

# Multi-Modality Transfer based on Multi-Graph Optimization for Domain Adaptive Video Concept Annotation

Shaoxi Xu<sup>1,2</sup>, Sheng Tang<sup>1</sup>, Yongdong Zhang<sup>1</sup>, Jintao Li<sup>1</sup>

<sup>1</sup>*Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, P.R. China*

<sup>2</sup>*Graduate University of Chinese Academy of Sciences, Beijing, 100190, P.R. China*

{xushaoxi,ts, jtli, zhyd}@ict.ac.cn

## Abstract

*Multi-modality, the unique and important property of video data, is typically ignored in existing video adaptation processes. To solve this problem, we propose a novel approach, named multi-modality transfer based on multi-graph optimization (MMT-MGO) in this paper, which leverages multi-modality knowledge generalized by auxiliary classifiers in the source domain to assist multi-graph optimization (a graph-based semi-supervised learning method) in the target domain for video concept annotation. To our best knowledge, it is the first time to introduce multi-modality transfer into domain adaptive video concept detection and annotation. Moreover, we propose an efficient incremental extension scheme to sequentially estimate a small batch of new emerging data without modifying the structure of multi-graph scheme. The proposed scheme can achieve a comparable accuracy with that of the brand-new round optimization which combines these data with the data corpus for the nearest round optimization, while the time for estimation has been greatly reduced. Extensive experiments over TRECVID2005-2007 data sets demonstrate the effectiveness of both the multi-modality transfer scheme and the incremental extension scheme.*

## 1. Introduction

Video concept detection and annotation is one of the most important techniques in automatic video content analysis. However, the emergence of explosive amount image/video data gives a big challenge to this task. On the one hand, expensive and time-consuming human labor is required for the collection of labels of new emerging training data. On the other hand, there exist a large amount of outdated models and labeled training data for previous tasks (usually referred to as auxiliary models/classifiers and auxiliary data respectively in the field of cross domain adaptation). Under the circumstances, domain adaptive concept detection and annotation emerges as one of major techniques to accommodate to such dilemma, which aims to adapt models built for concept detection in source domain to target domain. In general, in the field of video content analysis, a domain refers to a video genre, a content

provider, or a video program [6]. Different with conventional concept detection methods, the distinction of data distribution between source and target domain should be considered in the process of adaptation.

Recently, many domain adaptation techniques have been proposed to solve the problem of distribution mismatch in the field of video concept detection and annotation. However, in the process of adaptation all the features are treated as one modality and the characteristic of multiple modalities for video data is typically ignored. Actually this property is very important to understand the multimedia content in video analysis and it is also a unique attribute to distinguish video data from other data patterns. Therefore, it is quite necessary to fully utilize this unique property for domain adaptive video concept detection and annotation. Furthermore, as we know that different modalities splitting from video data, such as visual, audio, and caption track, can independently manifest different aspects of video archives. A lot of scientific research demonstrates that by properly fusing the evidence from different modalities, better understanding could be achieved than only using one modality or treating all features as one modality [27]. This conclusion is also applicable in the field of domain adaptive video concept detection and annotation. Moreover, different modalities play different roles according to their generalizabilities from source domain to target domain. How to elaborately fuse them to explore their complementary nature sufficiently in process of adaptation becomes a big challenge for multi-modality transfer in video concept detection and annotation.

In this paper, we propose a novel approach, named multi-modality transfer based on multi-graph optimization (MMT-MGO), to efficiently solve this problem. The most crucial point for this exploration is to find a reasonable model to simultaneously represent multi-modalities of the video data. We choose the Optimized Multi-Graph-based Semi-supervised Learning (OMG-SSL) scheme developed by Meng Wang et al. [11] as our basic classification scheme in the target domain due to its identified ability to deal with multiple modalities, multiple distance metrics, and video temporal consistency in a unified framework. However,

