig. 2. Visualization of a fraction of the semantic graph before and after domain adaptation. Edge width indicates the degree of correlation between concepts, which is also quantified by the values nearby the edges. (a) Initial concept relationship computed using the manual annotations on TRECVID 2005 training set. (b) Updated concept relationship for TRECVID 2007 test set after performing graph domain adaptation. (See more descriptions in Section VI-A.)

For simplicity, in most part of this paper, we will use the graph $\mathcal{G}^+$ and use $\mathcal{G}^+$ and $\mathcal{G}$ interchangeably without specific declaration. $\mathcal{G}^-$ will be evaluated later in Section IV-C. Fig. 2(a) visualizes a fraction of the semantic graph $\mathcal{G}^+$.

B. SD

Let us now introduce the SD algorithm. Recall that the function value $g(c_i) \in \mathcal{R}^{1\times n}$ on a semantic concept node c_i denotes the concept annotation scores on the test set $\mathbf{X}_{\mathrm{tst}}$, and $n = |\mathbf{X}_{\mathrm{tst}}|$ is the total number of test samples. Intuitively, the overall statistical relationship of the function values $g(c_i)$ and $g(c_j)$ should be consistent with the affinity between the concepts c_i and c_j, i.e., W_{ij}. In other words, strongly correlated concepts should own (statistically) similar concept annotation scores. Motivated by this semantic consistency intuition, we formulate the problem of contextual annotation refinement as a graph diffusion process and define a cost function on the semantic graph as

$$\mathcal{E}(g) = \frac{1}{2}\sum_{i,j=1}^{m} W_{ij}\left\| \frac{g(c_i)}{\sqrt{d(c_i)}} - \frac{g(c_j)}{\sqrt{d(c_j)}} \right\|^2. \tag{3}$$

This function evaluates the smoothness of g over the semantic graph $\mathcal{G}$. Therefore, reducing the function value of $\mathcal{E}$ makes the annotation results more consistent with the concept relationships captured by $\mathcal{G}$.

Gradient descending is adopted to gradually reduce the value of the cost function. Rewriting (3) into matrix formulation, we have

$$\mathcal{E}(\mathbf{g}) = \frac{1}{2}\mathrm{tr}(\mathbf{g}^T\mathbf{L}\mathbf{g}) \tag{4}$$

where $\mathbf{L} \in R^{m\times m}$ is the graph Laplacian of the semantic graph . From (4), the gradient of $\mathcal{E}$ with respect to g on the semantic graph is

$$\nabla_{\mathbf{g}}\mathcal{E} = \mathbf{L}\mathbf{g}. \tag{5}$$

Now, we can derive the following iterative diffusion process:

$$\begin{aligned}\mathbf{g}_t &= \mathbf{g}_{t-1} - \lambda\nabla_{\mathbf{g}_{t-1}}\mathcal{E} = \mathbf{g}_{t-1} - \lambda\mathbf{L}\mathbf{g}_{t-1}\\ &= (\mathbf{I} - \lambda\mathbf{L})\mathbf{g}_{t-1} = (\mathbf{I} - \lambda\mathbf{L})^2\mathbf{g}_{t-2}\\ &= \cdots = (\mathbf{I} - \lambda\mathbf{L})^t\mathbf{g}_0\end{aligned} \tag{6}$$

where $0 < \lambda \ll 1$ is the diffusion step size. Exponentiating the graph Laplacian with λ, we can get

$$\begin{aligned}\mathbf{g}_t &= \left(\mathbf{I} - t(\lambda\mathbf{L}) + \frac{t^2}{2!}(\lambda\mathbf{L})^2 - \frac{t^3}{3!}(\lambda\mathbf{L})^3 + \cdots\right)\mathbf{g}_0\\ &\approx \left(\mathbf{I} - t(\lambda\mathbf{L}) + \frac{1}{2}(\lambda\mathbf{L}t)^2 - \frac{1}{6}(\lambda\mathbf{L}t)^3\right)\mathbf{g}_0.\end{aligned} \tag{7}$$

Further omitting the high-order terms, the above cubic form approximates the exponential diffusion procedure. Instead of iterative diffusion on the initial function value $\mathbf{g}_0$, (6) and (7) give the one-step closed-form diffusion operation through applying the diffusion kernel $\mathcal{K}_t$ on $\mathbf{g}_0$, which is defined as

$$\mathcal{K}_t = (\mathbf{I} - \lambda\mathbf{L})^t. \tag{8}$$

Notice that the cost function in (3) has two main fundamental differences from the existing graph-based SSL techniques such as [39] and [40]. 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 [40]. 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 closed-form solution for efficient SD. 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.

C. Domain Adaptation

As mentioned in Section I, another challenge related to the scalability issue is the problem of data domain changes. In other words, the learnt semantic graph may not completely capture the contextual relationship of unseen video data. For example, the concept *weapon* frequently cooccurs with *desert* in news videos due to plenty of events about Iraq war. While such contextual relationship can be captured in $\mathcal{G}$, misleading annotation will be generated if applied to documentary videos where such relationship is seldom observed. In this section, we address this problem by introducing DASD, which is an extension of SD with a graph adaptation functionality. DASD captures the test domain knowledge by learning from the responses of $g(\cdot)$ over the new and previously unseen test data. The initial semantic graph learnt from training samples is online adapted to the new test domain.

DASD achieves SD and graph adaptation simultaneously by reducing the cost function value via alternatively updating $\mathbf{g}$ 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 $\mathbf{L} = \mathbf{I}-\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} = \mathbf{I}-\tilde{\mathbf{W}}$.

Fig. 3. Example images of a subset of the 81 NUS-WIDE concepts. All the NUS-WIDE images were downloaded from Flickr.com.

