

# Semi-Supervised Multi-Instance Multi-Label Learning for Video Annotation Task\*

Xin-Shun Xu<sup>1,2</sup>, Yuan Jiang<sup>1</sup>, Xiangyang Xue<sup>3</sup>, Zhi-Hua Zhou<sup>1</sup>  
{xuxs,jiangy,zhouzh}@lamda.nju.edu.cn xyxue@fudan.edu.cn

<sup>1</sup>National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210046, China

<sup>2</sup>School of Computer Science and Technology, Shandong University, Jinan 250101, China

<sup>3</sup>School of Computer Science, Fudan University, Shanghai 200433, China

## ABSTRACT

Traditional approaches for automatic video annotation usually represent one video clip with a flat feature vector, neglecting the fact that video data contain natural structures. It is also noteworthy that a video clip is often relevant to multiple concepts. Indeed, the video annotation task is inherently a Multi-Instance Multi-Label learning (MIML) problem. Considering that manually annotating videos is labor-intensive and time-consuming, this paper proposes a semi-supervised MIML approach, SSMIML, which is able to exploit abundant unannotated videos to help improve the annotation performance. This approach takes label correlations into account, and enforces similar instances to share similar multi-labels. Evaluation on TREVID 2005 show that the proposed approach outperforms several state-of-the-art methods.

## Categories and Subject Descriptors

I.2.10 [Artificial Intelligence]: Vision and Scene Understanding—*video analysis*; I.2.6 [Artificial Intelligence]: Learning

## General Terms

Algorithms, Design, Experimentation

## Keywords

Video Annotation, Multi-Instance Multi-Label Learning, Semi-Supervised Learning

## 1. INTRODUCTION

With the rapid development of storage devices, networks and compression techniques, large collections of digital videos have been generated and shared on the Internet. In order to effectively access large-scale video data, a common theme is to first annotate video clips with semantic concepts, and then use these concepts to index or retrieve the video. Annotating large archive of video data manually is quite labor-intensive and time-consuming, and thus,

\*Supported by the NSFC (60975043, 60970047, 61173068), 973 Program (2010CB327903), Jiangsu PPRF (1001001B) and CPSF (20100470063).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'12, October 29–November 2, 2012, Nara, Japan.

Copyright 2012 ACM 978-1-4503-1089-5/12/10 ...\$10.00.

automatic video annotation (also called concept detection or high-level feature extraction) has emerged as an active topic in the multimedia community.

A general routine for video annotation is to first extract textual or visual features of sample data, and then train one or multiple models to make predictions on the concepts for new video data. Many approaches have been developed, to name a few: Fan et al. [4] proposed a Bayesian framework for semantic video classification and indexing; Xie et al. [23] used Hidden Markov Models and dynamic programming to detect specific semantics in soccer videos; Yanagawa et al. [26] built a set of baseline detectors for 374 LSCOM concepts using SVMs; Tang et al. [20] proposed a graph-based learning approach with structure-sensitive manifold ranking.

Most previous studies represent a video clip with a flat feature vector, neglecting the fact that video data have natural structure and there are significant relations inside the structure. By using a single flat feature vector to represent a video clip, important information becomes collapsed and missed. This issue has been recently noticed. For example, Naphade et al. [13] proposed a generalized multi-instance learning algorithm for video concept detection, where one video clip is represented by a set of feature vectors (or called instances); Gu et al. [6, 7] proposed several models to capture the structural information in video data. On the other hand, video data are usually with very rich semantics, and one video clip is usually related to multiple concepts. For example, in TRECVID 2005 data set, according to LSCOM-Lite annotations [12], there are 39 concepts and one frame is related to up to 12 concepts. The concepts associated with one video clip are naturally related, and their correlations will help improve annotation performance; however, such information will be missed if the concepts are dealt with separately. Qi et al. [16] recently tried to exploit the correlations among concepts with a multi-label learning approach, and showed that such information are really helpful. It is also worth mentioning that it is expensive to get a large amount of annotated video data because manual annotating is labor-intensive and time-consuming; thus, semi-supervised learning has been applied to exploit unannotated video data to help improve the annotation performance [21, 25].