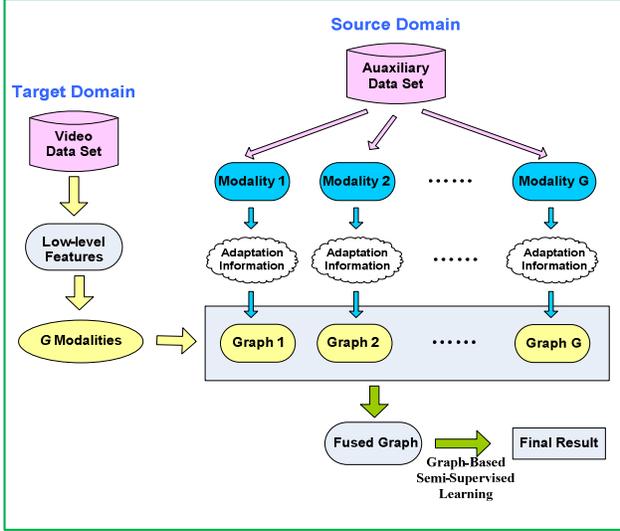


Figure 1: the proposed MMT-MGO scheme.

this framework has its own limitation, which is also well known in semi-supervised learning [22], that the unlabeled data do not always make active function in the process of learning, especially when the labeled data are insufficient. To solve this issue, we introduce the cross domain adaptation in the framework since the models generalized from the related auxiliary data in the source domains can have positive influences on the effect of unlabeled data in the target domain in semi-supervised learning process. Furthermore, in this paper we propose an incremental extension scheme for estimation when a small batch of new samples is emerging, while the OMG-SSL scheme does not possess a mechanism to deal with this situation. The optimization accuracy of new constructed multi-graphs based on the small batch of data is suspectable for graph-based semi-supervised learning. On the other hand, it is time-consuming to add the small batch of new data into previous multi-graph structure, because the graph structures are changed extensively. As the data distribution of this small batch data is similar to that of those in the nearest round multi-graph scheme, we propose to utilize the optimized parameters estimated in the nearest round optimization to estimate each sample in this small batch of new emerging data one by one for real time processing. Experimental results demonstrate that this approach can achieve comparable performance with that of incremental optimization by combining new data with those in previous multi-graph structure and the time for estimation has been reduced largely.

Figure 1 demonstrates the entire procedure of MMT-MGO. In source domain, the auxiliary data and pre-trained models are provided for the user. The adaptation information from  $G$  modalities in the source domain is transferred into multi-graph models in the target domain respectively. The auxiliary model in the source domain for one modality which generalizes to target domain better than those for other modalities will obtain a relatively larger

weight in the process of adaptation.

The organization of the rest of this paper is as follows. In section 2, we provide a short review on the related work. In section 3, we present our scheme in detail. Section 4 demonstrates the experiment results. The conclusions and future works are discussed in section 5.

## 2. Related Work

Cross Domain Adaptation or Transfer Learning is a very promising orientation in the machine learning community and attracts many attentions in recent years, which focuses on improving generalization across different distributions between source and target domains or tasks [16]. Recently a lot of domain adaptation methods have been applied in the field of multimedia content analysis [2, 13, 14, 15, 20, 23, 25]. In [2], a probabilistic transfer learning model by introducing task-level features to a hierarchical Bayesian model is proposed for information retrieval. In [23], a novel two-step (offline and online) semantic context transfer algorithm is proposed to firstly tackle domain change issue in video search. In [25], three cross domain learning strategies are proposed to cope with the statistical difference between web image data and consumer photos for textual query based retrieval.

In the field of video concept detection, domain adaptation techniques are also exploited to tackle the domain change among different video corpus to learn robust classifiers for concept detection in the target domain [6, 7, 19]. Jun Yang et al. in [6] try to learn adaptive SVM classifiers which are not “far from” the existing auxiliary classifiers and separate the labeled samples in the target domain well. Following this work, a more general formulation of adaptive SVM in [5] was proposed for function-level classifier adaptation. This framework is based on regularized loss minimization principle which measures the classification error of the target classifiers and controls the complexity of the hypothesis space. L. X. Duan [7] et al. proposed Domain Transfer SVM to simultaneously learn a kernel function and a robust SVM classifier by minimizing the both structural risk function of SVM and distribution mismatch of samples between source and target domains. In [8] data-dependent regularizer inspired from Manifold Regularization [1, 9] was used for domain adaptation, which assumes that the target classifier should have similar decision values with the pre-computed auxiliary classifiers when the source and target domain are closely relevant. The cross domain SVM developed in [19] tries to learn a new decision boundary by taking into consideration the classification impact of support vectors derived from source classifiers.