Recall the cost function in (4) and express it as a function of $\tilde{\mathbf{W}}$, i.e.,

$$\begin{aligned}\mathcal{E}(\mathbf{g}, \tilde{\mathbf{W}}) &= \frac{1}{2}\mathrm{tr}(\mathbf{g}^T\mathbf{L}\mathbf{g}) = \frac{1}{2}\mathrm{tr}\left(\mathbf{g}^T\left[\mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}\right]\mathbf{g}\right)\\ &= \frac{1}{2}\mathrm{tr}(\mathbf{g}^T\mathbf{g}) - \frac{1}{2}\mathrm{tr}\left(\mathbf{g}^T\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}\mathbf{g}\right)\\ &= \frac{1}{2}\mathrm{tr}(\mathbf{g}^T\mathbf{g}) - \frac{1}{2}\mathrm{tr}(\mathbf{g}^T\tilde{\mathbf{W}}\mathbf{g})\end{aligned} \quad (9)$$

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

$$\frac{\partial\mathcal{E}}{\partial\tilde{\mathbf{W}}} = -\mathbf{g}\mathbf{g}^T. \quad (10)$$

We use gradient descent to iteratively modify the normalized affinity matrix $\tilde{\mathbf{W}}$, i.e.,

$$\tilde{\mathbf{W}}_t = \tilde{\mathbf{W}}_{t-1} - \beta\frac{\partial\mathcal{E}}{\partial\tilde{\mathbf{W}}_{t-1}} = \tilde{\mathbf{W}}_{t-1} + \beta\mathbf{g}_{t-1}\mathbf{g}_{t-1}^T \quad (11)$$

where $0 < \beta \ll 1$ is the step size.

This graph adaptation process, i.e., the refinement of $\tilde{\mathbf{W}}$ with the aim of reducing $\mathcal{E}$, can be intuitively explained as follows. Remember that $\mathbf{g} \in R^{m\times n}$, where m is the number of concepts, and n is the number of test samples. The dot product $\mathbf{g}\mathbf{g}^T \in R^{m\times m}$ in the above equation implies the pairwise concept affinities predicted by the annotation scores from test domain. Thus, the above equation implicitly incorporates new domain knowledge into $\tilde{\mathbf{W}}$.

To inject the graph adaptation function into SD, we update the normalized graph Laplacian as

$$\begin{aligned}\mathbf{L}_t &= \mathbf{I} - \tilde{\mathbf{W}}_\mathbf{t} = \mathbf{I} - \tilde{\mathbf{W}}_{t-1} - \beta\mathbf{g}_{t-1}\mathbf{g}_{t-1}^T\\ &= \mathbf{L}_{t-1} - \beta\mathbf{g}_{t-1}\mathbf{g}_{t-1}^T.\end{aligned} \quad (12)$$

With that, we can derive the following iterative alternating optimization procedure:

$$\begin{aligned}\mathbf{g}_t &= \mathbf{g}_{t-1} - \lambda\mathbf{L}_{t-1}\mathbf{g}_{t-1}\\ \mathbf{L}_t &= \mathbf{L}_{t-1} - \beta\mathbf{g}_{t-1}\mathbf{g}_{t-1}^T.\end{aligned} \quad (13)$$

These two equations form the DASD process, which jointly imposes the SD of annotation scores and the adaptation of the semantic graph structure.

All the above derivations are based on the positive graph $\mathcal{G}^+$. The diffusion and adaptation on $\mathcal{G}^-$ can be done in a very similar manner as $\mathcal{G}^+$. Because in $\mathcal{G}^-$ the edge weights indicate the dissimilarity between the nodes (concepts), the graph diffusion on the negative graph is casted in the way of increasing the cost function value.

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

$$\begin{aligned}\mathbf{g}_t &= \mathbf{g}_{t-1} + \lambda\mathbf{L}_{t-1}^-\mathbf{g}_{t-1}\\ \mathbf{L}_t^- &= \mathbf{L}_{t-1}^- + \beta\mathbf{g}_{t-1}\mathbf{g}_{t-1}^T.\end{aligned} \quad (14)$$

The source codes of the proposed algorithms SD and DASD are publicly available online [1].

V. Experimental Setup

This section outlines our experimental setup. The proposed SD algorithm is assessed using an image benchmark and three video benchmarks. In addition to evaluating the proposed approach under different settings, we also perform simulations to study the effect of contextual detector quality on SD performance. The evaluation plan will be introduced in detail later.

A. Data Sets

We use NUS-WIDE data set [42] as the image annotation benchmark. NUS-WIDE contains 269 648 Flickr images, which are divided into a training set (161 789 images) and a test set (107 859 images). It is manually labeled with 81 semantic concepts/classes, covering a wide range of semantic topics from objects (e.g., *bird* and *car*) to scenes (e.g., *mountain* and *harbor*). Several example images are given in Fig. 3.

The video benchmarks are from NIST[1] TREC video retrieval evaluations (TRECVID) 2005–2007 [43]. In total, there are 340 h of video data. The videos are partitioned into shots, and representative keyframes are extracted from each shot. As shown in Table I, the 2005 and 2006 videos are broadcast news from different TV programs in English, Chinese, and Arabic, whereas the 2007 data set mainly consists of documentary videos in

[1] U.S. National Institute of Standards and Technology.

TABLE I
DETAILS 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 TRAINING DATA FOR TRECVID 2006

TRECVID-	Data domain	Training 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)

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 Large-Scale Concept Ontology for Multimedia (LSCOM) [44]. These concepts are defined according to criteria such as concept utility, observability, and the feasibility of developing classifiers for them using current technologies. Fig. 4 shows example keyframes for several concepts frequently evaluated by NIST.

Note that, for both image and video benchmarks, this is a multilabeling task, which means that each image or video shot can be labeled with more than one concept.

B. Baseline Annotation and Semantic Graph Construction

To apply the proposed SD algorithm, baseline annotation is needed. In other words, we need to set the initial value of $g(\cdot)$. For the Flickr image benchmark, we use original manual tagging as the baseline annotation. In that case, the initial function value of $g(\cdot)$ is binary (0/1 value indicating whether a concept was tagged to an image). While for the video benchmarks, we adopt some generic semantic concept detectors for baseline annotation. Specifically, we use the publicly available VIREO-374[2] [6], which includes SVM models of 374 LSCOM concepts. These models have been shown in TRECVID evaluations to achieve near-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 can be found in [6] and [45]. Later in this paper, we will also show the effectiveness of the proposed technique over different baseline annotation models.

The semantic graph $\mathcal{G}$ is primarily constructed based on the ground-truth labels on the training set of NUS-WIDE and the training labels provided by LSCOM[3] [44], for the image and video benchmarks, respectively, where the edge weights are calculated by (2). In the experiments, we will also test other alternative solutions for graph construction.

In practice, k-NN is commonly used to generate sparse graphs, which usually lead to better performance [46]. In our experiments, we empirically keep six strongest edges for each semantic node and break the remaining connections. In DASD where graph adaptation is performed, in order to keep the graph sparse, we round small gradients in the partial differential in (10) to zero and only keep the largest gradient for each node.