Notice that the three above issues, i.e., natural structure, rich semantics and small annotated samples, were separately dealt with in previous studies although they usually occur simultaneously. In this paper, we propose the SSMIML (Semi-Supervised Multi-Instance Multi-Label learning) approach which deals with these issues together. As its name suggests, SSMIML is an approach developed in MIML (Multi-Instance Multi-Label learning), a new machine learning framework [30, 31] which is suitable for learning with data objects with complicated semantics. In the MIML framework, each data object is represented by a set of instances and is allowed to

have multiple labels simultaneously. During the past few years, the MIML framework has been found well useful in diverse applications such as image annotation [14, 28, 30], text categorization [31, 27], bioimage informatics [10], video annotation [24], etc. MIML learning theory has been studied [22], and many MIML learning approaches have been developed [14, 27, 28, 30, 31]; however, semi-supervised MIML approaches are almost untouched. Recently, Feng et al. [5] proposed a MIML approach that can exploit unlabeled data, but it works in transductive setting and requires all unlabeled data be available for training. In contrast, our SSMIML approach works in inductive setting, and is able to make predictions for unseen data. Experiments on TREVID 2005 data set show that our SSMIML outperforms several state-of-the-art methods.

The rest of this paper is organized as follows: In Sections 2 we present the SSMIML formulation; in Section 3 we propose an iterative algorithm; in Section 4 we report on experiments; finally in Section 5, we conclude the paper.

## 2. THE FORMULATION

A direct solution to semi-supervised multi-instance multi-label learning is to degenerate the MIML problem into a series of multi-instance problems or multi-label problems in the way of [30], and then apply semi-supervised multi-instance learning [8, 17] or multi-label learning [11, 29] methods. The degeneration process, as indicated by [31], will definitely loss information. Therefore, we design the SSMIML approach.

Let  $\mathcal{X}$  be the instance space and  $\mathcal{Y}$  the label space. Then, given a data set  $\{\{X_i, Y_i\}_{i=1}^n, \{X_j\}_{j=n+1}^{n+m}\}$ ,  $n$  is the number of labeled bags,  $m$  is the number of unlabeled bags,  $X_i \subseteq \mathcal{X}$  is a set of instances  $\{x_{i1}, x_{i2}, \dots, x_{in_i}\}$ ,  $x_{ij} \in \mathcal{X}$  ( $j = 1, 2, \dots, n_i$ ),  $n_i$  the number of instances in bag  $X_i$ ,  $Y_i \subseteq \mathcal{Y}$  is a set of labels  $\{y_{i1}, y_{i2}, \dots, y_{il_i}\}$ ,  $y_{ik} \in \mathcal{Y}$  ( $k = 1, 2, \dots, l_i$ ),  $l_i$  is the number of labels in  $Y_i$ . The task is to find a function  $f : 2^{\mathcal{X}} \rightarrow 2^{\mathcal{Y}}$ . Suppose that the cardinality of  $\mathcal{Y}$  is  $T$ ,  $f$  can be decomposed as  $f = (f_1, f_2, \dots, f_T)$ , where  $f_t$  is the function for label  $t$ . Then, the labels of bag  $X$  can be predicted as  $Y = (\text{sign}(f_1(X)), \dots, \text{sign}(f_T(X)))$ . Inspired by manifold regularization [2], we try to get:

$$f^* = \arg \min_f \frac{1}{nT} \sum_{i=1}^n V(X_i, Y_i, f) + \frac{\gamma_A}{T} \Omega(f) + \frac{\gamma_I}{(n+m)^2 T^2} I(f). \quad (1)$$

where  $V(X_i, Y_i, f)$  is the loss function,  $\Omega(f)$  is a regularization term to control the model complexity,  $I(f)$  is a manifold regularization term to force similar samples to share similar labels, and  $\gamma_A$  and  $\gamma_I$  are parameters to balance the contribution of these two terms.

Let  $L_t^+$  denote the number of positive bags for label  $t$ . By sorting all positive bags before negative bags for each label, we implement the first term of Eq. 1 as

$$\begin{aligned} & \sum_{i=1}^n V(X_i, Y_i, f) \\ &= \sum_{t=1}^T \left( \sum_{i=1}^{L_t^+} \left( 1 - \max_{j:x_{ij} \in X_i} f_t(x_{ij}) \right)^2 + \sum_{i=L_t^++1}^n \sum_{j:x_{ij} \in X_i} (f_t(x_{ij}) + 1)^2 \right). \end{aligned} \quad (2)$$

where the second term means that all instances in a negative bag should be negative, and the first one defines the penalty term for a positive bag to be only related to the instance with the largest value, which also implicitly requires that:

$$f_t(X_i) = \max_{j:x_{ij} \in X_i} f_t(x_{ij}). \quad (3)$$

where  $x_{ij}$  is the  $j$ th instance of bag  $i$ , and  $f_t(x_{ij})$  is a model working on instance level. For simplicity, we assume that  $f_t(x_{ij})$  is a

linear model, i.e.,  $f_t(x_{ij}) = \langle w_t, \phi_t(x_{ij}) \rangle$ , where  $\phi_t(\cdot)$  is the mapping corresponding to a kernel  $\mathcal{K}_t$ .

