Unified Video Annotation via Multigraph Learning

Meng Wang, Xian-Sheng Hua, *Member, IEEE*, Richang Hong, Jinhui Tang, Guo-Jun Qi, and Yan Song

***Abstract*— Learning-based video annotation is a promising approach to facilitating video retrieval and it can avoid the intensive labor costs of pure manual annotation. But it frequently encounters several difficulties, such as insufficiency of training data and the curse of dimensionality. In this paper, we propose a method named optimized multigraph-based semi-supervised learning (OMG-SSL), which aims to simultaneously tackle these difficulties in a unified scheme. We show that various crucial factors in video annotation, including multiple modalities, multiple distance functions, and temporal consistency, all correspond to different relationships among video units, and hence they can be represented by different graphs. Therefore, these factors can be simultaneously dealt with by learning with multiple graphs, namely, the proposed OMG-SSL approach. Different from the existing graph-based semi-supervised learning methods that only utilize one graph, OMG-SSL integrates multiple graphs into a regularization framework in order to sufficiently explore their complementation. We show that this scheme is equivalent to first fusing multiple graphs and then conducting semi-supervised learning on the fused graph. Through an optimization approach, it is able to assign suitable weights to the graphs. Furthermore, we show that the proposed method can be implemented through a computationally efficient iterative process. Extensive experiments on the TREC video retrieval evaluation (TRECVID) benchmark have demonstrated the effectiveness and efficiency of our proposed approach.**

***Index Terms*— Multimodal fusion, semi-supervised learning, video annotation.**

I. Introduction

With rapid advances in storage devices, networks, and compression techniques, large-scale video data have become available to ordinary users. Content-based video search thus has become an increasingly active field. It is well known that a central problem of this field is the so-called semantic gap, namely, the gap between low-level (signal-level) features and high-level (semantic-level) queries. Recent studies reveal that annotating a large set of semantic concepts for the video data is a promising approach to bridging this gap [10], [11], [18], [22]. As noted by Hauptmann [10], "this splits the semantic gap between low level features and user information needs into two, hopefully smaller gaps: (a) mapping the low-level features into the intermediate semantic concepts and (b) mapping these concepts into user needs." Annotation is exactly the step to accomplish the first mapping. However, manual annotation for a large video archive is labor intensive and time consuming. For example, experiments in [20] prove that typically annotating 1 h of video with 100 concepts can take anywhere between 8 and 15 h. Therefore, efficient automatic annotation methods are highly desirable.

Generally, automatic video annotation (also referred to as "video concept detection" [25], "video semantic analysis" [31], or "high-level feature extraction" [17]) can be accomplished by machine learning methods. A typical learning-based video annotation method works as follows. First, videos are segmented into short units such as shots and sub-shots. Then, low-level features are extracted from each unit to describe its content. Video annotation is then formalized to learn a set of predefined concepts for each unit based on these low-level features. Since the to-be-annotated concepts may not be mutually exclusive (such as the concepts "street" and "outdoor"), a general scheme is to conduct a binary classification procedure for each concept. Given a concept, each unit is then annotated to be "positive" or "negative" according to whether it is associated with this concept. The National Institute of Standards and Technology (NIST) has also established "high-level feature extraction" as a task in TREC video retrieval evaluation (TRECVID) [1], [28], which aims to provide a benchmark for evaluating video annotation technologies. Naphade *et al.* [25] have presented a survey on the benchmark, where a great deal of different algorithms applied to this task can be found. Recent studies have demonstrated that video annotation could benefit from the investigation of a diverse set of features and learning methods. For example, Wang *et al.* [38] have shown the effectiveness of combining different features and Amir *et al.* [2] have integrated different learning algorithms, including support vector machine, Gaussian mixture model, maximum entropy methods, a modified nearest neighbor method, and multiple-instance learning. Snoek *et al.* have proposed a semantic pathfinder method which benefits from the exploitation of the video authoring process [30].

Although many different methods have been proposed for this task and several encouraging results have been reported [2], [30], [38], [19], we still frequently encounter the following difficulties which may result in the inaccurate annotation results.

1) *Insufficiency of training data*. To guarantee reasonable annotation accuracy, a large training set with enough sample prototypes is required in order to bridge the gap between low-level features and semantic concepts. However, this requirement is usually difficult to meet due to the high labor costs of manual annotation [7], [20], [40].
2) *Curse of Dimensionality*. To differentiate or describe a variety of semantic concepts, we have to extract a large

Manuscript received January 23, 2008; revised October 5, 2008. First version published March 16, 2009; current version published June 10, 2009. This paper was recommended by Associate Editor P. L. Correia.

M. Wang and X.-S. Hua are with the Microsoft Research Asia, Beijing 100080, P. R. China (e-mail: mengwang@microsoft.com; xshua@microsoft.com).

R. Hong, J. Tang, G.-J. Qi, and Y. Song are with the University of Science and Technology of China, Hefei 230027, P. R. China (e-mail: richong@mail.ustc.edu.cn; jhtang@mail.ustc.edu.cn; qgj@mail.ustc.edu.cn; songy@ustc.edu.cn).

Digital Object Identifier 10.1109/TCSVT.2009.2017400

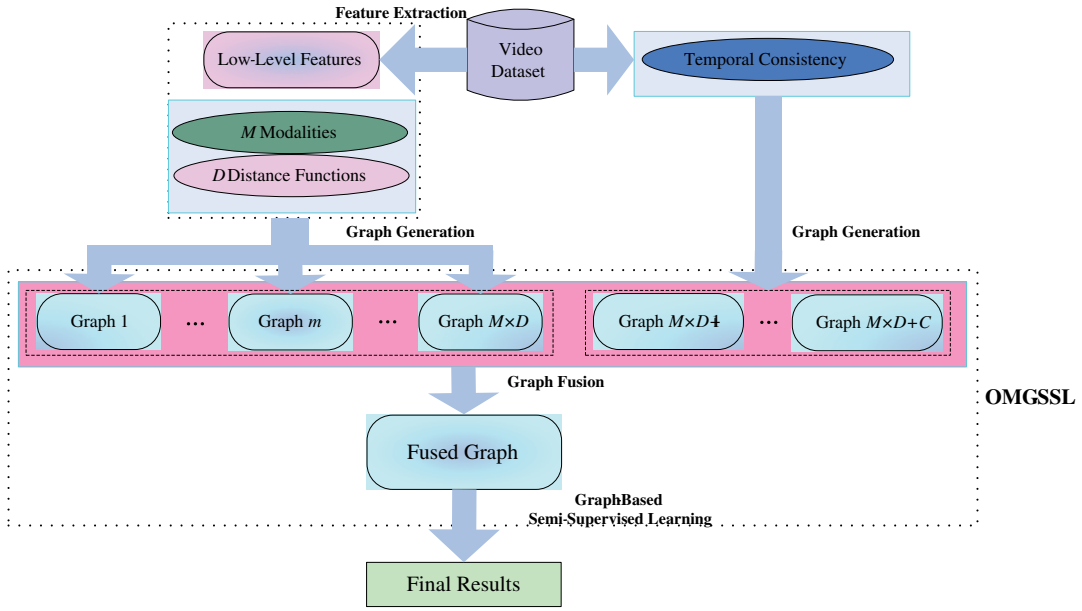

Fig. 1. Schematic illustration of the OMG-SSL-based video annotation process. It is equivalent to conducting semi-supervised learning on a graph fused from the graphs that encode the knowledge from multiple modalities, multiple distance functions, and temporal consistency.

amount of low-level features. But a high-dimensional feature space frequently leads to "curse of dimensionality" which may induce performance degradation [4], [42].

3) *Choice of distance Function*. It is well known that many learning methods heavily rely on the adopted distance function. However, the optimal distance function varies for different features and/or different semantic concepts. On the other hand, complementarity may exist among different distance functions [47]. However, the selection of best distance function and the combination of multiple distance functions are both challenging issues.
4) *Neglect of temporal consistency*. Temporal consistency is a widely noted property in video data, which means that the variation of semantic concepts within a continuous video segment is usually much smaller compared to that in different video segments [15], [45]. It indicates that adjacent video shots may share the same semantic concepts with high probability. This property can help improve annotation performance if it is exploited appropriately, but in the existing works it is often neglected or not sufficiently utilized.

Various methods have been proposed aiming to tackle the above problems, such as applying semi-supervised learning to deal with the training data insufficiency problem and utilizing multimodal fusion to avoid dimensionality curse. However, to the best of our knowledge, there is no unified scheme that can simultaneously deal with the above four problems. In this paper, we propose such an approach named optimized multigraph-based semi-supervised learning (OMG-SSL). Different from the traditional graph-based semi-supervised learning algorithms that mainly focus on learning from a single graph, OMG-SSL can handle multiple graphs simultaneously by integrating them into a regularization framework (here a graph can be simply understood as a similarity or correlation matrix). We will show that actually our approach is equivalent to fusing multiple graphs and then conducting semi-supervised learning on the fused graph. Thus, when applying it to integrate multiple modalities, the OMG-SSL scheme can also be viewed as a novel graph-based fusion approach which is different from the existing fusion strategies that perform fusion on features or the results learned from individual modalities [31].

Based on the proposed OMG-SSL algorithm, the video annotation scheme is able to deal with multiple modalities, multiple distance functions, and video temporal consistency in a unified manner, as illustrated in Fig. 1. Given M modalities and D distance functions, we can generate $M \times D$ graphs, following from the fact that the affinity matrix under each pair of modality and distance function corresponds to a graph. Moreover, temporal consistency also indicates the relationship of each sample with its adjacent ones, and it can thus be represented by a certain graph as well. Therefore, OMG-SSL is able to deal with the aforementioned four problems simultaneously, in which the insufficiency of training data is attacked by semi-supervised learning, curse of dimensionality is solved by multimodality learning, and multiple distance functions and temporal consistency are reflected in different graphs. Additionally, we will show that the proposed scheme is computationally more efficient compared with typical existing methods such as support vector machine (SVM), and this advantage is particularly encouraging when annotating a large lexicon of concepts, such as the large scale concept ontology for multimedia (LSCOM) that includes hundreds of concepts [23].