C. Evaluation Plan

Part 1—SD and DASD: The first set of experiments aims at extensively evaluating the proposed SD and its domain adaptive version DASD. Both image and video benchmarks will be used. We will evaluate a couple of settings in detail, including key parameters, graph construction methods, negative concept correlation, and baseline annotation models. Speed efficiency (run time) will be also reported at the end.

Part 2—How good are contextual detectors good enough? This experiment evaluates contextual diffusion performance by varying the quality of the contextual detectors in order to gain additional insights on the upper limit and trend of performance gain from contexts in the task of image and video annotation. We use the training sets of both image and video benchmarks, where the quality of contextual detectors is simulated by artificially adding various levels of noise to the ground-truth annotations of the contextual concepts.

D. Evaluation Criteria

Following the traditions on both the image and the video benchmarks, we use average precision (AP) to evaluate the performance of concept annotation. AP approximates the area under the precision–recall curve. For TRECVID 2006 and 2007 benchmarks, where complete ground-truth labels are not available on the test sets, we choose the NIST official measurement called inferred AP[4], which is an estimation of the traditional AP based on partial labels [47]. To aggregate the performance over multiple semantic concepts, mean AP or mean inferred AP (MAP) is adopted.

VI. RESULTS AND DISCUSSIONS

This section discusses experimental results and gives comparative studies with the state of the arts.

A. SD and DASD

We first evaluate the performance of SD and DASD, mostly using the semantic graph $\mathcal{G}^+$. $\mathcal{G}^-$ is only used in one trial to study the effect of negative correlation. There are some parameters in the proposed method, such as the diffusion step sizes λ and β. We use empirically determined suitable values for these parameters and will show the insensitivity of final performance to particular settings.

Table II summarizes the overall results on both image and video benchmarks. Note that the baseline annotations were obtained by original Flickr manual tagging and the VIREO-374 detectors, respectively, for the image and video benchmarks. Since there is no order information among Flickr images tagged with the same concept, we randomly rank them to form the baseline—this is repeated ten times to report the mean and std of both baseline and SD performance. All of the 81 concepts are evaluated over the image benchmark. While for the TRECVID video benchmarks, NIST selected a subset from the 374 concept pool each year and provided ground-truth labels on the test set—we follow this official setting to evaluate the same set of concepts

[2] http://vireo.cs.cityu.edu.hk/research/vireo374/

[3] LSCOM annotation effort was conducted on broadcast news videos from TRECVID 2005 training set.

[4] http://www-nlpir.nist.gov/projects/tv2006/infAP.html

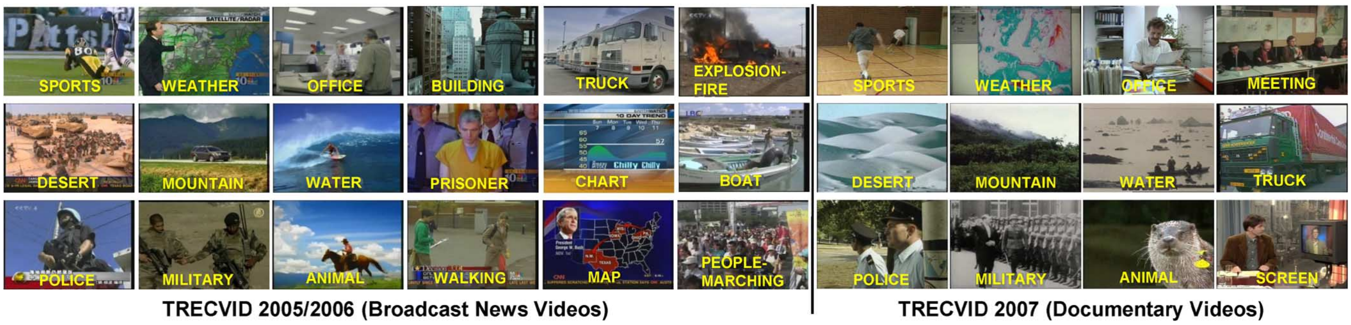


Fig. 4. Example video keyframes of several concepts evaluated in TRECVID 2005–2007. (Left) Examples from broadcast news videos in 2005 and 2006. (Right) Examples from documentary videos in 2007. Note that appearance of the same concept from the two data domains may be visually quite different.

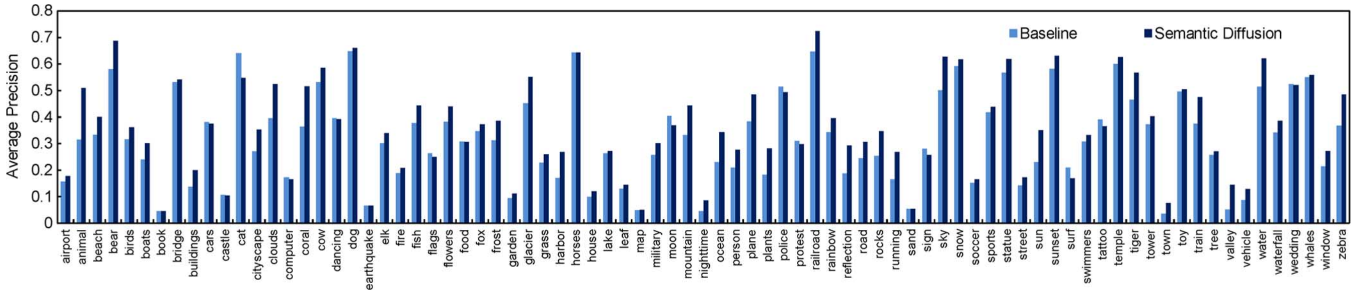


Fig. 5. Per-concept performance before and after SD on NUS-WIDE test set (mean APs over the ten runs). SD improves 66 concepts. Among the rest, very few concepts suffer from significant performance degradation.

TABLE II
Overall Performance Gain (Relative Improvement) on NUS-WIDE and TRECVID (TV) 2005–2007 Benchmarks. SD: Semantic Diffusion. DASD: Domain Adaptive Semantic Diffusion

	NUS-WIDE	TV '05	TV '06	TV '07
# evaluated concepts	81	39	20	20
Baseline (MAP)	0.316 ± 0.001	0.166	0.154	0.099
SD	14.0±0.4%	11.8%	15.6%	12.1%
DASD	9.4±0.3%	11.9%	17.5%	16.2%

over each year's test data. From the table, we see that, when SD is used, MAP gain (relative improvement) is consistently around 14.0% on NUS-WIDE. For the TRECVID 2005–2007 video benchmarks, MAP is improved by 11.8% to 15.6%. These results confirm the effectiveness of formulating graph diffusion for improving image and video annotation accuracy.