The model complexity is typically defined as:

$$\Omega(f) = \frac{1}{T} \sum_{t=1}^T \|w_t\|^2. \quad (4)$$

The regularization term which forces similar samples to share similar labels can be defined as:

$$I(f) = \sum_{i,j=1}^N (f(x_i) - f(x_j))' (f(x_i) - f(x_j)) W_{ij}. \quad (5)$$

where  $'$  denotes matrix transpose,  $N$  is the total number of instances of all bags, and  $W$  is the similarity matrix for instances. Note that the labels are treated independently in Eq. 5. To exploit label correlation, according to [9], we introduce a matrix  $C^{-1} = P'P$  to model the correlations between labels. Thus,  $I(f)$  becomes:

$$\begin{aligned} I(f) &= \sum_{i,j=1}^N (Pf(x_i) - Pf(x_j))' (Pf(x_i) - Pf(x_j)) W_{ij} \\ &= \sum_{i,j=1}^N (f(x_i) - f(x_j))' C^{-1} (f(x_i) - f(x_j)) W_{ij} \\ &= \langle F' L F, C^{-1} \rangle. \end{aligned} \quad (6)$$

where  $F$  is an  $N \times T$  matrix,  $F_{it} = f_t(x_i)$  (note that, for simplicity of representation, we neglect the bag number; thus,  $x_i$  is an instance, not a bag), and  $L = D - W$  is the Laplacian matrix of  $W$ ,  $D = \text{diag}(\sum_{j=1}^N W_{ij})$ .  $\langle X, Y \rangle = \text{trace}(X'Y)$ . Thus, by introducing  $C^{-1}$ , we distinguish the outputs after projecting them into a new label space in which the correlations are considered.

Substituting Eqs. 2, 4 and 6 into Eq. 1, we get:

$$\begin{aligned} f^* &= \arg \min_f \frac{\gamma_A}{T^2} \sum_{t=1}^T \|w_t\|^2 + \frac{\gamma_I}{(n+m)^2 T^2} \langle F' L F, C^{-1} \rangle \\ &+ \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^{L_t^+} \left( 1 - \max_{j:x_{ij} \in X_i} f_t(x_{ij}) \right)^2 \\ &+ \frac{1}{nT} \sum_{t=1}^T \sum_{i=L_t^++1}^n \sum_{j:x_{ij} \in X_i} (f_t(x_{ij}) + 1)^2. \end{aligned} \quad (7)$$

Note that, for any positive semi-definite matrix  $C$ , it can be shown that the Representer Theorem for manifold regularization [2] still holds based on the works [3, 18].

## 3. AN ITERATIVE SOLUTION

We can rewrite Eq. 7 by introducing relaxation parameters  $\xi_{it}$ :

$$\begin{aligned} f^* &= \arg \min_f \frac{\gamma_A}{T^2} \sum_{t=1}^T \|w_t\|^2 + \frac{\gamma_I}{(n+m)^2 T^2} \langle F' L F, C^{-1} \rangle \\ &+ \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^{L_t^+} \xi_{it}^2 + \frac{1}{nT} \sum_{t=1}^T \sum_{i=L_t^++1}^n \sum_{j:x_{ij} \in X_i} (f_t(x_{ij}) + 1)^2 \\ \text{s.t.} \quad & 1 - \max_{j:x_{ij} \in X_i} f_t(x_{ij}) \leq \xi_{it} \quad t = 1, \dots, T; i = 1, \dots, L_t^+ \\ & \xi_{it} \geq 0 \quad t = 1, \dots, T; i = 1, \dots, L_t^+. \end{aligned} \quad (8)$$

It is not hard to show that the optimal solutions to Eqs. 7 and 8 satisfy  $|f_t(x_{ij})| \leq 1$ . Moreover, these two problems are equivalent if  $|f_t(x_{ij})| \leq 1$ . So, the optimal solution to Eq. 8 is also the optimal one to Eq. 7. Note that  $C$  can be calculated directly when there are a good number of training examples. Since in our problem setting there are only a small amount of labeled data,  $C$  calculated from training data is unreliable. Thus, besides  $f$ , we have to learn  $C$  in Eq. 8. Here, we present an iterative solution.