The main contributions of this paper can be summarized as follows.

1) Propose the OMG-SSL algorithm. Different from the existing graph-based learning techniques, which deal with only one graph, the OMG-SSL method optimally explores multiple complementary graphs in the manner of semi-supervised learning.

2) Apply the OMG-SSL algorithm to video annotation, whereby a unified scheme is provided to simultaneously handle large-scale unlabeled data, multiple modalities, multiple distance functions, and video temporal consistency.
3) We demonstrate that the OMG-SSL algorithm can be viewed as a graph-based fusion approach when it is applied to integrate multiple modalities, and it has been demonstrated to be more effective than the existing fusion schemes.

The OMG-SSL approach was first introduced in our previous work [39]. Compared to the preliminary version [39], in this paper we have improvements in three aspects: 1) we performed a more comprehensive survey of existing related works; 2) we conducted more empirical evaluations; and 3) more discussions and analyses are provided. The organization of the rest of this paper is as follows. In Section II, we provide a short review on the related works. In Section III, we propose the OMG-SSL algorithm and its application in video annotation. Experimental results are presented in Section IV. Finally, we conclude this paper in Section V.

II. Related Work

A. Semi-Supervised Learning

Over the recent years, the availability of large data collections associated with only limited human annotation has turned the attention of a growing community of researchers to the topic of semi-supervised learning [5] and [50]. By leveraging unlabeled data based on certain assumptions, semi-supervised learning methods are expected to build more accurate models than those that can be achieved by purely supervised learning methods. Many different semi-supervised learning algorithms have been proposed. Some often-applied ones include self-training, co-training, transductive SVM, and graph-based methods. Extensive reviews of these methods can be found in [5] and [50]. Several of these methods have already been applied in image/video annotation and search. In [36], Tian *et al.* conducted a study on semi-supervised learning-based image retrieval. In [32], co-training is adopted for video annotation based on a careful splitting of visual features. In [43], Yan *et al.* pointed out the weakness of co-training in video annotation, and proposed an improved co-training-style algorithm named semi-supervised cross-feature learning. In [33], Song *et al.* adopted a semi-supervised ensemble learning method for video annotation. Ewerth *et al.* have proposed a semi-supervised video retrieval method that adapts the model trained on labeled samples based on unlabeled data [7]. More recently, graph-based semi-supervised methods have attracted the interest of researchers in this community due to their effectiveness and computational efficiency (most graph-based methods can be implemented with an efficient iterative process). Many works have demonstrated that the graph-based methods are computationally efficient with rather low computational costs. In [12] and [48], a graph-based semi-supervised learning method named learning with local and global consistency (LLGC) [49] is applied to image retrieval and video annotation, respectively. Tang *et al.* proposed a graph-based semi-supervised learning method named kernel linear neighborhood propagation and demonstrated its effectiveness in video annotation [35]. In [40], Wang *et al.* proposed a semi-supervised kernel density estimation method for video annotation and analyzed its relationship to graph-based methods. In [37], Tong *et al.* proposed a scheme to deal with two modalities in graph-based semi-supervised learning scheme. This directly motivates our work in this paper. But later we will show that, different from their approach that adopts fixed weights, our proposed method obtains optimal graph weights, and therefore it is capable of dealing with more graphs.

B. Multimodal Fusion

Existing studies reveal that the distances between sample pairs become increasingly similar when the dimension of the adopted feature space is high [4], [42]. This may introduce performance degradation if we directly apply the high-dimensional features in distance (or similarity)-based learning algorithms, such as the graph-based method adopted in this paper. In the multimedia field, a widely applied approach to addressing this issue is to replace the high-dimensional learning task by multiple low-dimensional learning tasks, i.e., separately apply different modalities to learning algorithms and then fuse the results [42]. Here, a modality can be viewed as a description to video data, such as color, edge, texture, audio, and text (Wu *et al.* [42] also proposed a statistical method to generate modalities without using such prior knowledge). This method is usually called "multimodal fusion" or "multi-modality learning." Sometimes it is also named "late fusion," whereas the approach of using concatenated high-dimensional global feature vector is named "early fusion" [31]. Although the multimodal fusion approach is heuristic, its effectiveness has been empirically demonstrated in many works. With a labeled fusion set, the task of multimodal fusion can actually be formulated as a learning issue. For example, Iyengar *et al.* [52] and Snoek *et al.* [31] have accomplished the fusion with SVM models. But Wang *et al.* [38] have reported that this approach may suffer from the over-fitting problem due to the limited size of fusion set (especially the limited positive samples). Thus generally linear fusion is regarded as a simple yet effective approach. Yan *et al.* have studied the theoretical upper bound of linear fusion [53]. Snoek *et al.* have given an empirical study to compare early fusion and late fusion [31]. Magalhães *et al.* [21] proposed a method to transform multimodal features based on the minimum description length criterion, and the multimodal fusion performance can thus be improved.

We will show that the proposed OMG-SSL method amounts to implementing semi-supervised learning on a fused graph, and it can thus be viewed as a novel "graph-based fusion" approach. Fig. 2 illustrates the schemes of early, late, and graph-based fusion for comparison. From the figure we can see that the graph-based fusion approach is different from early and late fusion in the sense that it explores the complementation of multiple modalities during the learning process. Experimental results will demonstrate the superiority of this approach.

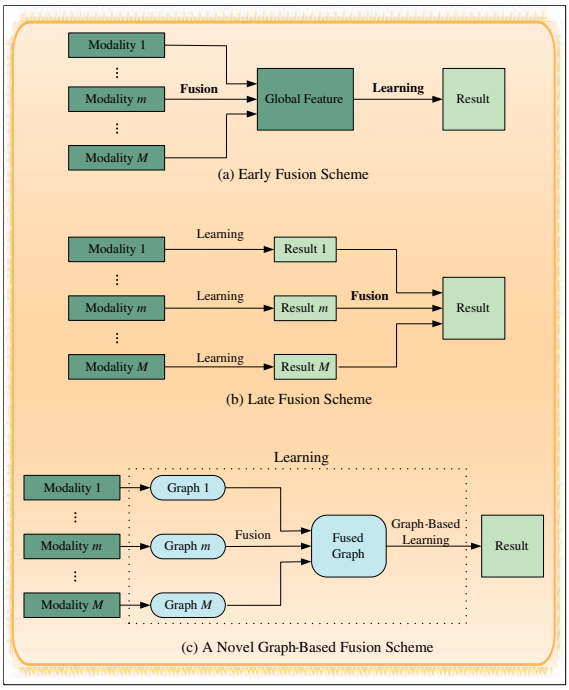

Fig. 2. Comparison of the early, late, and graph-based fusion schemes. We can see that the fusion is performed at different phases in the three approaches.

C. Choice of Distance Function

It is well-known that the distance function plays an important role in machine learning algorithms. In machine learning community, many distance metric learning algorithms have recently been proposed which aim to learn suitable distance functions from training data [9], [14], [46]. However, these methods are usually computationally intensive and prone to overfitting, especially when the training samples are limited and the dimension of feature space is high [46]. Therefore, practically many works tend to select a good distance function from the widely applied ones or combine them according to certain criteria. In image/video annotation tasks, a common sense is that L_1 distance is superior to the others in the Minkowski distance family, including the widely-applied L_2 distance [12], [34]. An explanation is that L_1 distance can better approximate the perceptual difference of visual features [34]. Sebe *et al.* [27] and Yu *et al.* [47] have studied this issue in the maximum likelihood perspective, and they show that the choice should depend on the data distribution. Yu *et al.* further proposed a boosting approach to construct distance function from multiple metrics [47]. This indicates that complementation may exist in different distance functions. Wang *et al.* have proposed a distribution-based distance that incorporates the structures around samples into the distance estimation [41]. In this paper we will explore the complementation of multiple distance functions, including Minkowski and distribution-based distances, in a graph-based learning scheme.

D. Temporal Consistency

It is usually believed that the temporal consistency property, which indicates the structure of video data, can be utilized to improve annotation performance [15], [45]. It indicates that a semantic concept has a large probability to repeat in a continuous video segment, as illustrated in Fig. 3. However, this property is not utilized in most of the previous works. This is because many popular learning methods, such as SVM, are based on i.i.d. assumption and they do not consider this special sample relationship. Song *et al.* have utilized this property for pre-clustering in home video annotation, whereby manual effort can be reduced by only labeling one sample for each cluster in the training set [32]. Kender *et al.* [15] and Yang *et al.* [45] proposed to utilize the property to refine the annotation results in a post-processing procedure. These works have shown considerable improvements in different aspects. In this paper, we will show that the relationship indicated by temporal consistency can be naturally represented in graph form, and therefore it can be directly explored in the OMG-SSL scheme instead of in a post-processing step.

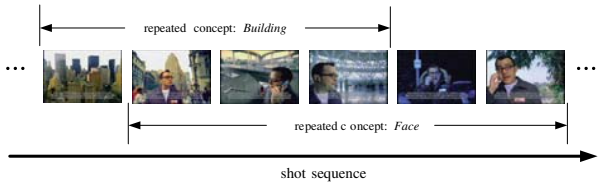

Fig. 3. Exemplary shot sequence from which we can see that semantic concepts have large probability to repeat in continuous video clips.

III. Optimized Multigraph-Based Semi-Supervised Learning

In this section, we present the formulation of OMG-SSL. First, we introduce the traditional single-graph-based semi-supervised learning methods developed on a regularization framework. Then, we show that multiple graphs can be integrated into the regularization framework as well. Tong *et al.* have shown the case of two graphs [37]. Here, we extend it to a general case. We also show that this framework amounts to firstly fusing graphs and then conducting semi-supervised learning on the fused graph using traditional methods. Finally, we further extend the framework to simultaneously optimize fusion weights and the sample labels, namely, the OMG-SSL method.