Fig. 5 gives per-concept performance of the 81 NUS-WIDE concepts, and Fig. 6 shows per-concept performance of the 20 evaluated concepts in TRECVID 2006. Our approach consistently improves 66 NUS-WIDE concepts and all the TRECVID 2006 concepts. In addition, among a total of 79 concepts from TRECVID 2005 to 2007 as reported in Table II, almost all concepts show improvement, except five concepts that have slight AP drop. For the concepts with performance degradation, one main reason is that the detector quality of their contextual concepts (i.e., those close to the target concepts according to the semantic graph) is not as good as that of the target concepts. As will be discussed in Section VI-B, in this case, it is difficult to attain performance gain from contextual diffusion.

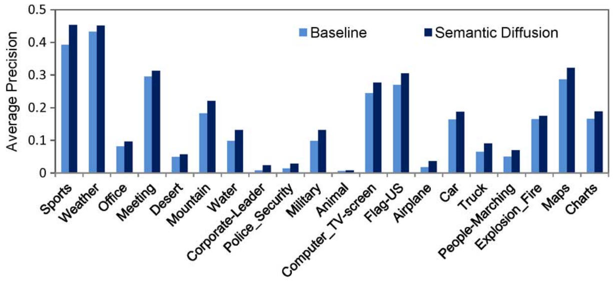


Fig. 6. Per-concept performance before and after SD on TRECVID 2006 test set. Consistent improvements are observed for all of the 20 semantic concepts.

Due to the video domain change 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 II, DASD further boosts the performance on TRECVID 2006 and 2007. On the other hand, there is no improvement on TRECVID 2005 and NUS-WIDE, which is due to the fact that there is no domain change in both data sets. Recall that the domain adaptation process incorporates new domain knowledge by adjusting concept affinities using the relationships of annotation scores from test domain (cf. Section IV-C). When training and test data are from the same domain, altering the concept affinities (initially computed from perfect training labels) using imperfect annotation scores on test data is apparently not useful. 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 changes a lot. We consider the graph adaptation process as a 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. One is suggested applying DASD when domain

TABLE III
PERFORMANCE COMPARISON OF DASD WITH SEVERAL EXISTING WORKS ON TRECVID BENCHMARKS

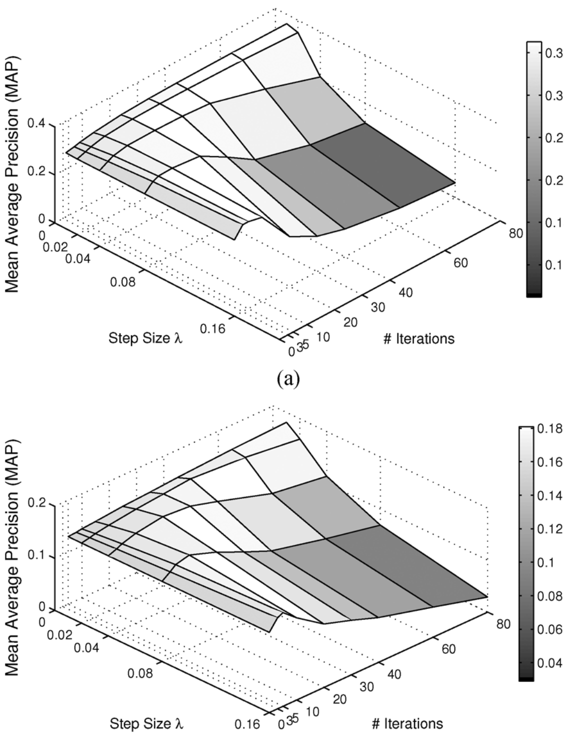
	Jiang et al. [34]	Aytar et al. [48]	Weng et al. [33]	DASD
2005	2.2%	4.0%	N/A	11.9%
2006	N/A	N/A	16.7%	17.5%

shift is expected to happen between training and test data. Fig. 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 cooccurs with *desert* scene in TRECVID 2005 broadcast news videos because there are many events about Iraq war, whereas in the documentary videos, this is rarely observed.

We compare our approach to three existing works in [33], [34], and [48] for context-based video annotation. As shown in Table III, 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 [34], and Aytar *et al.* improved 4% [48]. Another work on utilizing semantic context for video annotation is by Weng and Chuang [33], who reported a performance gain of 16.7% over the same VIREO-374 baseline on TRECVID 2006 test set. However, techniques in [33] did not show the domain adaptation ability, and they used a graphical model formulation, which is computationally much more expensive than our approach. In addition, the parameters in [33] were separately optimized for each of the concepts, whereas we use uniform parameter settings across all the concepts. We demonstrate a performance gain of 17.5%, which is, to our knowledge, the highest reported improvement on exploiting semantic context for video annotation.

Effect of Parameters: Fig. 7 shows the NUS-WIDE and TRECVID 2006 performance surfaces under different settings. As can be seen from results on both benchmarks, there is a tradeoff between the number of iterations and the diffusion step size λ. The finer λ is used, the more iterations are needed to reach the best diffusion performance. This empirical study also indicates that different λ, when combined with, correspondingly, the best iteration number, can lead to fairly consistent performance. Although NUS-WIDE is very much different from TRECVID in terms of data characteristics, we see a very consistent performance trend on both data sets. We also evaluate the same parameter setting ($\lambda = 0.04$) over different data sets, i.e., TRECVID 2005 and 2007, and verified that the optimal number of iteration (20) always can achieve the best or close-to-the-best performance.

Compared with SD, DASD has one more parameter, i.e., the step size of adaptation β. We set $\lambda = \beta$ in the experiments for simplicity and found that 20 remains a good choice for the number of iterations. These findings confirm the performance stability of the proposed SD and DASD over parameter settings. The same parameters ($\lambda = 0.04$ and #iteration $= 20$ for SD; $\lambda = \beta = 0.04$ and #iteration $= 20$ for DASD) are used throughout the experiments in this paper.