### 3.1 Estimating $f$ with Fixed $C$

From Eq. 8 we can see that the objective function is convex. The first constraints are non-convex, but each is the difference between 1 and a convex function  $\max(\cdot)$ . Hence, if  $C$  is fixed, inspired by [31], we can use the Constrained Concave Convex Procedure (CCCP) [19] to find the suboptimal solution to this problem. The CCCP was proposed for problems with a concave-convex objective function with concave-convex constraints, and it is theoretically guaranteed to converge. It works in an iterative way: At each iteration, the CCCP replaces the non-convex functions with their first-order Taylor expansions, and approximates the problem as a convex optimization problem.

Because  $\max(\cdot)$  function is non-smooth, the gradient in the first-order Taylor expansion needs to be replaced by the subgradient. Thus, at the  $k$ th iteration, for the  $\max(\cdot)$  function related to bag  $X_i$ , the first-order Taylor expansion is approximated as:

$$[\max_{j:x_{ij} \in X_i} f_i(x_{ij})]_{f_i}^{(k)} \approx \max_{j:x_{ij} \in X_i} f_i^{(k)}(x_{ij}) + \beta_{ii}'(f_i - f_i^{(k)}). \quad (9)$$

where  $f_i^{(k)}$  is the starting point at the  $k$ th iteration, which can be written as  $f_i^{(k)} = (f_i^{(k)}(x_1), \dots, f_i^{(k)}(x_N))$ ,  $x_i$  is an instance and  $N$  is the number of instances as defined in Eq. 6. The subgradient  $\beta_{ii}$  is an  $N \times 1$  vector, and its  $j$ th element is

$$\beta_{ij} = \begin{cases} 1/n_{ia} & x_{ij} \in X_i \text{ and } f_i^{(k)}(x_{ij}) = \max_{j:x_{ij} \in X_i} f_i^{(k)}(x_{ij}) \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

where  $\max_{j:x_{ij} \in X_i} f_i^{(k)}(x_{ij})$  is the largest label value of  $X_i$ ,  $n_{ia}$  is the number of active  $x_{ij}$ 's that have the largest label value. Note that, other choices for  $\beta_{ij}$  can also be used.

When getting  $f$ , we can also have the values of  $\xi_{ii}'$ 's.

### 3.2 Estimating $C$ with Fixed $f$

Once we get  $f$  and  $\xi$ , we can find that except the second one, all other items are known to us. Thus, the optimization problem of Eq. 8 becomes:

$$\min_{C \in \mathcal{A}} \langle F'LF, C^{-1} \rangle. \quad (11)$$

$C$  is a positive semi-definite matrix by definition. Without loss of generality, according to [9], we further suppose that the trace of  $C$  is 1, i.e.,  $C \in \mathcal{A} = \{C | C \geq 0, \text{trace}(C) = 1\}$ . Based on the work of [1, 9], this problem has a closed-form solution, which can be obtained by:

$$C = \frac{F'LF}{\text{trace}(F'LF)}. \quad (12)$$

## 4. EXPERIMENTS

In this section, we evaluate SSMIML on the development (DEV) set of TRECVID 2005. This original data set consists of 80 hours international broadcast news in English, Arabic and Chinese, and contains 43907 shots. Some shots have been segmented further into subshots. The final DEV set contains 61,901 subshots [15]. For each subshot, a keyframe has been extracted to represent the subshot, and it has been associated with one or more concepts in 39 concepts according to the LSCOM-Lite annotations [12]. Instead of segmenting each keyframe into regions and then treating each region as an instance, in our experiments, we treat each shot as a bag and each keyframe of a subshot as an instance. Totally, there are 43,907 bags; on average, there are 1.41 instances in a bag, and the maximum number of instances in a bag is 13. 82.5% bags have more than one concepts. Similar to the data partition routine in [26, 29], the data set is separated into three partitions. To be

specific, 90 videos are selected as training set, 16 videos as validation set, and the remaining 31 videos are used as test set. For each keyframe, SIFT descriptors are extracted by Harris-Laplace and dense sampling methods. Then, codebook is generated from randomly selected descriptors. Finally, each frame is represented as a histogram. The performance is evaluated by *Average Precision* (AP) and *Mean Average Precision* (MAP), the official performance metric in TRECVID evaluations.