A. Single-Graph-Based Learning

Graph-based learning is a large family among the existing semi-supervised methods [51]. They are conducted on a graph, where the vertices are labeled and unlabeled samples and the edges reflect the similarities between sample pairs. A function is estimated on the graph based on a label smoothness assumption. These methods have already been successfully applied in image and video content analysis on account of their effectiveness and efficiency [12], [37], [48]. We consider the method proposed in [49]. Denote by $\mathbf{W}$ an affinity matrix with W_{ij} indicating the similarity between the ith and jth sample. This similarity is often estimated based on a distance

function $d(.,.)$ and a positive radius parameter σ, i.e.,

$$W_{ij} = \begin{cases} \exp\left(-\dfrac{d(x_i, x_j)}{\sigma}\right), & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

then a regularization framework is formulated as follows [49]:

$$\arg\min_f \left\{ \sum_{i,j} W_{ij} \left| \frac{f_i}{\sqrt{D_{ii}}} - \frac{f_j}{\sqrt{D_{jj}}} \right|^2 + \mu \sum_i \left| f_i - Y_i \right|^2 \right\} \tag{2}$$

where $\mathbf{D}$ is a diagonal matrix with its (i,i) element equals to the sum of the ith row of $\mathbf{W}$, i.e., $D_{ii} = \sum_j W_{ij}$, and f_i can be regarded as a relevance score. There are two items in this regularization scheme, where the first item implies the smoothness of the labels on the graph and the second term indicates the constraint of training data. After obtaining f_i, we can classify x_i according to its sign, i.e., positive if $f_i > 0$ and negative otherwise. A noteworthy issue here is the setting of Y_i. For general classification task, Y_i is set to 1 if x_i is labeled positive, -1 if x_i is negative, and 0 if x_i is unlabeled. But in our work we set Y_i as follows:

$$Y_i = \begin{cases} 0, & \text{if } x_i \text{ is unlabeled} \\ \frac{1}{frequency} - 1, & \text{if } x_i \text{ is positive sample} \\ -1, & \text{if } x_i \text{ is negative sample} \end{cases} \tag{3}$$

where *frequency* = # *of labeled positive samples*/# *of labeled samples*, i.e., the percentage of positive samples in labeled set. This setting follows from the fact that positive samples are usually less than negative ones, and the distribution of negative samples are usually in a very broad domain. Therefore, positive samples are expected to contribute more in video concept learning. In fact, this setting is equivalent to duplicating $(1/frequency - 1)$ copies for each positive training sample, so that they are balanced with negative ones. It modulates the effect of positive samples and can yield better results.

Let $\mathbf{L} = \mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-1/2}$, which is usually named *normalized graph Laplacian*. Equation (2) then has a closed-form solution as

$$f = \left(\mathbf{I} + \frac{1}{\mu}\mathbf{L}\right)^{-1} Y. \tag{4}$$

However, directly solving (4) involves the inversion of an $n \times n$ matrix, where n is the number of all samples, and the computational cost scales as $O(n^3)$. For computational efficiency, the equation is usually solved by an iterative process as shown in Fig. 4.

The convergence of the iterative process in Fig. 4 can be easily proved based on the fact that the matrix $(\mathbf{I} - \mathbf{L})$, i.e., $\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, is symmetric and its eigenvalues are in $[-1, 1]$. This process is widely known as label propagation or manifold ranking [49].

B. Intuitive Extension to Multiple Graphs

Suppose we have G graphs $\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_G$. Now our problem is how to deal with multiple graphs in semi-supervised learning. Analogous to the approach in [37], we integrate the G graphs into the regularization framework in (2), which thus turns to

1: Initialize $f^{(t)}$ where $t = 0$.
2: Update f by

$$f^{(t+1)} = \frac{1}{1+\mu}(\mathbf{I} - \mathbf{L})f^{(t)} + \frac{\mu}{1+\mu}Y.$$

3: Let $t = t + 1$, and then jump to step 2 until convergence.

Fig. 4. Iterative solution process of the single-graph-based semi-supervised learning.

$$\arg\min_f \left\{ \sum_{g=1}^{G} \alpha_g \left(\sum_{i,j} W_{g,ij} \left| \frac{f_i}{\sqrt{D_{g,ii}}} - \frac{f_j}{\sqrt{D_{g,jj}}} \right|^2 + \mu_g \sum_i \left| f_i - Y_i \right|^2 \right) \right\} \tag{5}$$

where $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_G]$ is a weight vector which satisfies $\alpha_g \geq 0$ and $\sum_{g=1}^{G} \alpha_g = 1$. From (5) we can easily derive that

$$f = \left(\mathbf{I} + \frac{\sum_{g=1}^{G} \alpha_g \mathbf{L}_g}{\sum_{g=1}^{G} \alpha_g \mu_g} \right)^{-1} Y \tag{6}$$

where $\mathbf{L}_g$ is the normalized graph Laplacian obtained from $\mathbf{W}_g$. Then we can see that (6) amounts to firstly fusing $\mathbf{L}_g$ and μ_g as $\mathbf{L}_0 = \sum_{g=1}^{G} \alpha_g \mathbf{L}_g$ and $\mu_0 = \sum_{g=1}^{G} \alpha_g \mu_g$, and then computing f according to (4) by replacing $\mathbf{L}$ and μ with $\mathbf{L}_0$ and μ_0, respectively. Thus, we can conclude that this graph fusion actually amounts to combining normalized graph Laplacians.

C. Formulation of OMG-SSL

Up to now we have shown that multiple graphs can be integrated into a regularization framework, and its solution is equivalent to implementing semi-supervised learning on a fused graph. However, the decision of α_g is not considered in the above framework. This is crucial to the performance of this framework. Since the discriminative abilities may vary intensively among different modalities, α_g should vary as well according to their discriminative abilities. When G is small (say, $G = 2$, as in [37]), we can decide α_g by cross-validation. But when G is large, the searching space for cross-validation increases dramatically, and a more sophisticated strategy is thus required to obtain optimal α_g.

To decide α_g, a most straightforward way is to also regard α_g as variables in (5) and then optimize the regularization framework with respect to both f and α, i.e.,

$$Q(f, \alpha) = \sum_{g=1}^{G} \alpha_g \left(\sum_{i,j} W_{g,ij} \left| \frac{f_i}{\sqrt{D_{g,ii}}} - \frac{f_j}{\sqrt{D_{g,jj}}} \right|^2 + \mu_g \sum_i \left| f_i - Y_i \right|^2 \right)$$

$$[f, \alpha] = \arg\min_{f,\alpha} Q(f, \alpha), s.t. \sum_{g=1}^{G} \alpha_g = 1. \tag{7}$$

However, from (7) we can see that $Q(f,\alpha)$ is linear with respect to α, and its solution is $\alpha_g = 1$ if $g = \arg\min_g f^T \mathbf{L}_g f$ and otherwise $\alpha_g = 0$ (note that the optimal solution of linear programming will always be the extreme points). In other words, only one graph will be kept. Since $f\mathbf{L}_g f$ can be viewed as the smoothness degree of f on the gth graph, it means that a graph will be discarded even if it is merely a little less smooth than another graph. If all the graphs have the same smoothness degrees, i.e., $f^T\mathbf{L}_1 f = f^T\mathbf{L}_2 f = \cdots = f^T\mathbf{L}_G f$, then α_g can be set to arbitrary values, and of course this solution does not fit our goal. To tackle this problem, we make a relaxation by changing α_g to α_g^r, and we thus obtain the formulation of OMG-SSL

$$Q(f,\alpha) = \sum_{g=1}^{G} \alpha_g^r \left(\sum_{i,j} W_{g,ij} \left| \frac{f_i}{\sqrt{D_{g,ii}}} - \frac{f_j}{\sqrt{D_{g,jj}}} \right|^2 + \mu_g \sum_i |f_i - Y_i|^2 \right)$$

$$[f,\alpha] = \arg\min_{f,\alpha} Q(f,\alpha), s.t. \sum_{g=1}^{G} \alpha_g = 1 \tag{8}$$

where $r > 1$. Note that $\sum_{g=1}^{G} \alpha_g^r$ achieves a minimum when $\alpha_g = 1/G$ with the constraint $\sum_{g=1}^{G} \alpha_g = 1$. Therefore, (8) actually makes α_g potentially to be close to each other. The detailed effect of parameter r will be discussed later.

D. The Solution of OMG-SSL

We adopt a process that iteratively updates f and α to minimize $Q(f,\alpha)$, and we will demonstrate the convergence of the process based on the fact that Q is convex with respect to both f and α. Based on (8), we can obtain the partial derivative of Q with respect to f and α as follows:

$$\begin{cases} \dfrac{\partial Q(f,\alpha)}{\partial f} = 2\sum_{g=1}^{G} \alpha_g^r \left(\mathbf{L}_g f + \mu_g (f - Y) \right) \\ \dfrac{\partial Q(f,\alpha)}{\partial \alpha_g} = r\alpha_g^{r-1} \left(f^t \mathbf{L}_g f + \mu_g |f - Y|^2 \right). \end{cases} \tag{9}$$

Thus, when f is fixed, (9) turns to $\arg\min_\alpha Q(f,\alpha)$, $s.t. \sum_{g=1}^{G} \alpha_g = 1$, from which we can derive that

$$\alpha_g = \frac{\left(\frac{1}{f^T\mathbf{L}_g f + \mu_g |f-Y|^2} \right)^{\frac{1}{r-1}}}{\sum_{g=1}^{G} \left(\frac{1}{f^T\mathbf{L}_g f + \mu_g |f-Y|^2} \right)^{\frac{1}{r-1}}}. \tag{10}$$

On the other hand, if α is fixed, (8) turns to $\arg\min_f Q(f,\alpha)$, and f can be solved as

$$f = \left(\mathbf{I} + \frac{\sum_{g=1}^{G} \alpha_g^r \mathbf{L}_g}{\sum_{g=1}^{G} \alpha_g^r \mu_g} \right)^{-1} Y. \tag{11}$$

Now, we show that (11) can also be solved by the iterative solution process in Fig. 4. This is nontrivial because in our practical experiments we will apply the iterative process rather than the closed-form solution for reducing computational cost.

1: Initialize $f = Y$.
2: Update α according to (10).
3: Based on the updated α, re-calculate f according to (11) or the corresponding iterative solution method.
4: Repeat from step 2 until convergence.