(a)

(b)

Fig. 7. SD performance on (a) NUS-WIDE test set and (b) TRECVID 2006 test set under various parameter settings. Performance trends and optimal parameter values are similar on the two very different benchmarks.

Effect of concept affinity estimation method: The concept affinities in the semantic graphs are computed based on the ground-truth training labels in the above experiments. An alternative way is to estimate the concept affinities according to the baseline annotation scores on each benchmark's test data. Let $\mathcal{T}$ be the test set and g_k^i be the baseline annotation scores of concept c_i in test image or video shot k (g_k^i is a binary vector for the Flickr image benchmark). Similar to (2), 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 std 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 $\mathcal{G}_{\mathcal{T}}$ for each benchmark.

We rerun SD on NUS-WIDE and TRECVID 2007 test sets using the corresponding new $\mathcal{G}_{\mathcal{T}}$. The overall performance gains are 9.2% and 9.3%, respectively, for the image and video benchmarks, which are lower compared with the results (14.0% and 12.1%, respectively) from using $\mathcal{G}$ derived from the golden-labeled training sets. Nevertheless, this process is more economic than the construction of $\mathcal{G}$ using the fully labeled training sets. Manual labeled training data are difficult to obtain in practice, particularly when the number of concepts is in the order of thousands. Thus, constructing semantic graphs based on initial annotation scores (from either noisy manual tagging or machine learning) is a promising way when fully labeled training sets are unavailable.

Effect of Negative Correlation: In order to study the effect of negative correlations, we conduct another experiment on NUS-WIDE and TRECVID 2007 test sets. When using the negative

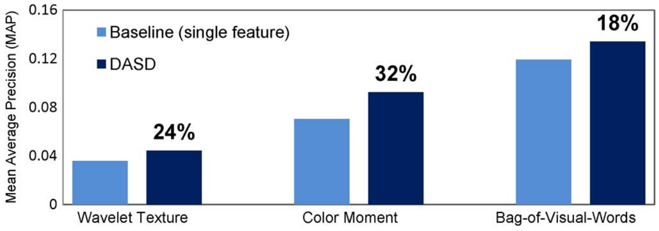


Fig. 8. DASD performance using various baselines on TRECVID 2006 test set. The text under light and dark blue bars indicates the feature used in the corresponding baseline detector.

graph $\mathcal{G}^-$ alone, the performance gain from SD is 5.9% on the image benchmark and merely 1.3% on the video benchmark.

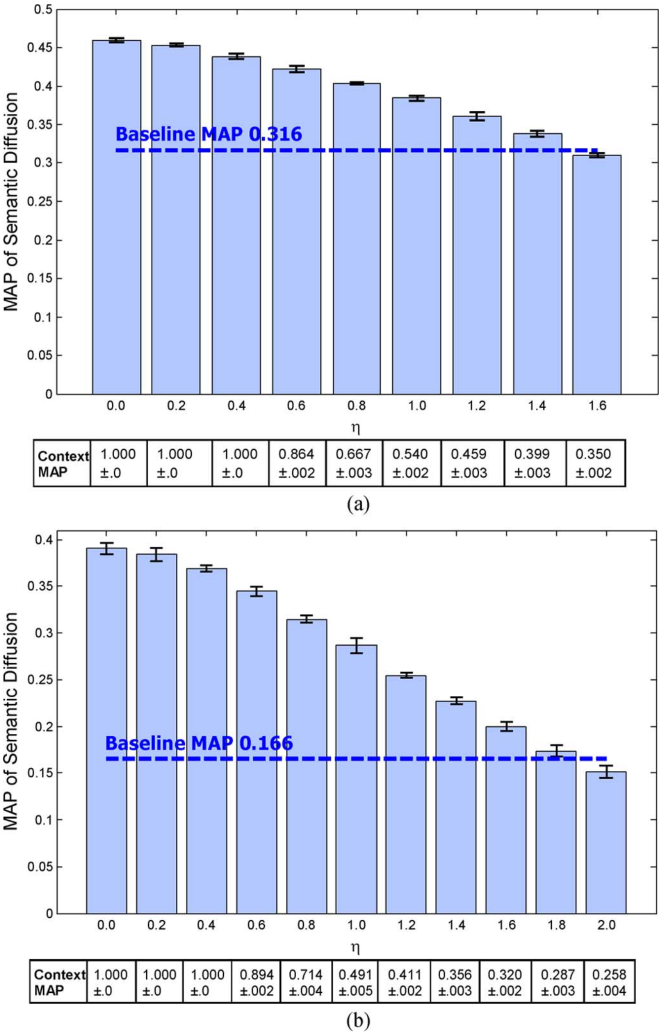
When both $\mathcal{G}^+$ and $\mathcal{G}^-$ are used (alternatively apply the first part of (13) and (14), i.e., without graph adaptation), the MAP gain is 18.1% on NUS-WIDE and 12% on TRECVID 2007. The result is about the same with that using $\mathcal{G}^+$ alone on TRECVID (12.1%). On the image benchmark, moderately higher performance gain is observed. Based on these results, we conclude that, although negative correlations captured by $\mathcal{G}^-$ may improve the performance, in practice, $\mathcal{G}^+$ alone is preferred for large-scale applications since it represents a good tradeoff between performance and speed.

Effect of Baseline Performance: Let us now evaluate the sensitivity of our approach to the performance of baseline detectors from machine learning. This experiment is conducted over TRECVID 2006 test set. As mentioned in Section V-B, instead of using fused output of the three SVMs as baseline annotation, here we adopt the prediction output of each single (and relatively weaker) modality as initial values for function $g(\cdot)$. The domain adaptive version DASD is then applied to these three weak baselines, respectively. Results are visualized in Fig. 8. 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 dark blue bar). From these results, we can see that the proposed approach is able to achieve consistently better performance over various baseline detectors, even for some fairly weak ones.

Speed Efficiency: The proposed approach is extremely efficient. Constructing the semantic graph $\mathcal{G}$ using NUS-WIDE and TRECVID 2005 training sets takes 59 and 284 s, respectively. The complexity of both SD and its adaptive version DASD is $O(m^2 n)$, where m is the number of concepts (normally ranging from several hundreds to a few thousands), and n is the number of data samples (usually a large number). Table IV lists the detailed run time on each data set. Take TRECVID 2006—which contains 79 484 video shots—as an example, the SD algorithm finishes in just 59 s, and DASD uses 165 s. In other words, running SD over the 374 concepts for each video shot only takes 1 ms. This is much faster than the existing works in [12], [33], and [34], in which tens or hundreds of hours are required due to the expensive training process involved in their methods.