Although these approaches designed for domain adaptive video concept detection and annotation have been recognized by the public, they all ignored the very important attribute of video data, i.e. multi-modalities, in the process of adaptation, which can guides the detection

towards a right orientation.

### 3. Multi-Modality Transfer Based on Multi-Graph Optimization (MMT-MGO)

In this section, we will have detailed description on the proposed approaches. In the subsection 3.1, the multi-modality transfer scheme will be described. Sub-section 3.2 demonstrates the incremental extension scheme.

#### 3.1. Multi-Modality Transfer Scheme

First of all, we define the notations used in this proposed scheme as follows. For each sample pair  $x_i$  and  $x_j$  ( $i=1, \dots, N, j=1, \dots, N$ ,  $N$  is the number of all the data including labeled and unlabeled data), their similarity, denoted by  $W_{ij}$ , is based on a distance metric  $d(\cdot, \cdot)$  and a positive radius parameter  $\sigma$ . All the  $W_{ij}$  consist of the elements of affine matrix  $W$

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

In source domain, for each concept the auxiliary models are pre-trained for each modality. We denote them in the  $G$  modalities by  $(F^1, F^2, \dots, F^G)$ . In the target domain, there are a small set of labeled data and a large amount of unlabeled data. We denote these data as  $D^T = D_l^T \cup D_u^T$ ,

$D_l^T = \{(x_i, Y_i)\}_{i=1}^{N_l}$ ,  $D_u^T = \{x_i\}_{i=1}^{N_u}$ . The total number of labeled and unlabeled data is  $N = N_l + N_u$ . For each model

$F^g$  ( $g=1, \dots, G$ ), we use it to predict samples (including labeled and unlabeled) in the target domain, and get a predicted score vector  $f^g = [f_1^g, f_2^g, \dots, f_N^g]_{g=1}^G$ . We get the regularized framework for multi-modality transfer as follows:

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

where  $W_{g,ij}$  represents the similarity between  $x_i$  and  $x_j$  in modality  $g$ .  $D_{g,ii} = \sum_j W_{g,ij}$  and  $f_i$  represents the relevance score of  $x_i$ . The first term of the right-hand side in the formulation (2) is the smoothness constraint, which means that a good classifying function should not change too much between nearby points. The second term is the fitting constraint, which means a good classifying function should not change too much from the initial label assignment [12]. The third term means that in modality  $g$ , if the data distribution of source domain is similar to that of target domain, the target estimation score  $f = [f_1, f_2, \dots, f_N]$  will be close to  $f^g$  and for each sample  $f_i$  should be close to  $f_i^g$ . For general classification task,  $Y_i$  is set to 1 if  $x_i$  is labeled as positive, -1 if  $x_i$  is labeled as negative, and 0

if  $x_i$  is unlabeled. We can derive from formulation (2) that:

$$f = \left( \frac{1}{\mu + \theta} \sum_{g=1}^G \alpha_g L_g + I \right)^{-1} \left( \frac{\mu}{\mu + \theta} Y + \frac{\theta}{\mu + \theta} \sum_{g=1}^G \alpha_g f^g \right) \quad (3)$$

where  $L_g = D_g^{-1/2} (D_g - W_g) D_g^{-1/2}$  is the normalized graph Laplacian. The decision of  $\alpha_g$  is predefined as constants in the above framework. From Eq. (3), we can obviously draw a very important conclusion that the introduction of auxiliary information from the source domain is irrelevant with multi-graph structure in the target domain. It is not difficult to discover that when some multi-modality based auxiliary models in the source domain are provided, we only need to use these models to predict samples in the target domain, and weighted average them with initial assigned labels  $Y$ . The weights assigned to the two components are determined proportionally according to the predefined parameter  $\mu$  and  $\theta$ . When the initial label  $Y$  is reliable, i.e. the proportion of labeled data is above a given threshold, we could assign  $\mu$  with a large value to make a strong reliance on  $Y$ ; otherwise, the multi-modality based prediction score vectors  $f^g = \{f_1^g, f_2^g, \dots, f_N^g\}_{g=1}^G$  will make a relatively larger function. When regarding  $\alpha_g$  ( $g=1, \dots, G$ ) as variables and optimize them with  $f^T$ , we substitute  $\alpha_g$  with  $\alpha_g^r$  to make  $\alpha_g$  potentially close to each other as it does in [11]. The regularized formulation is as follows:

$$f^* = \arg \min_f \left\{ \sum_{g=1}^G \alpha_g^r \left( \frac{1}{2} \sum_{i,j} W_{g,ij} \left| \frac{f_i}{\sqrt{D_{g,ii}}} - \frac{f_j}{\sqrt{D_{g,jj}}} \right|^2 + \mu \sum_i |f_i - Y_i|^2 + \theta \sum_i |f_i - f_i^g|^2 \right) \right\} \\ [f, \alpha] = \arg \min_{f, \alpha} Q(f, \alpha), \text{ s.t. } \sum_{g=1}^G \alpha_g = 1 \quad (4)$$

where  $r > 1$ . To solve the problem iteratively, we can derive that:

$$\begin{cases} \frac{\partial Q(f, \alpha)}{\partial f} = 2 \sum_{g=1}^G \alpha_g^r [L_g f + \mu (f - Y) + \theta (f - f^g)] \\ \frac{\partial Q(f, \alpha)}{\partial \alpha_g} = r \alpha_g^{r-1} [f^T L_g f + \mu |f - Y|^2 + \theta |f - f^g|^2] \end{cases} \quad (5)$$

Firstly, to obtain the values of  $\alpha_g$  ( $g=1, \dots, G$ ),  $f$  is fixed, and the formulation is as follows:

$$\alpha_g = \frac{\left[ \frac{1}{f^T L_g f + \mu |f - Y|^2 + \theta |f - f^g|^2} \right]^{\frac{1}{r-1}}}{\sum_{g=1}^G \left[ \frac{1}{f^T L_g f + \mu |f - Y|^2 + \theta |f - f^g|^2} \right]^{\frac{1}{r-1}}} \quad (6)$$

when  $\alpha_g$  is obtained, we could refresh  $f$  by:

$$f = \left[ I + \frac{1}{\mu + \theta} \sum_{g=1}^G \alpha_g^r L_g \right]^{-1} \left[ \frac{\mu}{\mu + \theta} Y + \frac{\theta}{\mu + \theta} \sum_{g=1}^G \alpha_g^r f^g \right] \quad (7)$$

According to convergence theorem in [11], the convergence of iterative process is related to the matrix

$L_0 = \sum_g \alpha'_g L_g / \sum_g \alpha'_g$  and the eigenvalues of  $I - L_0 = \sum_g \alpha_g W_g$ . If  $L_0$  is symmetric and the eigenvalues of  $I - L_0$  are in  $[-1, 1]$ , the iterative process can converge. Consequently, we can conclude that the aforementioned property of matrix  $L_0$  and  $I - L_0$  does not change, because the introduction of auxiliary information from the source domain do not have any influence on the multi-graph structure in the target domain. Therefore, the iterative process in our proposed framework can still converge.

### 3.2. Incremental Extension Scheme

In OMG-SSL framework, the structures of multi-graphs are built after all the labeled and unlabeled data are known. If some new data are coming, it has to re-construct the affine matrixes and then re-optimize the parameters for a brand-new round of estimation. It is very inconvenient when the number of new emerging data is small. In this subsection, we will introduce a novel approach to solve this problem. Suppose that  $\{x_{N+1}, \dots, x_{N+M}\}$  denotes a small batch of new emerging data ( $M \ll N$ , if  $M$  is large enough and some labeled samples are provided, the cost of constructing multiple new graphs is deserved for optimization) and the data distribution of the  $M$  samples is similar to that of  $N$  previous samples in the nearest round optimization. In our incremental extension scheme, we estimate the  $M$  samples one by one. Therefore in the following discussion, we only estimate  $x_{N+1}$  as an example and the estimation of other samples can be the same. If  $f_{N+1}$  is optimized together with  $N$  previous samples, then the regularization framework is as follows:

$$f^* = \arg \min_f \left\{ \sum_{g=1}^G \alpha_g \left( \frac{1}{2} \sum_{i=1}^{N+1} \sum_{j=1}^{N+1} W'_{g,ij} \left| \frac{f_i}{\sqrt{D'_{g,ii}}} - \frac{f_j}{\sqrt{D'_{g,jj}}} \right|^2 + \mu \sum_{i=1}^{N+1} |f_i - Y_i|^2 + \theta \sum_{i=1}^{N+1} |f_i - f_i^g|^2 \right) \right\} \quad (8)$$

$$W'_{g,ij} = \begin{cases} W_{g,ij} & \text{if } i \leq N \text{ and } j \leq N \\ \exp\left(-\frac{d(x_j^g, x_{N+1}^g)}{\sigma^g}\right) & j = 1, \dots, N \text{ else} \end{cases} \quad (9)$$

$$D'_{g,ii} = \begin{cases} D_{g,ii} + W_{i,N+1} & \text{if } i \leq N \\ \sum_{j=1}^{N+1} W_{N+1,j} & \text{if } i = N + 1 \end{cases} \quad (10)$$

split  $f_{N+1}$  from other  $N$  samples we can obtain that:

$$f^* = \arg \min_f \left\{ \sum_{g=1}^G \alpha_g \left[ \sum_{i=1}^N f_i^2 - \sum_{i=1}^N \sum_{j=1}^N \frac{W'_{g,ij} f_i f_j}{\sqrt{D'_{g,ii}} \sqrt{D'_{g,jj}}} + \mu \sum_{i=1}^N |f_i - Y_i|^2 + \theta \sum_{i=1}^N |f_i - f_i^g|^2 \right] + \left( f_{N+1}^2 - 2f_{N+1} \sum_{j=1}^{N+1} \frac{W'_{g,N+1,j} f_j}{\sqrt{D'_{g,N+1,N+1}} \sqrt{D'_{g,jj}}} + \mu |f_{N+1} - Y_{N+1}|^2 + \theta |f_{N+1} - f_{N+1}^g|^2 \right) \right\} \quad (11)$$

Demanding the partial derivative of Eq. (11) on  $f_{N+1}$  to be zero, we can approximately estimate  $f_{N+1}$  by:

$$f_{N+1} = \frac{\sum_{g=1}^G \alpha_g \sum_{j=1}^N \frac{W'_{g,N+1,j} f_j}{\sqrt{D'_{g,N+1,N+1}} \sqrt{D'_{g,jj}}} + \mu Y_{N+1} + \theta \sum_{g=1}^G \alpha_g f_{N+1}^g}{1 + \mu + \theta - \sum_{g=1}^G \alpha_g \frac{W'_{g,N+1,N+1}}{D'_{g,N+1,N+1}}} \quad (12)$$

where the values  $\alpha_g (g=1, \dots, G)$  and  $f_i (i=1, \dots, N)$  can be obtained through the nearest round optimization since the data distribution to which  $f_{N+1}$  conforms is similar to that of  $N$  previous samples.  $f_{N+1}^g (g=1, \dots, G)$  denote the prediction values given by auxiliary models in source domain.

