Multi-Feature Fusion via Hierarchical Regression for Multimedia Analysis

Yi Yang[1], Jingkuan Song[2], Zi Huang[2], Zhigang Ma[3], Nicu Sebe[3], Alexander G. Hauptmann[1].

[1]School of Computer Science, Carnegie Mellon University.

[2]School of Information Technology and Electrical Engineering, The University of Queensland.

[3]Department of Information Engineering and Computer Science, University of Trento.

***Abstract*—Multimedia data are usually represented by multiple features. In this paper, we propose a new algorithm, namely Multi-feature Learning via Hierarchical Regression for multimedia semantics understanding, where two issues are considered. First, labeling large amount of training data is labor intensive. It is meaningful to effectively leverage unlabeled data to facilitate multimedia semantics understanding. Second, given that multimedia data can be represented by multiple features, it is advantageous to develop an algorithm which combines evidence obtained from different features to infer reliable multimedia semantic concept classifiers. We design a hierarchical regression model to exploit the information derived from each type of feature, which is then collaboratively fused to obtain a multimedia semantic concept classifier. Both label information and data distribution of different features representing multimedia data are considered. The algorithm can be applied to a wide range of multimedia applications and experiments are conducted on video data for video concept annotation and action recognition. Using Trecvid and CareMedia video datasets, the experimental results show that it is beneficial to combine multiple features. The performance of the proposed algorithm is remarkable when only a small amount of labeled training data are available.**

***Index Terms*—Multiple feature fusion, semi-supervised learning, video concept annotation, action recognition.**

I. Introduction

Multimedia content is usually represented by multiple features. For example, given a video frame, its visual content can be represented by different features such as color histogram, SIFT, etc. It therefore turns an interesting research challenge to effectively utilize the multiple information sources of independent or heterogeneous features. Intuitively, analyzing different features simultaneously is beneficial for disambiguation [14]. Previous research efforts have also shown that better performance could be achieved for multimedia content analysis if we properly fuse the evidences from different features when compared to only using one type of feature or simply using all types of feature as one feature [21], [27], [30], [31], [33], [34], etc.

Late fusion and early fusion are two straightforward ways of dealing with multi-feature data [26]. However, it remains unclear which fusion is more reliable [26]. It has been shown that feature concatenation is less effective in multimedia content analysis, especially when the features are independent or heterogeneous [34]. In the field of machine learning, researchers have developed many multi-view learning algorithms to address this problem. Representative works include Canonical Correlation Analysis (CCA) [11], [28], two-view support vector machines, i.e., SVM-2k [6] and their variants [13], [18]. These algorithms have been applied to different applications, resulting in better performance than feature concatenation in a broad range of applications, such as cross-language text analysis, object recognition, image annotation, image-audio clustering, and so on. However, these algorithms require a large amount of labeled data for training, which is often expensive and seldom available.

Multimedia semantics understanding is to associate multimedia data with a single or multiple semantic concepts. For example, video concept annotation associates videos with labels/concepts to provide effective and efficient tools for managing video resources [10]. There are many ways to improve the performance of multimedia semantics understanding. One well-known method is to define more accurate features for multimedia representation, such as a visual thesaurus [24]. Another typical and effective approach is to apply machine learning algorithms. Generally speaking, multimedia semantics understanding related task can be usually regarded as a classification problem. Many supervised classification algorithms can be used for multimedia semantics understanding, such as multimedia event detection [23]. However, a typical supervised classification algorithm may require a large amount of labeled data and collecting this is time consuming and labor intensive. For example, 111 researchers from 23 institutes spent 220+ hours to annotate only 63 hours of Trecvid 2003 development corpus [19].

There are three main strategies to relieve the tedious work in labeling a large amount of training data for multimedia content analysis. The first strategy is known as active learning [17], [37], which selects the most informative data as the training data to be labeled. The second one is transfer learning, which utilizes the labeled data from another domain, e.g., in [23], Ma et al. employ annotated video frames to facilitate multimedia event detection of video clips. The third one is semi-supervised learning [22], [3], [36], [39], which leverages unlabeled data to infer a more accurate classifier. Previous studies have shown that simultaneously utilizing labeled and unlabeled data is beneficial for multimedia semantics understanding.