TABLE IV
RUN TIME ON NUS-WIDE AND TRECVID (TV) 2005–2007 BENCHMARKS. THE EXPERIMENTS WERE CONDUCTED ON AN INTEL CORE-2 DUO 2.2-GHZ PC WITH 2G RAM

	NUS-WIDE	TV '05	TV '06	TV '07
SD	24s	59s	84s	12s
DASD	42s	89s	165s	28s



Context MAP	1.000 ±.0	1.000 ±.0	1.000 ±.0	0.864 ±.002	0.667 ±.003	0.540 ±.002	0.459 ±.003	0.399 ±.003	0.350 ±.002

(a)

Context MAP	1.000 ±.0	1.000 ±.0	1.000 ±.0	0.894 ±.002	0.714 ±.004	0.491 ±.005	0.411 ±.002	0.356 ±.003	0.320 ±.002	0.287 ±.003	0.258 ±.004

(b)

Fig. 9. SD performance on (a) NUS-WIDE and (b) TRECVID 2005, using contextual detectors at various performance levels [simulated with varying η; see (15)]. Baseline annotation can be improved when η ranges from 0 (perfect context) to 1.4 for NUS-WIDE and 1.8 for TRECVID. The table below each chart gives MAP performance of the contextual detectors simulated using the corresponding η on top of each cell.

In Web-scale applications, there may be millions of test images or video shots. Baseline annotation scores ($\mathbf{g} \in R^{m \times n}$; n is the number of samples) of data samples in such a huge scale cannot be loaded in memory. Since the proposed algorithm diffuses detection scores in each column of $\mathbf{g}$ independently [see eq. (6)], we can get around this scalability issue easily by splitting $\mathbf{g}$ columnwise into an arbitrary number of smaller matrices and then by running SD on each of them separately.

B. How Good Are Contextual Detectors Good Enough?

This experiment evaluates contextual diffusion performance by varying the quality of the contextual detectors. Note that this study is different from the experiment in Section VI-A-4, where we have tested SD over different baseline annotation methods.

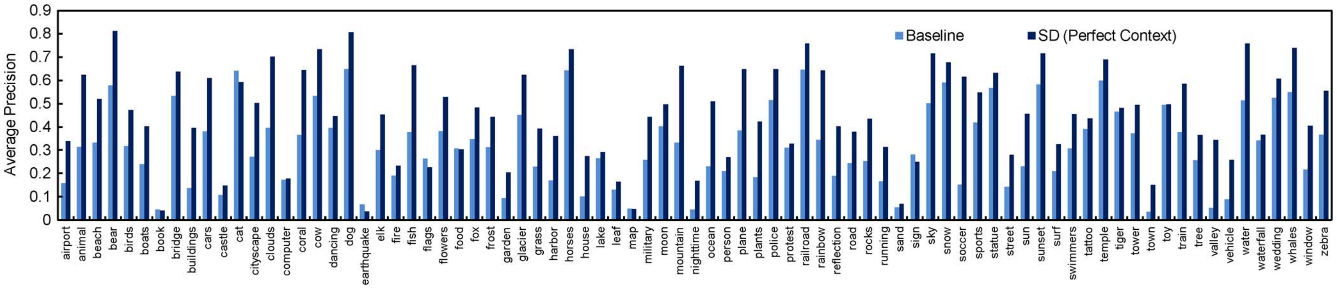


Fig. 10. Per-concept performance before and after SD on NUS-WIDE test set using perfect contextual annotations ($\eta = 0$).

The main purposes here are to study the effect of *contextual detector reliability* on SD (and, potentially, also other contextual annotation approaches) and to learn the upper limit of context as auxiliary cues for image and video annotation.

We use NUS-WIDE and TRECVID 2005[5] benchmarks. Instead of using manual tagging (image benchmark) or machine learning classification (video benchmarks) as baseline annotation of contextual concepts, we adopt ground-truth manual labels. The quality of contextual detectors is then simulated by artificially adding various levels of noise to the ground-truth annotations. Formally, the simulated contextual concept annotations are obtained as follows:

$$\mathbf{g}_{\text{simu}} = \mathbf{gt} + \eta \cdot \mathbf{r} \tag{15}$$

where $\mathbf{gt} \in \mathcal{R}^{m\times n}$ is the ground-truth annotation matrix, $\mathbf{r} \in \mathcal{R}^{m\times n}$ is a randomly generated noise matrix with values in $[-1, 1]$, and η is a parameter controlling the degree of noise imposed on $\mathbf{gt}$.

During SD, given a target concept c_i, we use its original baseline annotation $g(c_i)$ as initial values of the ith row in $\mathbf{g}$, whereas initial values of the other contextual concepts are all drawn from corresponding rows of $\mathbf{g}_{\text{simu}}$. This way, we are able to reach the scenario that contexts with varying accuracy values (adjusted by η) are used for improving a target concept annotated by manual tagging or machine learning.

Fig. 9 shows the results on the image benchmark NUS-WIDE and the video benchmark TRECVID 2005, where the bars indicate SD performance using contexts simulated with various η. The table below each bar chart gives the performance of the context detectors simulated with the η value on top of each cell. Since there is a random factor (matrix $\mathbf{r}$) in the generation of $\mathbf{g}_{\text{simu}}$, for each choice of η, we repeat the experiment ten times and compute the mean and std of both contextual detector performance (reported in the tables) and final SD performance (std shown on top of the bars). From the figure, we see that the std values are quite small, indicating that the performance is not sensitive to the randomization of noise matrix $\mathbf{r}$.

As shown in the figure, when perfect context is in use ($\eta = 0$), the MAP performance can be significantly improved to 0.459 for NUS-WIDE (relative gain 45.3% over the baseline with a MAP of 0.316). The improvement is even more significant for TRECVID 2005, with a MAP gain of 135%. This clearly shows the great potential of context for visual annotation. Fig. 10 further shows the per-concept AP on NUS-WIDE using perfect context, where the performance of most concepts is improved very significantly.