Fig. 5. Iterative solution method for OMG-SSL.

To prove this, we let $\mathbf{L}_0 = \sum_{g=1}^{G} \alpha_g^r \mathbf{L}_g / \sum_{g=1}^{G} \alpha_g^r$ and $\mu_0 = \sum_{g=1}^{G} \alpha_g^r \mu_g / \sum_{g=1}^{G} \alpha_g^r$, and (11) then turns to

$$f = \left(\mathbf{I} + \frac{1}{\mu_0} \mathbf{L}_0 \right)^{-1} Y. \tag{12}$$

We replace $\mathbf{L}$ and μ with $\mathbf{L}_0$ and μ_0 in Fig. 4. Since $\mathbf{L}_0$ is symmetric, to prove the convergence of the iterative process, we only need to prove the following fact.

Theorem 1: The eigenvalues of $(\mathbf{I} - \mathbf{L}_0)$ are in $[-1,\ 1]$.

Proof: Let $\beta_g = \alpha_g^r / \sum_{g=1}^{G} \alpha_g^r$, and consequently we have $\mathbf{I} - \mathbf{L}_0 = \sum_{g=1}^{G} \beta_g (\mathbf{I} - \mathbf{L}_g)$ and $\sum_{g=1}^{G} \beta_g = 1$.

Since $(\mathbf{I} - \mathbf{L}_g)$ is symmetric and its eigenvalues are in $[-1, 1]$, $(\mathbf{I} \pm (\mathbf{I} - \mathbf{L}_g))$ are positive semi-definite. Thus, we can derive that $(\mathbf{I} \pm (\mathbf{I} - \mathbf{L}_0)) = \sum_{g=1}^{G} (\mathbf{I} \pm (\mathbf{I} - \mathbf{L}_g))$ are positive semi-definite. Consequently, the eigenvalues of $(\mathbf{I} - \mathbf{L}_0)$ are in $[-1, 1]$. ■

From the above derivation, we can easily form an iterative process to solve f and α by repeatedly updating them as in Fig. 5.

Now, we prove the convergence of this iterative solution process. Denote by f^t and α^t the values of f and α in tth repetition in the process, then we have

$$Q(f^{t+1}, \alpha^{t+1}) < Q(f^t, \alpha^{t+1}) < Q(f^t, \alpha^t) \tag{13}$$

which implies that the cost function $Q(f,\alpha)$ decreases monotonically. Since $Q(f,\alpha) \geq 0$ and it is convex with respect to both f and α, this process is guaranteed to converge to the solution of (8).

We now observe the impact of parameter r. From (10) we can find that r modulates the effect of the smoothness difference of graphs. If $r \to 1$, then the effect of this difference is expanded and only α_g of the smoothest graph is close to 1. Contrarily, if $r \to \infty$, the effect of this difference is reduced, and α_g are close to each other. Therefore, the optimal choice of r should depend on the complementation of these graphs. If rich complementation exists, then r should be large and therefore all graphs can be comprehensively explored, and otherwise r should be small to keep the performance of the "best" graph. In practice, this parameter is decided by cross-validation.

E. Video Annotation Based on OMG-SSL

In this section, we present the OMG-SSL-based video annotation scheme, in which unlabeled data, multiple modalities, multiple distance functions, and temporal consistency are simultaneously taken into consideration. To this end, we show that each modality with a distance function can be represented by a graph, and the temporal consistency property can be explored in graph form as well.

Suppose we have M modalities, and each sample x_i is represented by $x_{i1}, x_{i2}, \ldots, x_{iM}$ for these M modalities, respectively. Consider we have D distance functions $d_1(.,.), d_2(.,.), \ldots, d_D(.,.)$. Then from these M modalities and D distance functions we can generate $M \times D$ graphs as follows:

$$W_{(m-1)\times D+k,ij} = \begin{cases} \exp\left(-\dfrac{d_k(x_i^m, x_j^m)}{\sigma_{(m-1)\times D+k}}\right), & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

where $W_{(m-1)\times D+k}$ is the graph generated by the mth modality and kth distance function.

In this paper we adopt two distance functions: the well-known L_1 distance and the distribution-based distance introduced in [41]. The distribution-based distance between two samples is defined as the symmetric Kullback–Leibler divergence of the neighborhood distributions around the corresponding samples. We use a multivariate normal distribution with mean vector x_i to model the neighbors around x_i, i.e.,

$$p_i(x) = \frac{1}{(2\pi)^{d/2}\left|\mathbf{C}_i\right|^{1/2}} \exp\left(-\frac{1}{2}(x - x_i)^T \mathbf{C}_i^{-1}(x - x_i)\right). \tag{15}$$

The covariance matrix $\mathbf{C}_i$ is estimated as

$$\mathbf{C}_i = \frac{1}{N} \sum_{x_k \in \mathcal{N}_i} (x_k - x_i)(x_k - x_i)^T \tag{16}$$

where $\mathcal{N}_i$ is the set of K neighbors of x_i. The distribution-based distance between x_i and x_j can thus be computed as

$$D_{KL}(p_i, p_j) = \frac{1}{2} tr(\mathbf{C}_i - \mathbf{C}_j)(\mathbf{C}_j^{-1} - \mathbf{C}_i^{-1}) + \frac{1}{2}(x_i - x_j)^T(\mathbf{C}_i^{-1} + \mathbf{C}_j^{-1})(x_i - x_j). \tag{17}$$

From (17), we can see that the distribution-based distance can simultaneously take into account the geometric distance between samples and the structure difference around them, and this makes them potentially superior to the traditional widely applied distances, such as Minkowski distances.

In terms of temporal consistency, we can construct C graphs. Here we use two graphs, i.e., $C = 2$. The first graph simply considers the relationships between every two adjacent units (can be shot or sub-shot [16]), i.e., a unit has high probability to have the same concepts with the previous and the next units. If the indices of these samples are arranged according to temporal relationship, then this sample relationship can be indicated in graph form as

$$W_{M\times D+1,ij} = \begin{cases} 1, & \text{if } i = j+1 \text{ or } i = j-1 \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

the other graph considers the connections of each unit with adjacent six units and assigns different weights to them according to their positions. Specifically, it is defined as

TABLE I
SIX MODALITIES USED IN VIDEO ANNOTATION EXPERIMENTS

Modality 1	225D block-wise color moment
Modality 2	144D HSV correlogram
Modality 3	128D wavelet texture
Modality 4	64D HSV histogram
Modality 5	75D edge direction histogram
Modality 6	16D co-occurrence texture

$$W_{M\times D+2,ij} = \begin{cases} 1, & \text{if } i = j+1 \text{ or } i = j-1 \\ 0.5, & \text{if } i = j+2 \text{ or } i = j-2 \\ 0.25, & \text{if } i = j+3 \text{ or } i = j-3 \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

it is noteworthy that we can also design other graphs to indicate temporal consistency, and all these graphs can be easily integrated since OMG-SSL is a general scheme.

Therefore, the OMG-SSL-based video annotation process consists of two steps: 1) construct $M \times D + C$ graphs, including the $M \times D$ graphs generated from M modalities and D distance metrics and the C graphs indicating temporal consistency; and 2) implement the OMG-SSL algorithm with these $M \times D + C$ graphs.

IV. EXPERIMENTS

A. Experimental Settings

To evaluate the performance of the proposed approach, we conduct experiments on the benchmark video corpus of TRECVID 2005 [1], [28]. The dataset consists of 137 news videos recorded from 13 different programs in English, Arabic, and Chinese [1]. The videos are about 160 h in duration and they are segmented into 49 532 shots and 61 901 sub-shots (the results of shot segmentation have been provided by Petersohn *et al.* [26]). We annotate 39 concepts in the experiments, namely, the LSCOM-Lite concepts [24].

We regard sub-shot as the unit for annotation. A key-frame is selected from each sub-shot, and from each key-frame we extract the following feature sets: 1) block-wise color moment based on 5 by 5 division of the image (225D); 2) HSV correlogram (144D); 3) wavelet texture (128D); 4) HSV histogram (64D); 5) lay-out edge direction histogram (75D); and 6) co-occurrence texture. These six feature sets are regarded as six different modalities, as illustrated in Table I.

As mentioned earlier, we adopt two distance functions, i.e., L_1 distance and distribution-based distance. In the computation of distribution-based distance, we set the neighborhood size K to 20, and the detailed implementations can found in [41].

Following the guideline in [44], we separate the dataset into four partitions, i.e., a "training set" with 90 videos, a "validation set" with 16 videos, a "fusion set" with 16 videos (the fusion set is only used for late fusion), and a "test set" with 15 videos. The four dataset contains 41 847, 7022, 6525 and 6507 sub-shots, respectively. Details about the data partition can be found in [44].

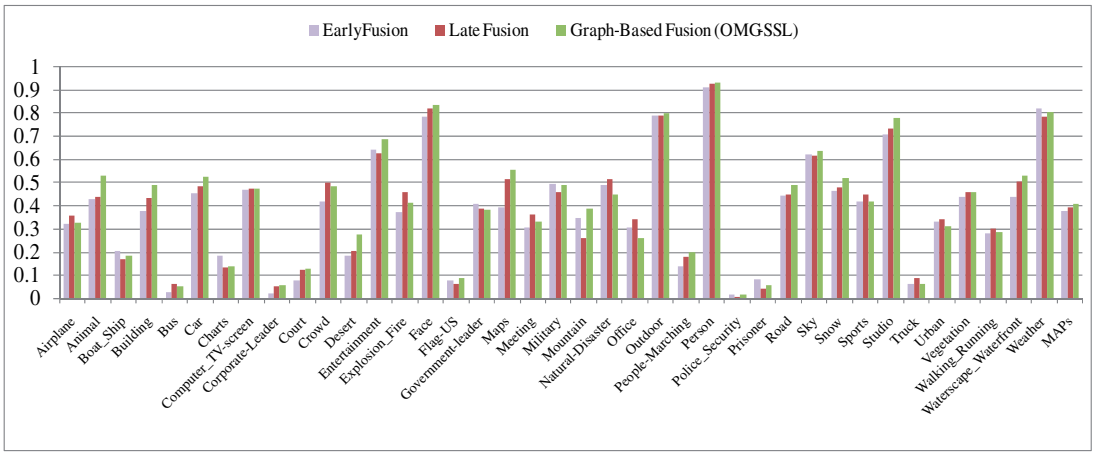

Fig. 6. Performance comparison of early fusion, late fusion, and graph-based fusion with six modalities (using L_1 distance).