## 4. Experiments

### 4.1. Data Sets and Evaluation

To evaluate the performance of our proposed approaches, we conduct experiments by using the development set of TRECVID 2005 and 2007 video benchmark collection [18] referred to as TV05Dev and TV07Dev. The TV05Dev contains 86 hours of broadcast news videos. These news videos are segmented into 43873 shots and 61901 keyframes are extracted from these shots. The TV07Dev collection contains 60 hours of documentary videos partitioned into 18142 shots and 21532 keyframes are extracted [10]. The program structure and production value are quite different between the TV05Dev and TV07Dev [19]. Four categories of low-level features are extracted to represent the characteristics of keyframes: (1) ColorMoments based on  $5 \times 5$  grid (225D); (2) Edge Direction Histogram (73D); (3) Gabor Texture (48D); (4) Audio (36D, including short-time energy, shot-time average zero-crossing rate, sub-brand short-time energy, sub-brand shot time energy rate, MFCC). These four features are regarded as four modalities. All the keyframes are manually labeled for 36 semantic concepts chosen from LSCOM-lite [26]. The features (1)(2)(3) and corresponding pre-trained auxiliary classifiers from source domains are provided by Columbia 374 [24], while the audio features and the auxiliary audio detectors are produced by S. Tang et al. [17]. For OMG-SSL and MMT-MGO,  $L_1$  metric is used to calculate similarities between sample pairs in the affine matrix  $W$ . We make matrices  $L_g$  sparse by only keeping  $n$  largest values each row as it does in [11], which can significantly reduce the computational cost while keeping comparable performance. The non-interpolated Average Precision is used as performance evaluation, which is the standard performance metric in TRECVID.

### 4.2. Multi-Modality Transfer Scheme

In this subsection, we compare our proposed MMT-MGO with the OMG-SSL which utilizes multi-modalities for annotation without domain adaptation and other three domain adaptive concept detection methods without multi-modality transfer, that is, A-SVM (Adaptive SVM) [6], CD-SVM [19] and DT-SVM [7]. The comparisons of these

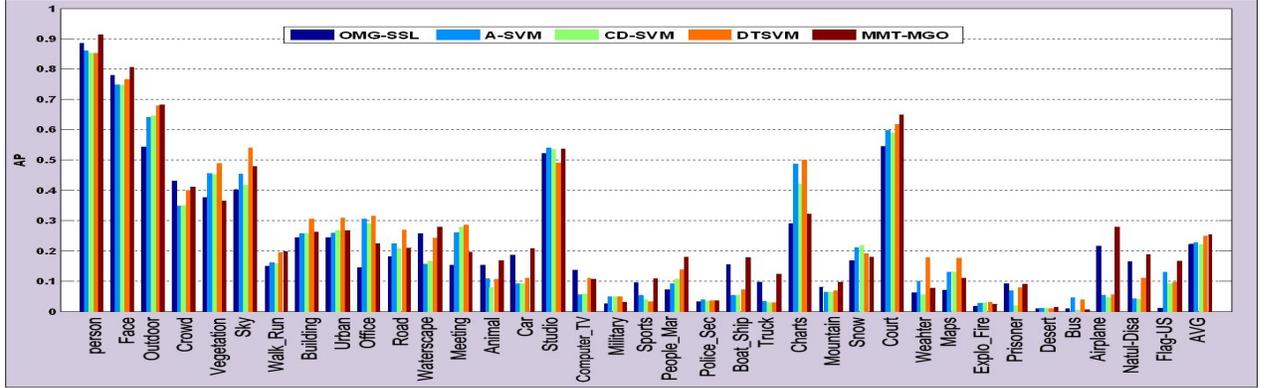


Figure 2. Performance comparison on DT05Dev-to DT07Dev data set in individual APs for 36 concepts

methods are conducted on the two data sets with TV05Dev as the source domain and TV07Dev as the target domain. The performance results of the three adaptation methods without multi-modality transfer are replicated from [7]. To compare MMT-MGO with them more intuitively, the experimental settings for data of MMT-MGO are the same with that in [7], that is, the auxiliary data set  $D^A$  is obtained by randomly sampling 100 positive samples per concept from the TRECVID 2005 dataset, and 10 positive samples per concept randomly sampled from TRECVID2007 are used as the labeled samples in the target domain  $D_t^l$ . The remaining data in target domain are used as test data. The MAPs of these different approaches are demonstrated in Table 1 and the APs for individual concepts are in Figure 2.