We compare SSMIML with several state-of-the-art methods, including a semi-supervised multi-label learning approach ML-GRF [29], semi-supervised multi-instance learning approach MISSL [17], and a supervised MIML approach MIMLSVM [30]. For ML-GRF, we aggregate all instances in a bag into one feature vector; for MISSL, we run it for multiple times each for one label; for MIML-SVM, we simply neglect unlabeled data. We use Gaussian kernel to calculate the weight matrix  $W$ , and select all parameters of all algorithms on validation set. Note that the weight matrix  $W$  is very large ( $43,907 \times 43,907$  or  $61,901 \times 61,901$ ) and difficult for storage and inversion calculation. Thus, we use K-Nearest-Neighbor (K-NN) to find the  $K$  ( $K = 300$ ) nearest neighbors for each instance or bag, and construct a sparse representation of  $W$ .

The comparison results are shown in Figure 1. It can be observed that the semi-supervised approaches ML-GRF, MISSL and SSMIML beat MIMLSVM on 35 concepts, verifying the usefulness of exploitation of unannotated video clips. MIMLSVM works well on concepts such as ‘‘People’’ and ‘‘Studio’’, and we find that these concepts have a good number of labeled training examples. Among the semi-supervised approaches, SSMIML and ML-GRF beat MISSL on 24 concepts, such as ‘‘Watscape\_Waterfront’’, ‘‘Airplane’’, ‘‘Car’’, ‘‘Face’’, ‘‘Outdoor’’, etc. We attribute this observation to the fact that SSMIML and ML-GRF considers label correlation but MISSL treats the labels independently, whereas these concepts are with strong correlations. More importantly, it can be found that SSMIML beats ML-GRF on 30 concepts, validating the usefulness of MIML framework for semi-supervised video annotation.

## 5. CONCLUSION

In this paper, we propose SSMIML, a semi-supervised multi-instance multi-label learning approach for video annotation. This approach works in the framework of multi-instance multi-label learning (MIML), which explicitly considers the facts that video data usually contain natural structures and a video clip is often related to multiple concepts. SSMIML is able to exploit unannotated videos to help improve the annotation performance, and it can make predictions on unlabeled video clips that do not appear in training data owing to its inductive nature. Experiments on TRECVID 2005 data show that SSMIML outperforms several state-of-the-art methods. An interesting future work is to incorporate temporal information into the approach.

## 6. REFERENCES

- [1] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS 19*, pages 41–48. 2007.
- [2] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434, 2006.
- [3] T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *J. Mach. Learn. Res.*, 6:615–637, 2005.