Compared with the existing graph-based semi-supervised learning methods that all have parameters σ and μ, OMG-SSL adds only one new parameter r. The parameters are tuned on the validation set. We first decide σ_g and μ_g for each graph, and then decide r for OMG-SSL. In all experiments we adopt the iterative solutions rather than direct solutions. We make matrices $\mathbf{L}_g$ sparse by only keeping N largest values in each row. In our study the parameter N is empirically set to 20. In fact, this parameter can also be further tuned with the validation set such that better performance can be achieved (of course, it will lead to larger computational cost). This is a frequently used strategy in graph-based learning methods, which significantly reduces the computational cost while retaining comparable performance. For performance evaluation, NIST has defined non-interpolated average precision (AP) over a set of retrieved shot as a measure of retrieval effectiveness [23]. Let R be the number of true relevant shots in a set of size S. At any given index j, let R_j be the number of relevant shots in the top j shots. Let $I_j = 1$ if the jth shot is relevant and 0 otherwise. Assuming $R < S$, the AP is then defined as $\frac{1}{R}\sum_{j=1}^{S} I_j R_j/j$. Mean average precision (MAP) is the average of average precisions over all concepts.

B. Experimental Results

1) OMG-SSL With Multiple Modalities: As previously mentioned, OMG-SSL can be viewed as a graph-based fusion approach when dealing with multiple modalities. Thus, here we compare its performance with traditional early and linear late fusion methods (here we have only applied L_1 distance function). In late fusion, the linear weights are tuned on the "fusion set." The results are illustrated in Fig. 6. From the figure we can see that the graph-based fusion approach outperforms the other two fusion methods for most concepts, and the superiority is evident in MAP. Fig. 7 further illustrates the MAP results obtained by learning from six different modalities and those achieved by early, late and graph-based fusion approaches. From the figure we can see that all of the three fusion methods outperform using only one modality, and the graph-based fusion method performs the best.

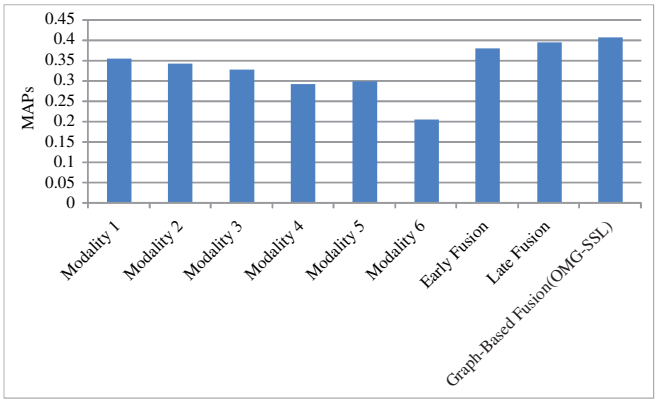

Fig. 7. MAP results obtained by learning from six modalities and early, late and graph-based fusion.

2) OMG-SSL With Multiple Distance Functions: Table II presents the results attained by OMG-SSL with each distance function alone and with two functions together (with all the six modalities). From the table we can see that the distribution-based distance performs better than L_1 distance, which is consistent with the analysis in [41]. We can also see that OMG-SSL with the two distance functions together performs better than that with each individual distance, which indicates that it successfully integrates the two functions to improve performance.

3) Exploiting Temporal Consistency: Based on the results of OMG-SSL with multiple modalities and multiple distance functions, we now further investigate the effectiveness of temporal consistency. Table III shows the performance comparison of the following four methods:

1) OMG-SSL without considering temporal consistency;
2) OMG-SSL with 1st temporal graph, i.e., integrating the graph generated by (18);
3) OMG-SSL with 2nd temporal graph, i.e., integrating the graph generated by (19);
4) OMG-SSL with both two temporal graphs.

From the table it is clear that integrating temporal graphs can improve annotation performance. Although for some

TABLE II
PERFORMANCE COMPARISON OF OMG-SSL WITH DIFFERENT DISTANCE FUNCTIONS. FROM THE TABLE WE CAN SEE THAT OMG-SSL CAN SUCCESSFULLY INTEGRATE MULTIPLE DISTANCE MEASURES. THE BEST RESULT FOR EACH CONCEPT IS SHOWN IN BOLDFACE

Concept	L_1	Distribution -based	Two distance functions
Airplane	0.325	0.307	**0.331**
Animal	**0.530**	0.513	0.520
Boat_Ship	0.182	0.169	**0.183**
Building	0.489	0.486	**0.497**
Bus	0.051	**0.091**	0.057
Car	0.525	0.556	**0.558**
Charts	0.139	0.144	**0.146**
Computer_TV-screen	**0.472**	0.459	0.468
Corporate-Leader	0.055	0.049	**0.056**
Court	0.129	**0.228**	0.221
Crowd	0.484	0.491	**0.499**
Desert	0.277	0.278	**0.284**
Entertainment	0.688	**0.699**	0.694
Explosion_Fire	0.414	0.385	**0.424**
Face	**0.835**	0.830	0.831
Flag-US	0.086	0.098	**0.109**
Government-leader	0.380	**0.394**	0.391
Maps	0.558	0.613	**0.636**
Meeting	0.331	**0.340**	0.338
Military	0.489	0.510	**0.521**
Mountain	0.390	0.428	**0.439**
Natural-Disaster	**0.451**	0.342	0.441
Office	0.261	**0.337**	0.329
Outdoor	0.799	**0.810**	0.808
People-Marching	0.198	**0.206**	**0.206**
Person	**0.934**	0.929	0.926
Police_Security	0.013	0.020	**0.025**
Prisoner	**0.057**	0.056	0.056
Road	0.489	0.505	**0.506**
Sky	0.637	**0.667**	0.660
Snow	0.520	0.508	**0.522**
Sports	0.418	0.443	**0.451**
Studio	0.780	**0.788**	0.782
Truck	**0.063**	0.050	0.069
Urban	0.309	0.341	**0.345**
Vegetation	0.457	0.461	**0.462**
Walking_Running	0.284	0.334	**0.335**
Waterscape_Waterfront	**0.528**	0.491	0.513
Weather	0.807	**0.831**	0.826
MAP	0.406	0.415	**0.422**

TABLE III
PERFORMANCE COMPARISON OF OMG-SSL WITHOUT TEMPORAL GRAPH (NTG), USING 1ST TEMPORAL GRAPH (TG1), USING 2ND TEMPORAL GRAPH (TG2), AND USING TWO TEMPORAL GRAPHS (TG1+TG2). FROM THE RESULTS WE CAN SEE THAT OMG-SSL CAN EXPLORE THE PROPERTY OF TEMPORAL CONSISTENCY TO IMPROVE ANNOTATION PERFORMANCE. THE BEST RESULT FOR EACH CONCEPT IS SHOWN IN BOLDFACE

Concept	nTG	TG1	TG2	TG1+TG2
Airplane	0.331	0.362	0.354	**0.369**
Animal	0.520	0.536	0.537	**0.539**
Boat_Ship	0.183	0.184	0.185	**0.186**
Building	0.497	0.496	0.499	**0.501**
Bus	**0.057**	**0.057**	**0.057**	**0.057**
Car	0.558	0.568	0.573	**0.571**
Charts	0.146	0.147	0.144	**0.148**
Computer_TV-screen	0.468	0.476	0.475	**0.479**
Corporate-Leader	0.056	0.057	0.058	**0.060**
Court	0.221	0.227	**0.235**	0.229
Crowd	0.499	0.501	0.501	**0.502**
Desert	0.284	0.303	0.302	**0.306**
Entertainment	0.694	0.707	**0.718**	0.711
Explosion_Fire	0.424	0.447	0.440	**0.454**
Face	**0.831**	0.828	0.829	0.827
Flag-US	**0.109**	0.098	0.098	0.097
Government-leader	0.391	0.431	0.429	**0.435**
Maps	**0.636**	0.621	0.624	0.628
Meeting	0.338	0.382	**0.389**	0.386
Military	0.521	0.534	**0.545**	0.536
Mountain	0.439	0.440	**0.443**	0.439
Natural-Disaster	0.441	0.481	0.478	**0.483**
Office	0.329	0.338	**0.343**	0.330
Outdoor	0.808	0.813	**0.815**	0.814
People-Marching	0.206	0.204	0.215	**0.217**
Person	0.926	0.925	0.926	**0.930**
Police_Security	0.025	**0.026**	0.025	0.025
Prisoner	**0.056**	0.054	0.055	0.054
Road	0.506	0.520	0.518	**0.522**
Sky	0.660	0.666	0.665	**0.670**
Snow	0.522	**0.525**	0.524	0.524
Sports	0.451	0.481	**0.494**	0.493
Studio	0.782	0.772	0.773	**0.783**
Truck	**0.069**	0.064	0.066	0.064
Urban	0.345	0.343	0.342	**0.346**
Vegetation	0.462	0.478	0.474	**0.483**
Walking_Running	0.335	0.341	0.340	**0.349**
Waterscape_Waterfront	0.513	0.513	**0.517**	0.515
Weather	0.826	**0.869**	0.867	0.867
MAP	0.422	0.431	0.432	**0.434**

concepts the improvements are small in magnitude, they are fairly consistent in sign. Note that the MAP measures obtained by using 1st temporal graph and using 2nd temporal graph are 0.431 and 0.432, respectively, whereas the MAP obtained by using two temporal graphs is 0.434. This indicates that the complementation exists in the two temporal graphs as well. To be clear, MAP of 0.434 is the final performance achieved by OMG-SSL with all the 14 graphs on this video annotation task.