Another observation is that SD performance drops approximately linearly to the amount of noise added. No improvement is observed when η increases to 1.6 (context MAP 0.350) for NUS-WIDE, which is a bit surprising because an improvement of 14% was obtained using the manual tagging baseline as context (see Table II), whose MAP is lower than that of the simulated context here (0.316 versus 0.350). This indicates that the artificial noise added to the simulated contexts affects SD more significantly than its effect on the context MAP. A similar observation also holds for the TRECVID 2005 benchmark. Overall, from the SD performance trend observed in these experiments, we can conclude that context is likely to be helpful when its annotation/detection quality is similar to or better than that of the target concept baseline before contextual fusion.

VII. Conclusion and Discussion

We have presented a novel and efficient approach, which is named SD, to exploit semantic contexts for improving image and video annotation accuracy. The semantic context is embedded in an undirected and weighted concept graph, on top of which we recover the consistency and smoothness of video annotation results using a function-level graph diffusion process. Extensive experiments on both image and video benchmark data sets show that the semantic context is powerful for enhancing annotation accuracy, and the proposed SD algorithm consistently and significantly improves the performance over the vast majority of the evaluated concepts. Furthermore, an adaptive version of the proposed algorithm, which is called DASD, 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 annotation accuracy. Additionally, in the experiments on graph node affinity estimation, we demonstrated that the semantic graph can be bootstrapped using initial annotation results from noisy manual tagging or machine learning, which is very helpful when an exhaustively labeled training set is not available to construct the semantic graph.

Another research question we had is "how good are contextual detectors good enough?". We conducted a set of experiments to simulate contextual detectors at various performance levels. We observed that, with perfect contextual detectors, SD

[5] TRECVID 2005 is the only data set containing full annotations of all the 374 LSCOM concepts, which allows us to simulate context detectors at various performance levels.

is able to improve a decent baseline by 45% on the image benchmark and 135% on the video benchmark, which, again, signifies the great potential of contexts in visual annotation. Our answer to the question is that context is very likely to be helpful when the quality of contextual detectors is better than or at least on the same level as the target concept baseline annotation. Since, in most cases, we have similar baseline and contextual annotation quality, this requirement should not be a bottleneck of context-based image and video annotation.

Currently, our approach models concept affinities using an undirected graph. This is not always ideal since, in some cases, the contextual relationship of concepts may be *directional* (e.g., seeing a *car* indicates the existence of a *road* but not vice versa). Therefore, one promising idea for future research is to adopt a directed graph to model the semantic contexts. In addition, since not all the concepts can benefit from contextual diffusion, how to predict which concept detector could be improved from contextual modeling is another interesting topic that deserves future investigations.

References

[1] Fast Semantic Diffusion for Context-Based Image and Video Annotation: Project Page and Source Codes [Online]. Available: http://www.ee.columbia.edu/ln/dvmm/researchProjects/MultimediaIndexing/DASD/dasd.htm

[2] 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 *Proc. CVPR Workshop Generative-Model Based Vis.*, 2004, p. 178.

[3] 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 *Proc. ACM Int. Conf. Multimedia*, Oct. 2006, pp. 421–430.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, *The PASCAL visual object classes (VOC) challenge*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[5] J. Fan, Y. Gao, and H. Luo, "Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation," *IEEE Trans. Image Process.*, vol. 17, no. 3, pp. 407–426, Mar. 2008.

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

[7] Z. Lu, H. Ip, and Y. Peng, "Contextual kernel and spectral methods for learning the semantics of images," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1739–1750, Jun. 2011.

[8] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar, "Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 958–966, Aug. 2007.

[9] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, pp. 1458–1465.

[10] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 2169–2178.

[11] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[12] G. J. Qi, X. S. Hua, Y. Rui, J. Tang, T. Mei, and H. J. Zhang, "Correlative multi-label video annotation," in *Proc. ACM Multimedia*, 2007, pp. 17–26.

[13] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, May/Jun. 2001.

[14] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[15] A. C. Berg and J. Malik, "Geometric blur for template matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2001, pp. I-607–I-614.

[16] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2223–2231.

[17] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2003, pp. II-264–II-271.

[18] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[19] L. Cao, J. Luo, F. Liang, and T. Huang, "Heterogeneous feature machines for visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1095–1102.

[20] A. Torralba, "Contextual priming for object detection," *Int. J. Comput. Vis.*, vol. 53, no. 2, pp. 169–191, Jul. 2003.

[21] K. Murphy, A. Torralba, and W. Freeman, "Using the forest to see the tree: A graphical model relating features, objects and the scenes," *Adv. Neural Inf. Process. Syst.*, pp. 1–8, 2003.

[22] 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," *Int. J. Comput. Vis.*, vol. 81, no. 1, pp. 2–23, Jan. 2009.

[23] A. Torralba, K. Murphy, W. Freeman, and M. Rubin, "Context-based vision system for place and object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2003, pp. 273–280.

[24] L. Wolf and S. Bileschi, "A critical view of context," *Int. J. Comput. Vis.*, vol. 69, no. 2, pp. 251–261, Aug. 2006.

[25] P. Carbonetto, N. de Freitas, O. de Freitas, and K. Barnard, "A statistical model for general contextual object recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 350–362.

[26] T. Malisiewicz and A. A. Efros, "Beyond categories: The visual memex model for reasoning about object relationships," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1222–1230.

[27] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[28] Y. J. Lee and K. Grauman, "Object-graphs for context-aware category discovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1–8.

[29] J. J. Lim, P. Arbeláez, C. Gu, and J. Malik, "Context by region ancestry," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1978–1985.

[30] D. Lin, A. Kapoor, G. Hua, and S. Baker, "Joint people, event, and location recognition in personal photo collections using cross-domain context," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 243–256.

[31] N. Rasiwasia and N. Vasconcelos, "Holistic context modeling using semantic co-occurrences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1889–1895.

[32] J. Tang, X.-S. Hua, M. Wang, Z. Gu, G.-J. Qi, and X. Wu, "Correlative linear neighborhood propagation for video annotation," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 2, pp. 409–416, Apr. 2009.

[33] M.-F. Weng and Y.-Y. Chuang, "Multi-cue fusion for semantic video indexing," in *Proc. ACM Multimedia*, 2008, pp. 71–80.

[34] W. Jiang, S. F. Chang, and A. Loui, "Context-based concept fusion with boosted conditional random fields," in *Proc. ICASSP*, 2007, pp. I-949–I-952.

[35] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan, "Contextualizing object detection and classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1585–1592.