Table 1. Performance Comparison on MAPs

OMG-SSL	22.2%
A-SVM	22.8%
CD-SVM	21.9%
DT-SVM	24.9%
MMT-MGO	25.4%

### 4.3. Incremental Extension Scheme

When some new data are emerging, we firstly evaluate the scores of these data  $f_{IE}$  through our proposed incremental extension. Then we incorporate these new data with labeled and unlabeled data in the target domain for the nearest round optimization, and use the MMT-MGO to obtain the evaluation scores  $f_{MMT-MGO}$  through a brand-new round optimization. We compare the accuracy of  $f_{IE}$  with that of  $f_{MMT-MGO}$  to evaluate the performance of our incremental extension. For a small batch set  $A$ , we calculate:

$$Error(A) = \frac{1}{N_A} \sum_{i \in A} |f_i^{MMT-MGO} - f_i^{IE}| \quad (13)$$

since the Eq. (13) can reflect the difference between  $f_{IE}$  and  $f_{MMT-MGO}$  well. The new emerging data are randomly selected from the test set of TRECVID 2007. The performance of incremental intension scheme is illustrated

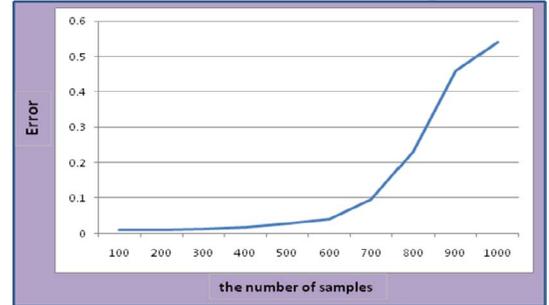


Figure 3: The performance of Incremental Extension Scheme for different scales of new emerging data.

in Figure 3. We can see that when the scale of samples is small, the errors are limited in a small range and increase slowly. The approximation by incremental extension is acceptable. When the size of sample exceeds some bound, the errors increase sharply and can't be tolerant.

It is easy to derived from Eq. (12) that the computational cost for estimating one new data is  $O(G \times N)$  and thus for  $M$  new emerging data is  $O(M \times G \times N)$ . However, when combining the  $M$  new emerging data with previous data in target domain for re-optimization, the computational cost becomes  $O(G \times (M+N)^2)$ . The reduction of computational cost is obvious.

## 5. Conclusion and Future Work

We have proposed an adaptive video concept annotation method by introducing multi-modality transfer based on multi-graph optimization. From the experiments presented in section 4, we have the following observations: First, the proposed MMT-MGO approach outperform some existing recognized adaptive video concept detection and annotation approaches which do NOT consider the multi-modality scheme for video data in the process of domain adaptation. Second, compared with its origin OMG-SSL, the MMT- MGO achieves better performance which validates the effectiveness of introducing domain adaptation in graph- based semi-supervised learning. Third, the incremental extension scheme can acquire comparable accuracy with precise optimization for a small batch of data

while much less computational time is required. Many aspects in our proposed framework can be improved in our future work. First, we can consider some more reasonable distance metrics in the graphs to reflect the pattern similarities [3] between sample pairs, or semantic relations by using sparse graphs [4] to derive datum-adaptive local structures which can robustly handle heterogeneously distributed data and remove concept-unrelated links. Second, due to increasingly growing up of video on the internet and the intensive application of web videos [21, 28, 29], we should exploit the domain adaptation techniques considering some special characteristics of web videos. Especially, the noisy tags and comments associated with web videos can be utilized as one kind of very important modality. We plan to explore these aspects in future work.

## 7. Acknowledgement

This work was supported by National Basic Research Program of China (973 Program, 2007CB311100); National Nature Science Foundation of China (60873165).