To further demonstrate the effectiveness, we compare the results obtained by OMG-SSL with the Columbia374 concept detectors [44]. Columbia374 is a public baseline system[1] that is developed using SVM with three visual feature sets, including block-wise color moment, edge direction histogram, and Gabor texture. To make a fair comparison, in OMG-SSL we only apply the color moment and edge direction histogram, features. We use the L_1 distance and the distribution-based distance as well as the two temporal graphs, i.e., six graphs have been used in all. The results are illustrated in Table IV. From the table we can see that, even with fewer features, OMG-SSL can outperform the Columbia374 for most concepts.

4) Impact of Parameter r: To investigate the effect of r, we illustrate the performance variations of OMG-SSL with respect to r for several concepts in Fig. 8 (with all the 14 graphs). Here we have only illustrated the results of three

[1]There are also several other such public baselines, such as VIREO-374 [13] and Mediamill-101 [29]. But in this paper we only compare our results with Columbia374 since they are under the same experimental settings.

TABLE IV
PERFORMANCE COMPARISON OF OMG-SSL AND COLUMBIA374. THE BEST RESULT FOR EACH CONCEPT IS SHOWN IN BOLDFACE

Concept	Columbia374 (SVM)	OMG-SSL
Airplane	0.361	**0.366**
Animal	0.311	**0.534**
Boat_Ship	**0.208**	0.183
Building	0.454	**0.459**
Bus	**0.092**	0.029
Car	0.456	**0.542**
Charts	0.132	**0.134**
Computer_TV-screen	0.434	**0.462**
Corporate-Leader	0.029	**0.052**
Court	0.113	**0.226**
Crowd	**0.481**	0.475
Desert	0.277	**0.307**
Entertainment	0.630	**0.699**
Explosion_Fire	0.420	**0.470**
Face	0.795	**0.806**
Flag-US	**0.080**	0.074
Government-leader	0.412	**0.421**
Maps	**0.699**	0.560
Meeting	**0.391**	0.357
Military	0.392	**0.515**
Mountain	0.314	**0.412**
Natural-Disaster	0.248	**0.483**
Office	0.288	**0.320**
Outdoor	0.786	**0.802**
People-Marching	0.138	**0.183**
Person	**0.936**	0.911
Police_Security	0.014	**0.015**
Prisoner	0.004	**0.054**
Road	0.447	**0.487**
Sky	0.600	**0.654**
Snow	0.499	**0.558**
Sports	0.407	**0.451**
Studio	**0.786**	0.751
Truck	**0.135**	0.058
Urban	0.301	**0.312**
Vegetation	0.423	**0.443**
Walking_Running	0.280	**0.345**
Waterscape_Waterfront	**0.494**	**0.494**
Weather	0.763	**0.860**
MAP	0.388	**0.418**

concepts, namely, *Airplane*, *Building*, and *Maps*, but similar phenomena can be observed for other concepts as well. From the figure we can see that the optimal choice of r is concept-dependent and the performance curves exhibit a "$\wedge$" shape as r increases from 1 to ∞. As discussed in Section III-D, this is because the complementation of graphs has not been sufficiently explored when r is near 1, and contrarily the graphs are nearly averagely fused when r is too large. Thus, we have to tune the parameter r for each concept using cross-validation in practical experiments.

5) Performance Variation in the Iterative Solution Process: Fig. 9 presents the MAP results with different iterations in the iterative process of OMG-SSL. From the figure we can see that the performance consistently improves as the iteration number increases. But the performance curve converges fast, and the improvement becomes very limited after five iterations. In our experiments, we set the iteration time to 6.

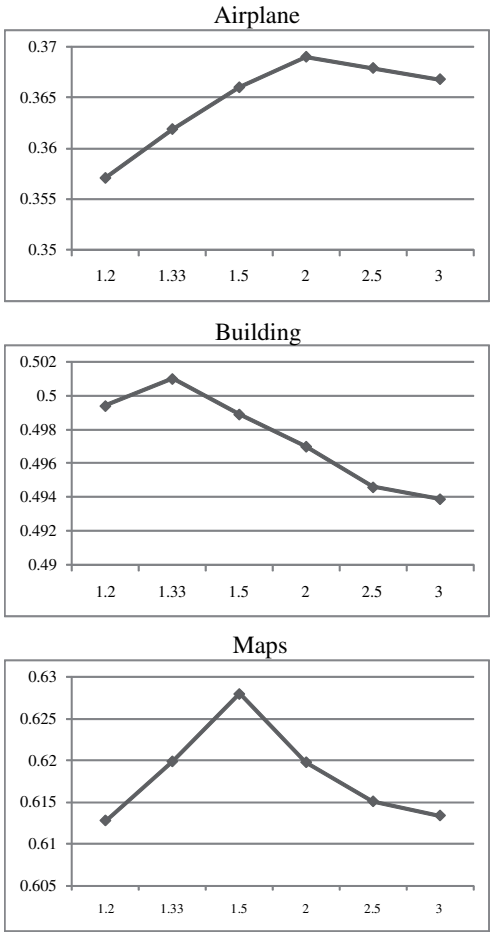

Fig. 8. Performance curves with respect to r for different concepts.

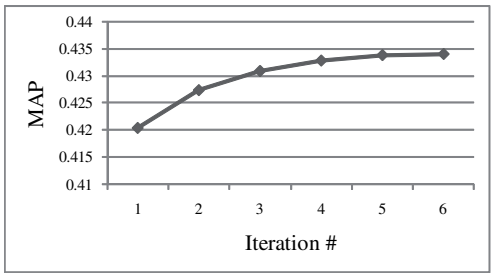

Fig. 9. Performance comparison with different iteration time.

6) Performance of OMG-SSL With Different Sizes of Labeled Data: We also conduct experiments to study whether the effectiveness of OMG-SSL will depend on the size of training data and the relative percentages of labeled and unlabeled data. We randomly select l labeled samples from the original training set and the other samples are regarded as unlabeled. For the consistency of comparison, we use the same experimental settings of the other parts except that we have reduced the number of labeled samples, i.e., we use the original validation set, fusion set, and testing set to tune parameters, fuse multiple modalities and evaluate performance, respectively. We set different l and perform 10 trials for each l to obtain average results. Fig. 10 illustrates the MAP curves of OMG-SSL

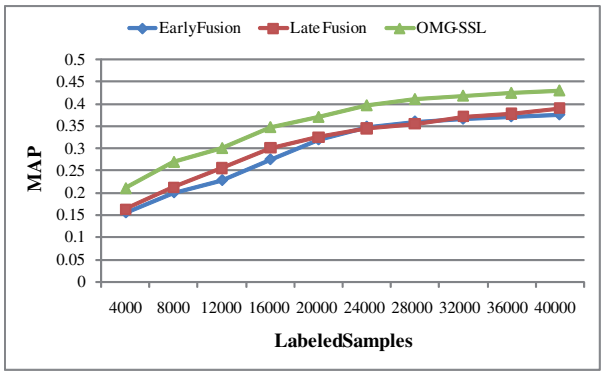

Fig. 10. Performance variation of OMG-SSL with respect to the sizes of labeled data and its comparison with early fusion and late fusion approaches.

TABLE V
PRACTICAL VALUES OF THE NOTATIONS IN THE EXPERIMENTS OF VIDEO ANNOTATION

Notation	Description	Value
n	Number of samples	61901
d	Dimension of low-level feature space	652
D	Number of modalities	6
G	Number of graphs	14
N	Nonzero entries in each row in $\mathbf{L}_g$	20
T_1	Iteration times in the process in Fig. 4	50
T_2	Iteration times in the process in Fig. 5	6

with all the 14 graphs and the early fusion and late fusion approaches with the six modalities. From the figure we can see that the performance of the three approaches keep improving as the labeled data increase, and the OMG-SSL consistently outperforms the other two methods.

C. Computational Efficiency

The computational cost of OMG-SSL mainly consists of two parts, one is for graph construction, and the other is for the iterative solution of the regularization framework. In fact, these two steps can be viewed as "construction" and "inference" procedures of OMG-SSL, respectively. We can easily derive that the computational cost of graph construction is $O(D \times d \times n^2)$, where d is the dimension of global low-level feature vector (including M modalities), and the cost of the iterative solution method is $O(G \times T_1 \times T_2 \times n \times N)$, where G is the number of graphs, and T_1 and T_2 are the respective iteration times in the processes in Figs. 4 and 5, respectively. We illustrate the definitions of all these notations and their detailed values in our video annotation experiments in Table V for clarity.

Obviously "inference" is much more rapid than the "construction" procedure. But an encouraging property of OMG-SSL is that the "construction" is a concept-independent step, i.e., the graphs only have to be constructed once and then they can be utilized for all concepts. Compared with traditional methods those need to train a model for each individual concept, such as SVM, OMG-SSL has great advantage in terms of efficiency when dealing with multiple concepts. For instance, the computational cost of training a SVM model scales as nearly $O(l^3)$, where l is the size of training set. Furthermore, the cost is proportional to the lexicon size, and it would thus be prohibitive if we have to annotate a large lexicon of concepts, such as the LSCOM [23]. Contrarily, OMG-SSL only needs to repeat its efficient testing procedure for different concepts, and thus its computational cost will not increase dramatically. This property makes OMG-SSL particularly appropriate for large-scale annotation, in terms of both dataset size and lexicon size.

It is worth noting that OMG-SSL also has certain weakness in terms of computation in comparison with the traditional methods such as SVM. As a semi-supervised method, OMG-SSL has mixed the training and testing phrases and it has difficulty in dealing with newly coming data, i.e., out-of-sample data, since it has to reconstruct graphs for modeling. On the contrary, most supervised methods only have to test the new samples with the existing model. However, recently several semi-supervised induction methods have been investigated which are able to directly induce the labels of out-of-sample data without the model reconstruction process [6], and these methods can be directly applied with OMG-SSL to address the difficulty.

D. Generic Applicability

In this section, we apply it to another task, i.e., person identification from webcam images. This test will demonstrate that OMG-SSL is actually a general framework that can be applied in many applications besides video annotation.