Motivated by the recent success of semi-supervised learning, in this paper we address the problem of effectively exploring the information contained in multiple features of both labeled and unlabeled data for multimedia content analysis and propose a new semi-supervised multi-feature learning algorithm, namely Multi-feature Learning via Hierarchical Regression (MLHR). Different from most of the existing semi-supervised algorithms [3], [25], [35], [38], [39], the manifold structure of each feature type is preserved during the training phase. MLHR is a general algorithm, which can be applied to a variety

 Yi Yang and Alexander G. Hauptmann are with the School of Computer Science, Carnegie Mellon University. Jingkuan Song, and Zi Huang re with School of Information Technology and Electrical Engineering, The University of Queensland. Zhigang Ma and Nicu Sebe are with Department of Information Engineering and Computer Science, University of Trento. This material is based upon work supported in part by the National Science Foundation under Grants No.IIS-0812465, No.IIS-0917072, in part by the National Institutes of Health (NIH) Grant No.1RC1MH090021-01, in part by the European Commission under the contract FP7-248984 GLOCAL.

of applications related to multimedia content analysis, where the multimedia data are represented by multiple features. In this paper, we apply the proposed MLHR algorithm to video concept annotation and action recognition to test its effectiveness. Compared with the existing algorithms, MLHR has the following two main advantages. First, MLHR leverages unlabeled data represented by multiple features to improve the performance in multimedia semantic understanding. The manifold structure of each feature type is preserved, resulting in a more faithful learning result. Second, in order to exploit the manifold structure of the training data, we propose a statistical approach to better exploit the manifold structure of the training data, which is more robust than simply using the pairwise distances of the training data.

The rest of this paper is organized as follows. In Section 2, we briefly discuss related work. The proposed algorithm is detailed in Section 3, followed by experiment. Conclusion is drawn in Section 5.

II. Related Work

A. Semi-supervised Learning for Multimedia Understanding

Machine learning and data mining have been shown to be effective in bridging the semantic gap [16]. A typical example is Support Vector Machine (SVM) and its variants, which have been widely used for multimedia semantic understanding, e.g., video and image annotation [8], [24], [32]. Recently, semi-supervised learning and its applications in multimedia have attracted much research attention [5], [3], [25], [35], [38], [39].

In the rest of this paper, $\|\cdot\|_F$ denotes the Frobenius Norm. Suppose there are n training data $\{x_1, ..., x_t, x_{t+1}, ..., x_n\}$ from c classes, in which the first t ($t < n$) data are labeled samples. Denote $Y = [Y_1, ..., Y_n]^T \in \{0,1\}^{n\times c}$ as the label information provided by human supervisors. Given a labeled datum x_i, if it belongs to the j-th class, $Y_{ij} = 1$, otherwise $Y_{ij} = 0$. If x_i is not a labeled datum, $Y_{ij} = 0$ for any j that $1 \le j \le c$. Let $F = [F_1, ..., F_n]^T \in \mathbb{R}^{n\times c}$, where $F_i \in \mathbb{R}^c$ is the predicted label vector of x_i. A larger value of F_{ij} indicates a higher possibility that x_i is associated with the j-th class. The affinity matrix $A \in \mathbb{R}^{n\times n}$ is defined as follows:

$$A_{ij} = \begin{cases} \exp\frac{-\|x_i - x_j\|^2}{\sigma}, & x_i \text{ and } x_j \text{ are } k\text{-nearest neighbors;} \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where σ is a parameter. In [38], a graph based classification algorithms, namely learning with Local and Global Consistency (LGC), was proposed, whose objective function is shown in (2).

$$\min_F \sum_{i,j=1}^{n} A_{ij} \left\| \frac{F_i}{\sqrt{D_{ii}}} - \frac{F_j}{\sqrt{D_{jj}}} \right\|_F^2 + \mu \sum\nolimits_{i=1}^{n} \|F_i - Y_i\|_F^2, \quad (2)$$

where μ is a parameter and D is a diagonal matrix with its diagonal element $D_{ii} = \sum_j A_{ij}$.

In the field of multimedia, the graph based algorithm LGC and its variants have been applied to different applications, resulting in remarkable performance. In [29], Wang et al. have demonstrated that a limitation of LGC is that the affinity matrix A in (1) is defined completely according to the distance between data. Besides, as a transductive classification algorithm, LGC is not able to predict the labels of the data which are outside the training set. Each time new data are added into the database, one needs to re-run the whole algorithm.

Other semi-supervised learning algorithms, such as Manifold Regularized Least Square Regression [3] and Flexible Manifold Embedding [25], are able to predict labels which are outside the training set. The objective function of Manifold Regularized Least Square Regression [3] is as follows.

$$\min_W \lambda_1 \sum\nolimits_{i,j=1}^{n} \left(A_{ij} \left\| W^T \tilde{x}_i - W^T \tilde{x}_j \right\|_F^2 \right) + \lambda_2 \|W\|_F^2 + \left\| \tilde{X}_{Train}^T W - Y_t \right\|_F^2, \quad (3)$$

where $\tilde{x}_i$ is the i-th datum after subtracting the mean of all the training data, $\tilde{X}_{Train} = [\tilde{x}_1, ..., \tilde{x}_t]$, and λ_1 and λ_2 are regularization parameters. Given a multimedia datum to be annotated, its label can be directly obtained by the classifier W. However, it is designed to deal with a single feature. A frequently used method to exploit multiple feature data is feature concatenation but the structural information of each type of feature is lost. Besides, the performance is sensitive to the parameter σ [35].