## References

- [1] A. B. Goldberg, M. Li and X. J. Zhu. Online Manifold Regularization: A New Learning Setting and Empirical Study. *Machine Learning and Knowledge Discovery in Databases*, 2008.
- [2] A. Quattoni, M. Collins and T. Darrell. Transfer Learning for Image Classification with Sparse Prototype Representations. *IEEE CVPR*, 2008.
- [3] H. X. Wang and J. Pei. Clustering by Pattern Similarity. *Journal of Computer Science and Technology*, 2008.
- [4] J. Tang, S. Yan, R. Hong, G. Qi and T. Chua. Inferring Semantic Concepts from Community-Contributed Images and Noisy Tags. *ACM MM*, 2009.
- [5] J. Yang, and A. G. Hauptmann. A Framework for Classification Adaptation and its Application in Concept Detection. *ACM MIR*, 2008.
- [6] J. Yang, R. Yan, and A. G. Hauptmann. Cross-Domain Video Concept Detection Using Adaptive SVMs. *ACM MM*, 2007.
- [7] L. X. Duan, I. W. Tsang, and D. Xu. Domain Transfer SVM for Video Concept Detection. *IEEE CVPR*, 2009.
- [8] L. X. Duan, I. W. Tsang, D. Xu and T. S. Chua. Domain Adaptation from Multiple Sources via Auxiliary Classifiers. *ICML*, 2009.
- [9] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 2006.
- [10] S. Tang, J.-T. Li, M. Li, and et al. *Trecvid 2008 participation by mcg-ict-cas*. In *TRECVID Workshop*, 2008.
- [11] M. Wang, X. S. Hua, X. Yuan, Y. Song, and L. R. Dai. Optimizing Multi-Graph Learning: Towards a Unified Video Annotation Scheme. *ACM MM*, 2007.
- [12] D. Y. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with Local and Global Consistency. *NIPS*, 2004.
- [13] Q. An, C. Wang, I. Shterev, E. Wang, L. Carin, and D. B. Dunson. Hierarchical Kernel Stick-Breaking Process for Multi-Task Image Analysis. *ICML*, 2008.
- [14] Q. Yang, Y. Chen, G. R. Xue, W. Dai, and Y. Yu. Heterogeneous Transfer Learning for Image Clustering via the Social Web. *The 47<sup>th</sup> Annual Meeting of the ACL and the 4<sup>th</sup> IJCNLP of the AFNLP*, 2009.
- [15] R. Yan, and J. Zhang. Transfer Learning Using Task-Level Features with Application to Information Retrieval. In *21<sup>th</sup> IJCAI*, 2009.
- [16] S. J. Pan and Q. Yang. A Survey on Transfer Learning. 2009
- [17] S. Tang, Y. D. Zhang, J. T. Li, and et al. Trecvid 2007 High Level Feature Extraction by mcg-ict-cas. In *TRECVID Workshop*, 2007.
- [18] TRECVID: <http://www-nlpir.nist.gov/Project/trecvid/>.
- [19] W. Jiang, E. Zavesky, S. F. Chuang and A. Loui. Cross-Domain Learning Methods for High Level Visual Concept Classification. *IEEE ICIP*, 2008.
- [20] X. Wang, C. Zhang, and Z. Zhang. Boosted Multi-Task Learning for Face Verification with Application to Web Image and Video Search. *IEEE CVPR*, 2009.
- [21] X. Zhang, Y. C. Song, J. Cao, Y. D. Zhang and J. T. Li. Large Scale Incremental Web Video Categorization, *ACM MM WSMC* 2009.
- [22] X. Zhu. Semi-supervised Learning Literature Survey. 2008.
- [23] Y. G. Jiang, C. W. Ngo, S. F. Chang. Semantic Context Transfer across Heterogeneous Sources for Domain Adaptive Video Search. *ACM MM*, 2009.
- [24] A. Yanagawa, S. F. Chang, L. Kennedy, and W. Hsu. Columbia University's Baseline detectors for 374 LSCOM Semantic Visual Concepts. *Columbia University ADVENT Technical Report*, 2007
- [25] Y. Liu, D. Xu, I. W. Tsang, and J. Luo. Using Large-Scale Web Data to Facilitate Textual Query Based Retrieval of Consumer Photos. *ACM MM*, 2009.
- [26] M. R. Naphade, L. Kennedy, J.R. Kender, S.F. Chang, J. Smith, P. Over, and A. Hauptmann. A light scale concept ontology for multimedia understanding for TRECVID 2005. The IBM Research Technical Report, 2005.
- [27] H. Tong, J. R. He, M. J. Li, C. S. Zhang, and W. Y. Ma. Graph-based multi-modality learning. *ACM MM*, 2005.
- [28] Richang Hong, Jinhui Tang, and Hung-Khoon Tan. Event Driven Summarization for Web Videos. *ACM MM*, 2009.
- [29] Richang Hong, Guangda Li, Liqiang Nie, Jinhui Tang, Tat-Seng Chua. Exploring Large Scales Data for Multimedia QA: An Initial Study. *ACM CIVR*, 2010.