In [3], Balcan *et al.* have demonstrated the application of graph-based semi-supervised learning in person identification of webcam images. They have shown that the knowledge from different domains should be sufficiently explored in the designed graph. Here we conduct experiments on the same dataset as used in [3], i.e., FreeFoodCam, to show that the performance can be further improved if we develop multiple graphs to encode knowledge from different domains and apply OMG-SSL to integrate these graphs.

The FreeFoodCam dataset consists of 5254 images, which are captured in a public lounge in the Carnegie Mellon University. In each image there is one and only one person, and there are 10 different persons in the whole image set in all. Thus the person identification problem from these images is naturally a 10-way classification task. More information about the dataset can be found in [3]. Balcan *et al.* [3] proposed to adopt graph-based semi-supervised learning in this task, and the graph is designed based on the following knowledge.

1) Time. Two images are connected if their time difference is less than t_1 (note that the capturing date and time of each image have been recorded).
2) Color. The 100D color histogram is extracted from each image. The cosine similarities between histograms are estimated. Then two images are connected if their time difference is less than t_2 and one is in the k_c-neighborhood of the other.
3) Face. A square face image is extracted by a face detector from each image. Then two images are connected if one face image is in k_f-neighborhood of the other (in terms of pixel-wise Euclidean distance).

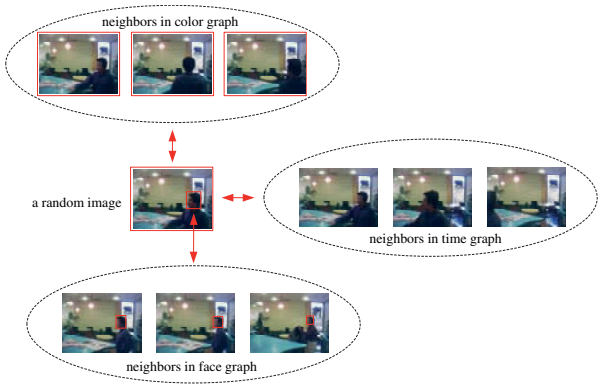

Fig. 11. Random image and its neighbors in three graphs. We can see that a sample has different neighbors in different graphs and this indicates the complementary nature of the graphs.

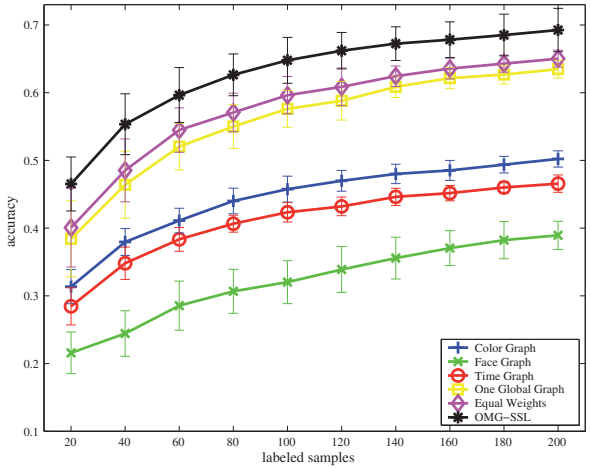

Fig. 12. Performance comparison of six methods for person identification from webcam images.

In [3], equal weights are assigned to all the edges in the graph. But Balcan *et al.* also mentioned that appropriately modulating the effect of different knowledge can further improve performance. According to our previous analysis, OMG-SSL is capable of dealing with this issue. To verify it, we develop three graphs, i.e., time graph, color graph and face graph, according to the corresponding domain knowledge, and then apply OMG-SSL to integrate them. Fig. 11 illustrates an image and its neighbors in three different graphs. From the figure we can see that rich complementation exists in these three graphs.

We compare the following six methods:

1) time graph only;
2) color graph only;
3) face graph only;
4) one global graph integrating all knowledge, i.e., the method proposed in [3];
5) graph fusion with equal weights, i.e., $\alpha_g = 1/3$;
6) OMG-SSL.

The first three methods only utilize an individual graph, and the other three methods are different knowledge fusion approaches. In all experiments we adopt the similar settings as those applied in [3], i.e., $t_1 = 2$ s, $t_2 = 12$ h, $k_c = 3$, $k_f = 3$. The parameter μ is empirically set to 50 and the parameter r is decided by 10-fold cross-validation. We gradually increase the labeled set size from 20 to 200. For each size, we perform 20 trials and in each trial we randomly select labeled samples from *the first day of a person's appearance only*, which follows the guideline of [3]. It is worth mentioning that in the previous discussion about OMG-SSL, we have only considered the case of binary classification. This is because video annotation is always formulated as a binary classification task for each concept. But OMG-SSL is capable of dealing with multiple classes as well. We only have to extend f_i to be a vector, and details can be found in [40].

Fig. 12 illustrates the classification performance of different methods. Consistent with intuition, we can see that the last three methods remarkably outperform the first three, i.e., integrating knowledge from different domains is beneficial. Meanwhile, we can see that OMG-SSL performs much better than the other two knowledge fusion methods. This indicates that OMG-SSL is able to appropriately modulate the effects of different knowledge sources and thus leads to much better performance than fusing them equally.

V. Conclusion

In this paper we have proposed an OMG-SSL algorithm, which is able to integrate multiple complementary graphs into a regularization framework. We have proven that it is equivalent to conducting semi-supervised learning on an optimally fused graph. In this way, the complementation of multiple graphs can be explored and the learning performance can be thus improved. Based on this algorithm, we provided a novel efficient video annotation scheme, in which large-scale unlabeled data, multiple modalities, multiple distance functions, and video temporal consistency could be simultaneously tackled in a unified manner. We have also shown that the proposed method could be viewed as a graph-based fusion approach when it is applied to fuse multiple modalities. Extensive experiments have demonstrated the effectiveness of the proposed approach.

It is worth noting that the OMG-SSL is actually a general approach and can be applied in many domains besides video annotation. In this paper we have also demonstrated its application in a person identification task. Furthermore, the proposed scheme is flexible and can be easily extended through utilizing more graphs. For example, the demonstrated video annotation performance can be easily improved by extracting more features, integrating more distance functions, and designing more graphs to explore temporal consistency.

References

[1] TRECVID: TREC Video Retrieval Evaluation. [Online]. Available: http://www.nlpir.nist.gov/projects/trecvid

[2] A. Amir, J. Argillander, M. Campbell, A. Haubold, G. Iyengar, S. Ebadollahi, F. Kang, M. R. Naphade, A. Natsev, J. R. Smith, J. Tesic, and T. Volkmer, "IBM research TRECVID-2005 video retrieval system," in *Proc. TREC Video Retrieval Evaluation*, Gaithersburg, MD, 2005, pp. 1–17.

[3] M. F. Balcan, A. Blum, P. P. Choi, J. Lafferty, B. Pantano, M. R. Rwebangira, and X. Zhu, "Person identification in webcam images: An application of semi-supervised learning," in *Proc. Int. Conf. Machine Learning Workshop on Learning from Partially Classified Training Data*, Bonn, Germany, 2005, pp. 1–9.

[4] K. Beyer, J. Goldstein, and U. Shaft, "When is nearest neighbor meaningful?" in *Proc. Int. Conf. on Database Theory*, Jerusalem, Israel, 1999, pp. 217–235.

[5] O. Chapelle, A. Zien, and B. Scholkopf, *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006.

[6] O. Delalleau, Y. Bengio, and N. L. Roux, "Efficient non-parametric function induction in semi-supervised learning," in *Proc. Artificial Intell. and Statist.*, Barbados, 2005, pp. 96–103.

[7] R. Ewerth and B. Freisleben, "Semi-supervised learning for semantic video retrieval," in *Proc. ACM Int. Conf. Image and Video Retrieval*, Amsterdam, The Netherlands, 2007, pp. 154–161.

[8] S. L. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *Proc. Int. Conf. Comput. Vision and Pattern Recognition*, Washington, DC, 2004, pp. 1002–1009.

[9] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighborhood component analysis," in *Proc. Advances Neural Inform. Process.*, Whistler, BC, 2005, pp. 571–577.

[10] A. G. Hauptmann, "Lessons for the future from a decade of informedia video analysis research," in *Proc. ACM Int. Conf. Image and Video Retrieval*, Singapore, 2005, pp. 1–10.

[11] A. G. 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.

[12] J. R. He, M. J. Li, H. J. Zhang, H. H. Tong, and C. S. Zhang, "Manifoldranking based image retrieval," in *Proc. ACM Multimedia*, New York, NY, 2004, pp. 9–16.

[13] Y. G. Jiang, C. W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. ACM Int. Conf. Image and Video Retrieval*, Amsterdam, The Netherlands, 2007, pp. 494–501.

[14] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. Advances Neural Inform. Process.*, Whistler, BC, 2006, pp. 1473–1480.

[15] J. R. Kender and M. R. Naphade, "Video news shot labeling refinement via shot rhythm models," in *Proc. Int. Conf. Multimedia & Expo*, Toronto, ON, 2003, pp. 37–40.

[16] J. G. Kim, H. S. Chang, J. Kim, and H. M. Kim, "Efficient camera motion characterization for MPEG video indexing," in *Proc. Int. Conf. Multimedia & Expo*, vol. 2.New York, NY, 2000, pp. 1171–1174.

[17] W. Kraaij and P. Over. "TRECVID-2005 high-level feature task: Overview," in *Proc. TRECVID*, Gaithersburg, MD, 2005.

[18] X. Li, D. Wang, J. Li, and B. Zhang, "Video search in concept subspace: A text-like paradigm," in *Proc. ACM Int. Conf. Image and Video Retrieval*, Amsterdam, The Netherlands, 2007, pp. 603–610.

[19] C. Y. Lin, M. Naphade, A. Natsev, C. Neti, J. R. Smith, B. Tseng, H. J. Nock, and W. Adams, "User-trainable video annotation using multimodal cues," in *Proc. ACM SIGIR Conf. Research and Development Inform. Retrieval*, Toronto, Canada, 2003, pp. 403–404.