[36] Y.-G. Jiang, J. Wang, S.-F. Chang, and C.-W. Ngo, "Domain adaptive semantic diffusion for large scale context-based video annotation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1420–1427.

[37] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 7, pp. 629–639, Jul. 1990.

[38] M. Proesmans, L. J. van Gool, E. Pauwels, and A. Oosterlinck, "Determination of optical flow and its discontinuities using nonlinear diffusion," in *Proc. Eur. Conf. Comput. Vis.*, 1994, pp. 295–304.

[39] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 321–328.

[40] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. Int. Conf. Mach. Learn.*, 2003, pp. 912–919.

[41] A. D. Szlam, M. Maggioni, and R. R. Coifman, "Regularization on graphs with function-adapted diffusion processes," *J. Mach. Learn. Res.*, vol. 9, pp. 1711–1739, Jun. 2008.

[42] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM CIVR*, 2009, pp. 1–9.

[43] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVID," in *Proc. ACM MIR*, 2006, pp. 321–330.

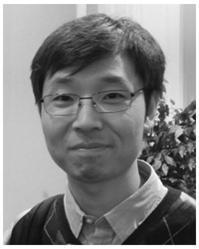
[44] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large scale concept ontology for multimedia," *IEEE Multimedia*, vol. 13, no. 3, pp. 86–91, Jul.–Sep. 2006.

[45] Y.-G. Jiang, A. Yanagawa, S.-F. Chang, and C.-W. Ngo, CU-VIREO374: Fusing Columbia374 and VIREO374 for large scale semantic concept detection Columbia Univ., Manhattan, NY, ADVENT Tech. Rep. 223-2008-1, Aug. 2008.

[46] T. Jebara, J. Wang, and S.-F. Chang, "Graph construction and b-matching for semi-supervised learning," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 441–448.

[47] E. Yilmaz and J. A. Aslam, "Estimating average precision with incomplete and imperfect judgments," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2006, pp. 102–111.

[48] Y. Aytar, O. B. Orhan, and M. Shah, "Improving semantic concept detection and retrieval using contextual estimates," in *Proc. ICME*, 2007, pp. 536–539.



Yu-Gang Jiang received the Ph.D. degree in computer science from the City University of Hong Kong, Kowloon, Hong Kong, in 2009.

During 2008–2011, he was with the Department of Electrical Engineering, Columbia University, New York, as a Visiting Scholar the first year and later as a Postdoctoral Research Scientist. He is currently an Associate Professor of computer science with Fudan University, Shanghai, China. His research interests include multimedia retrieval and computer vision. He has authored more than 30 papers in these fields.

Dr. Jiang is an active participant of the Annual NIST TRECVID Evaluation and has designed a few top-performing video retrieval systems over the years. He has served on the technical program committees of many international conferences and is a Guest Editor of a forthcoming special issue on Socio-Video Semantics, IEEE TRANSACTIONS ON MULTIMEDIA.

Qi Dai received the B.Sc. degree in computer science from East China University of Science and Technology, Shanghai, China, in 2011. He is currently working toward the M.Sc. degree with the School of Computer Science, Fudan University, Shanghai.

His research interests include multimedia retrieval and computer vision.

Jun Wang (M'12) received the M.Phil. and Ph.D. degrees from Columbia University, New York, in 2010 and 2011, respectively.

Currently, he is a Research Staff Member with the business analytics and mathematical sciences department at IBM T. J. Watson Research Center, Yorktown Heights, NY. He also worked as an Intern at Google Research in 2009 and as a Research Assistant at Harvard Medical School, Harvard University, Cambridge, MA, in 2006. His research interests include machine learning, business analytics, information retrieval, and hybrid neural-computer vision systems.

Dr. Wang is the recipient of several awards and scholarships, including the Jury thesis award from the Department of Electrical Engineering, Columbia University, in 2011; the Google global intern scholarship in 2009; and a Chinese government scholarship for outstanding self-financed students abroad in 2009.

Chong-Wah Ngo (M'02) received the Ph.D. degree in computer science from Hong Kong University of Science and Technology, Kowloon, Hong Kong, and the M.Sc. and B.Sc. degrees, both in computer engineering, from Nanyang Technological University, Singapore.

He is an Associate Professor with the Department of Computer Science, City University of Hong Kong, Kowloon. Before joining City University, he was a Postdoctoral Scholar with the Beckman Institute, University of Illinois in Urbana-Champaign. He was also a Visiting Researcher with Microsoft Research Asia. During 2008–2009, he served as the Chairman of ACM (Hong Kong Chapter). His recent research interests include large-scale multimedia information retrieval, video computing, and multimedia mining.

Dr. Ngo is currently an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA. He is also the Program Cochair of the ACM Multimedia Modeling 2012 and the International Conference on Multimedia Retrieval 2012.

Xiangyang Xue (M'05) received the B.S., M.S., and Ph.D. degrees in communication engineering from Xidian University, Xi'an, China, in 1989, 1992, and 1995, respectively.

He joined the Department of Computer Science, Fudan University, Shanghai, China, in 1995. Since 2000, he has been a Full Professor with the Department of Computer Science, Fudan University. He has authored more than 100 research papers in these fields. His current research interests include multimedia information processing and retrieval, pattern recognition, and machine learning.

Dr. Xue is an Associate Editor of the IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT. He is also an Editorial Board Member of the Journal of Computer Research and Development and the Journal of Frontiers of Computer Science and Technology.

Shih-Fu Chang (S'89–M'90–SM'01–F'04) is the Richard Dicker Professor with the Departments of Electrical Engineering and Computer Science and the Director of Digital Video and Multimedia Lab with Columbia University, New York. He has made significant contributions to multimedia search, visual communication, media forensics, and international standards. He has worked in different advising/consulting capacities for industry research labs and international institutions.

Prof. Chang is a Fellow of the American Association for the Advancement of Science. He served as Editor-in-Chief for the IEEE Signal Processing Magazine (2006–2008) and as Chair of Columbia's Electrical Engineering Department (2007–2010). He has been recognized with ACM SIGMM Technical Achievement Award, IEEE Kiyo Tomiyasu Award, Navy ONR Young Investigator Award, IBM Faculty Award, ACM Recognition of Service Award, and NSF CAREER Award. He and his students have received many Best Paper Awards, including the Most Cited Paper of the Decade Award from the Journal of Visual Communication and Image Representation.