B. Multi-Feature Learning for Multimedia Understanding

Suppose a multimedia object is represented by v features $x_1^i, x_2^i, ..., x_v^i$. A common way of dealing with the multiple features is to concatenate each feature vector and represent the multimedia object as $[x_1^i, x_2^i, ..., x_v^i]$. It has been shown that such representation may degrade the performance of multimedia content analysis algorithms, especially when the features are independent or heterogenous [26], [34]. A possible reason could be that the individual structural information of each feature may be lost in feature concatenation.

In [26], Snoek et al. have classified the feature fusion strategies into two groups, which are early fusion and late fusion. They have shown that if SVM classifier is used, late fusion tends to give better performance in video analysis [26]. However, more computation burden is required in late fusion. In [31], researchers have analyzed the limitation of linear combination. In [7], a multi-kernel learning algorithm is proposed for object classification. The major limitation is that it requires the computation of multiple kernel matrices. In [34], cartoon image features are discussed and classified into different feature classes. A bi-distance metric learning algorithm was then proposed to learn a better distance metric from heterogeneous features for cartoon image retrieval [34]. However, the algorithm proposed in [34] is application dependent and cannot be directly applied to other applications. In [27], a multiple feature hashing algorithms is proposed for near duplicate video retrieval.

Canonical Correlation Analysis (CCA) is a statistical approach which maximizes the correlations between two modalities in a compact subspace [11]. Vinokourov et al. [28] have shown that CCA gains good performance in cross-language text analysis. In [13], CCA and Linear Discriminant Analysis are combined for face and object recognition. SVM-2K [6] is another well-known algorithm of multi-feature learning. This family of algorithms has also been applied to different applications [6], [18].

C. Notes of Caution

Despite the success of semi-supervised multi-feature learning for multimedia analysis, we should also emphasize the following notes of caution:

- While semi-supervised learning has shown great potential for multimedia content analysis, in some cases, utilizing unlabeled data for training can degrade performance, especially when the manifold assumption does not hold. It remains unclear how to automatically decide if and when it is advisable to exploit unlabeled data for training.
- Combining multiple features is generally beneficial for multimedia analysis. The combination of a "weak" feature and a "strong" feature usually yields better performance than using one feature only, if the features are complementary. Yet, feature fusion

may hurt performance if the multiple features are contradictory or a feature is very weak. Evaluating the appropriateness of combining multiple features has not been sufficiently studied.

III. The proposed algorithm

In this section, we give the details of the proposed algorithm. We begin with the terms and notations. t is the number of labeled training data and n is the number of the training data. In semi-supervised learning, $t \ll n$, that is, only a small amount of training data are labeled. Suppose each datum is represented by v features. Given an integer $g \leq v$, we denote x_g^i as the g-th feature of the i-th datum and $X_g = [x_g^1, ..., x_g^n] \in \mathbb{R}^{d_g \times n}$, where d_g is the dimension of the g-th feature. I is the identity matrix. $1_m \in \mathbb{R}^m$ is a vector of all ones for an arbitrary number m. $Tr(\cdot)$ is the trace operator. Following [25], [38], we define $F = [F_1, ..., F_n]^T \in \mathbb{R}^{n \times c}$ as the predicted matrix of training data, which is the same as (2). The definition of Y is the same as in (2) as well. Further, we define $f_g = [f_{g1}, ..., f_{gn}]^T \in \mathbb{R}^{n \times c}$ as the predicted matrix of training data derived from the g-th feature. As indicated by previous graph based approaches [3], [25], [35], [38], the manifold structure of input data plays an important role in pattern recognition and multimedia analysis. To exploit the manifold structure, we construct v local sets for a datum, each of which is computed according to one feature type. More specifically, given the i-th datum represented by its g-th feature x_g^i, we construct a local set, denoted as $\mathcal{N}_g^i$, which consists of x_g^i and its k-nearest neighbors according to the distance derived from the g-th feature.