[20] C. Y. Lin, B. Tseng, and J. R. Smith, "VideoAnnEx: IBM MPEG-7 annotation tool for multimedia indexing and concept learning," in *Proc. Int. Conf. Multimedia & Expo*, Baltimore, MD, 2003.

[21] J. Magalhaes and S. Ruger, "Information-theoretic semantic multimedia indexing," in *Proc. ACM Int. Conf. Image and Video Retrieval*, Amsterdam, The Netherlands, 2007, pp. 619–626.

[22] X. Mu, "Content-based video retrieval: Does video's semantic visual feature matter?" in *Proc. ACM SIGIR Conf. Research and Development Informa. Retrieval*, Seattle, WA, 2006, pp. 679–680.

[23] M. Naphade, J. R. Smith, J. Tesic, S. F. Chang, W. Hsu, A. Hauptmann, and J. Curtis, "LSCOM lexicon definitions and annotations version 1.0. dto challenge workshop on large scale concept ontology for multimedia," *IEEE Multimedia*, vol. 13, no. 3, pp. 86–91, Jul.–Sep. 2006.

[24] M. R. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over and A. Hauptmann, "A light scale concept ontology for multimedia understanding for TRECVID 2005." IBM, Yorktown Heights, NY, IBM Research Tech. Rep., 2005.

[25] M. R. Naphade and J. R. Smith, "On the detection of semantic concepts at TRECVID," in *Proc. ACM Multimedia*, New York, NY, 2004, pp. 660–667.

[26] C. Petersohn, "Fraunhofer from HHI at TRECVID 2004: Shot boundary detection system," in *Proc. TRECVID Workshop*, Gaithersburg, MD, 2004, pp. 1–7.

[27] N. Sebe, M. S. Lew, and D. P. Huijsmans, "Toward improved ranking metrics," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 22, no. 10, pp. 1132–1143, Oct. 2000.

[28] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proc. ACM Workshop on Multimedia Inform. Retrieval*, Santa Barbara, CA, 2000, pp. 321–330.

[29] C. G. Snoek, M. Worring, J. C. Gemert, J. M. Geusebroek, and A. W. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *Proc. ACM Multimedia*, Santa Barbara, CA, 2006, pp. 421–430.

[30] C. G. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. W. M. Smeulders, "The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1678–1689, Oct. 2006.

[31] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. ACM Multimedia*, Singapore, 2005, pp. 399–402.

[32] Y. Song, X. S. Hua, L. R. Dai, and M. Wang, "Semi-automatic video annotation based on active learning with multiple complementary pre dictors," in *Proc. ACM Workshop Multimedia Inform. Retrieval*, Singapore, 2005, pp. 97–104.

[33] Y. Song, X. S. Hua, G. J. Qi, L. R. Dai, M. Wang, and H. J. Zhang, "Efficient semantic annotation method for indexing large personal video database," in *Proc. ACM Workshop on Multimedia Inform. Retrieval*, Santa Barbara, CA, 2006, pp. 289–296.

[34] M. Stricker and M. Orengo, "Similarity of color images," in *Proc. Storage and Retrieval for Image and Video Databases (SPIE 2420)*, San Diego, CA, 1995, pp. 381–392.

[35] J. Tang, X. S. Hua, G. J. Qi, Y. Song, and X. Wu, "Video annotation based on kernel linear neighborhood propagation," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 620–628, Jun. 2008.

[36] Q. Tian, J. Yu, Q. Xue, and N. Sebe, "A new analysis of the value of unlabeled data in semi-supervised learning in image retrieval," in *Proc. Int. Conf. Multimedia & Expo*, vol. 2. Taipei, Taiwan, Jun. 2004, pp. 1019–1022.

[37] H. Tong, J. R. He, M. J. Li, C. S. Zhang, and W. Y. Ma, "Graph-based multi-modality learning," in *Proc. ACM Multimedia*, Singapore, 2005, pp. 862–871.

[38] D. Wang, X. Liu, L. Luo, J. Li, and B. Zhang, "Video diver: Generic video indexing with diverse features," in *Proc. ACM Workshop Multimedia Inform. Retrieval*, Augsburg, Germany, 2007, pp. 61–70.

[39] M. Wang, X. S. Hua, X. Yuan, Y. Song, and L. R. Dai, "Optimizing multi-graph learning: Towards a unified video annotation scheme," in *Proc. ACM Multimedia*, Augsburg, Germany, 2007, pp. 862–871.

[40] M. Wang, X. S Hua, T. Mei, R. Hong, G. Qi, Y. Song, and L. R. Dai, "Semi-supervised kernel density estimation for video annotation," *Comput. Vision and Image Understanding*, vol. 113, no. 3, pp. 384–396, 2009.

[41] M. Wang, T. Mei, X. Yuan, Y. Song, and L. R. Dai, "Video annotation by graph-based learning with neighborhood similarity," in *Proc. ACM Multimedia*, Augsburg, Germany, 2007, pp. 325–328.

[42] Y. Wu, E. Y. Chang, K. C. C. Chang, and J. R. Smith, "Optimal multimodal fusion for multimedia data analysis," in *Proc. ACM Multimedia*, New York, NY, 2004, pp. 572–579.

[43] R. Yan and M. R. Naphade, "Semi-supervised cross feature learning for semantic concept detection in videos," in *Proc. Int. Conf. Comput. Vision and Pattern Recognition*, San Diego, CA, 2005, pp. 657–663.

[44] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu, "Columbia university's baseline detectors for 374 LSCOM semantic visual concepts." Columbia University, New York, NY, ADVENT Tech. Rep. #222-2006-8, 2007.

[45] J. Yang and A. G. Hauptmann, "Exploring temporal consistency for video analysis and retrieval," in *Proc. ACM Workshop Multimedia Inform. Retrieval*, Santa Barbara, CA, 2006, pp. 33–42.

[46] L. Yang, R. Jin, R. Sukthankar, and Y. Liu, "An efficient algorithm for local distance metric learning," in *Proc. AAAI Conf. Artificial Intell.*, Boston, MA, 2006, pp. 543–548.

[47] J. Yu, J. Amores, N. Sebe, and Q. Tian, "Toward robust distance metric analysis for similarity estimation," in *Proc. Int. Conf. Comput. Vision and Pattern Recognition*, New York, NY, 2006, pp. 316–322.

[48] X. Yuan, X. S. Hua, M. Wang, and X. Wu, "Manifold-ranking based video concept detection on large database and feature pool," in *Proc. ACM Multimedia*, Santa Barbara, CA, 2006, pp. 623–626.

[49] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," in *Proc. Advances of Neural Inform. Process.*, Whistler, BC, 2004.
[50] X. Zhu, "Semi-supervised learning literature survey," Univ. Wisconsion-Madison, Madison, WI, Tech. Rep. 1530, 2008.
[51] X. Zhu, "Semi-supervised learning with graphs," Ph.D. Thesis, Department of Psychology, Carnegie Mellon Univ., Pittsburgh, PA, 2005.
[52] G. Iyengar, H. J. Nock, and C. Neti, "Discriminative model fusion for semantic concept detection and annotation in video," in *Proc. ACM Multimedia*, Berkeley, CA, 2003, pp. 255–258.
[53] R. Yan and A. G. Hauptmann, "The combination limit in multimedia retrieval," in *Proc. ACM Multimedia*, Berkeley, CA, 2003, pp. 339–342.

Meng Wang received the B.E. and Ph.D. degrees in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively.

Since July 2008, he has been an Associate Researcher in Microsoft Research Asia, Beijing, China. His current research interests include multimedia content analysis, computer vision and pattern recognition.

Xian-Sheng Hua (M') received the B.S. and Ph.D. degrees from Peking University, Beijing, China, in 1996 and 2001, respectively, both in applied mathematics. When he was in Peking University, his major research interests were in the areas of image processing and multimedia watermarking.

Since 2001, he has been with Microsoft Research Asia, Beijing, China, where he is currently a Lead Researcher with the Internet Media group. His current interests are in the areas of video content analysis, multimedia search, management, authoring, sharing, and advertising. He has published more than 130 publications in these areas and has more than 30 filed patents or pending applications He is also an Adjunct Professor of the University of Science and Technology of China, Hefei, China. He won the Best Paper Award and Best Demonstration Award in ACM Multimedia 2007 and the Best Poster Paper Award in the 2008 IEEE International Workshop on Multimedia Signal Processing. He also won the 2008 MIT Technology Review TR35 Young Innovator Award.

Dr. Hua serves as an Associate Editor of *IEEE Transactions on Multimedia* and as an Editorial Board Member of *Multimedia Tools and Applications*. He is a member of the Association for Computing Machinery.

Richard Hong received the Ph.D. degree in March 2008 from the University of Science and Technology of China (HKUST), Hefei, China.

He is currently with HKUST. He is also a Research Assistant in the School of Computing, National University of Singapore. From Feb. 2006 to Jun. 2006, he worked as a research intern in the Web Search and Data Mining group at Microsoft Research Asia, Beijing, China. His current research interests include content-based image retrieval, video content analysis, and pattern recognition.

Dr. Hong is a member of ACM.

Jinhui Tang received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively, both in electronic engineering and information science.

Since July 2008, he has been a Research Fellow in the School of Computing, National University of Singapore. His current research interests include content-based image retrieval, video content analysis and pattern recognition. He is a recipient of the 2008 President's Scholarship of the Chinese Academy of Science, and a co-recipient of the Best Paper Award in ACM Multimedia 2007.

Guo-Jun Qi received the B.E. degree in automation from the University of Science and Technology of China, Hefei, in 2005.

He is now working in the Internet Media Group at Microsoft Research Asia, Beijing, China, as a Research Intern. His research interests include computer vision, multimedia, and machine learning, especially content-based image/video retrieval, analysis, management and sharing.

Mr. Qi was the winner of the Best Paper Award at the 15th ACM International Conference on Multimedia, Augsburg, Germany, 2007. He is a Student Member of the Association for Computing Machinery.

Yan Song received the Ph.D. degree in Electronic Engineering from the University of Science and Technology of China in 2006.

Since 1997, he has been an Assistant Professor in the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China. His main research interests include multimedia information processing and video content analysis.