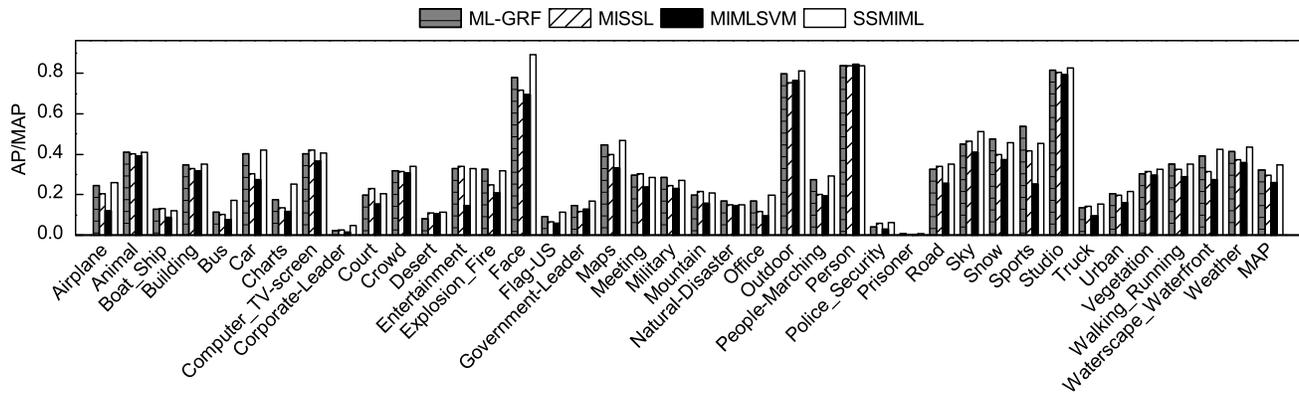


Figure 1: Comparison of SSMIML with several state-of-the-art approaches

- [4] J. Fan, H. Luo, and X. Lin. Semantic video classification by integrating flexible mixture model with adaptive em algorithm. In *MIR*, pages 9–16, 2003.
- [5] S. Feng and D. Xu. Transductive multi-instance multi-label learning algorithm with application to automatic image annotation. *Expert Syst. Appl.*, 37(1):661–670, 2010.
- [6] Z. Gu, T. Mei, X.-S. Hua, J. Tang, and X. Wu. Multi-layer multi-instance kernel for video concept detection. In *MM*, pages 349–352, 2007.
- [7] Z. Gu, T. Mei, X.-S. Hua, J. Tang, and X. Wu. Multi-layer multi-instance learning for video concept detection. *IEEE Trans. Multimedia*, 10(8):1605–1616, 2008.
- [8] Y. Jia and C. Zhang. Instance-level semisupervised multiple instance learning. In *AAAI*, pages 640–645, 2008.
- [9] Y.-F. Li, S.-J. Huang, and Z.-H. Zhou. Regularized semi-supervised multi-label learning. *J. Comp. Res. Dev.*, 49(6):1272–1278, 2012 (in Chinese).
- [10] Y.-X. Li, S. Ji, J. Ye, S. Kumar, and Z.-H. Zhou. Drosophila gene expression pattern annotation through multi-instance multi-label learning. *ACM/IEEE Trans. Comput. Biol. Bioinf.*, 9(1):98–112, 2012.
- [11] Y. Liu, R. Jin, and L. Yang. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *AAAI*, pages 421–426, 2006.
- [12] M. Naphade, L. Kennedy, J. R. Kender, S. F. Chang, P. Over, and A. Hauptmann. A light scale concept ontology for multimedia understanding for TRECVID 2005. Technical report, IBM Research, 2005.
- [13] M. Naphade and J. Smith. A generalized multiple instance learning algorithm for large scale modeling of multimedia semantics. In *ICASSP*, pages 341–344, 2005.
- [14] N. Nguyen. A new svm approach to multi-instance multi-label learning. In *ICDM*, pages 384–392, 2010.
- [15] C. Petersohn. Fraunhofer HHI at TRECVID 2004: Shot boundary detection system. In *TRECVID*, 2004.
- [16] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, M. Wang, and H.-J. Zhang. Correlative multilabel video annotation with temporal kernels. *ACM Trans. Multimedia Comput., Comm. Appl.*, 5(1):1–27, 2008.
- [17] R. Rahmani and S. A. Goldman. MISSL: Multiple-instance semi-supervised learning. In *ICML*, pages 705–712, 2006.
- [18] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.
- [19] A. J. Smola, S. Vishwanathan, and T. Hofmann. Kernel methods for missing variables. In *AISTATS*, pages 325–332, 2005.
- [20] J. Tang, X.-S. Hua, M. Wang, T. Mei, G.-J. Qi, and X. Wu. Structure-sensitive manifold ranking for video concept detection. In *MM*, pages 852–861, 2007.
- [21] 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. Vis. Image Und.*, 113(3):384–396, 2009.
- [22] W. Wang and Z.-H. Zhou. Learnability of multi-instance multi-label learning. *Chinese Sci. Bull.*, 57(19):2488–2491, 2012.
- [23] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with domain knowledge and hidden markov models. *Pattern Recogn. Lett.*, 25(7):767–775, 2004.
- [24] X.-S. Xu, X. Xue, and Z.-H. Zhou. Ensemble multi-instance multi-label learning approach for video annotation task. In *MM*, pages 1153–1156, 2011.
- [25] R. Yan and M. Naphade. Semi-supervised cross feature learning for semantic concept detection in videos. In *CVPR*, pages 657–663, 2005.
- [26] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia university’s baseline detectors for 374 LSCOM semantic visual concepts. Technical Report # 222-2006-8, Columbia University ADVENT, Mar. 2007.
- [27] S.-H. Yang, H. Zha, and B.-G. Hu. Dirichlet-bernoulli alignment: A generative model for multi-class multi-label multi-instance corpora. In *NIPS 23*, pages 2143–2150. 2009.
- [28] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. In *CVPR*, pages 1–8, 2008.
- [29] Z.-J. Zha, T. Mei, J. Wang, Z. Wang, and X.-S. Hua. Graph-based semi-supervised learning with multiple labels. *J. Vis. Comm. Image Rep.*, 20(2):97–103, 2009.
- [30] Z.-H. Zhou and M.-L. Zhang. Multi-instance multi-label learning with application to scene classification. In *NIPS 19*, pages 1609–1616. 2007.
- [31] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li. Multi-instance multi-label learning. *Artif. Intell.*, 176(1):2291–2320, 2012.