Instead of computing the affinity matrix directly, we propose a statistical approach to exploit the manifold structure of the input data for semi-supervised learning. Inspired by [36], we assume there is a local classifier c_g^i, which classifies all the training data in $\mathcal{N}_g^i$ to c classes. The prediction error *w.r.t.* c_g^i is be formulated as

$$\sum_{x_g^j \in \mathcal{N}_g^i} \ell oss(c_g^i(x_g^j), f_{gj}), \tag{4}$$

where $\ell oss(\cdot)$ is a loss function. To exploit the structural information derived from the g-th feature, we propose to minimize the total prediction error *w.r.t.* the g-th feature [36], i.e.,

$$\min_{f_g, c_g^i} \sum_{i=1}^{n} \sum_{x_g^j \in \mathcal{N}_g^i} \Big(\ell oss(c_g^i(x_g^j), f_{gj}) + \lambda \Omega(c_g^i) \Big), \tag{5}$$

where $\Omega(c_g^i)$ is a regularization function on c_g^i and λ is a parameter. We minimize (6) to combine the evidences from all the v features.

$$\min_{f_g, c_g^i} \sum_{g=1}^{v} \sum_{i=1}^{n} \sum_{x_g^j \in \mathcal{N}_g^i} \Big(\ell oss(c_g^i(x_g^j), f_{gj}) + \lambda \Omega(c_g^i) \Big). \tag{6}$$

Different from concatenating the v features, the individual structural information is preserved. The predicted label matrix F of the training data should be consistent with each evidence f_g $(1 \leq g \leq v)$. We then minimize $\min_{F, f_g} \sum_{g=1}^{v} \|F - f_g\|_F^2$. To classify the data outside the training set, we train a set of global classifiers $\{C_1, ..., C_v\}$, in which C_g $(1 \leq g \leq v)$ is able to predict the labels of the data according to the g-th feature. We propose to simultaneously learn the predicted label matrix F of the training data and the v global classifiers $C_g|_{g=1}^v$. Specifically, we minimize the following objective.

$$\begin{aligned} \min_{F, f_g, c_g^i, C_g} & \sum_{g=1}^{v} \sum_{i=1}^{n} \sum_{x_g^j \in \mathcal{N}_g^i} \Big(\ell oss(c_g^i(x_g^j), f_{gj}) + \lambda \Omega(c_g^i) \Big) \\ & + \mu_1 \sum_{g=1}^{v} \sum_{i=1}^{n} \Big(\ell oss(C_g(x_g^i), f_{gi}) + \gamma \Omega(C_g) \Big) \\ & + \mu_2 \sum_{g=1}^{v} \|F - f_g\|_F^2, \\ s.t. \quad & F_i = Y_i, \text{ if } x_i \text{ is a labeled training data,} \end{aligned} \tag{7}$$

where μ_1, μ_2, and γ are parameters. Denote U as a diagonal matrix. If x_i is a labeled datum $U_{ii} = \infty$, and $U_{ii} = 0$ otherwise. Then we arrive at

$$\begin{aligned} \min_{F, f_g, c_g^i, C_g} & \sum_{g=1}^{v} \sum_{i=1}^{n} \sum_{x_g^j \in \mathcal{N}_g^i} \Big(\ell oss(c_g^i(x_g^j), f_{gj}) + \lambda \Omega(c_g^i) \Big) \\ & + \mu_1 \sum_{g=1}^{v} \sum_{i=1}^{n} \Big(\ell oss(C_g(x_g^i), f_{gi}) + \gamma \Omega(C_g) \Big) \\ & + \mu_2 \sum_{g=1}^{v} \|F - f_g\|_F^2 + Tr\left((F-Y)^T U (F-Y) \right). \end{aligned} \tag{8}$$

We use the least square loss in our model as the loss function. The objective function of MLHR is shown as follows

$$\begin{aligned} \min_{\substack{F, f_g, w_g^i, \\ W_g, b_g^i, B_g}} & \sum_{g=1}^{v} \sum_{i=1}^{n} \sum_{x_g^j \in \mathcal{N}_g^i} \left(\left\| (w_g^i)^T x_g^j + b_g^i - f_{gj} \right\|_F^2 + \lambda \left\| w_g^i \right\|_F^2 \right) \\ & + \mu_1 \sum_{g=1}^{v} \sum_{i=1}^{n} \left(\left\| (W_g)^T x_g^i + B_g - f_{gi} \right\|_F^2 + \gamma \|W_g\|_F^2 \right) \\ & + \mu_2 \sum_{g=1}^{v} \|F - f_g\|_F^2 + Tr\left((F-Y)^T U (F-Y) \right), \end{aligned}$$

where $w_g^i \in \mathbb{R}^{d_g \times c}$ and $b_g^i \in \mathbb{R}^c$ are the local classifier and bias term of x_g^i *w.r.t.* the g-th feature, and $W_g \in \mathbb{R}^{d_g \times c}$ and $B_g \in \mathbb{R}^c$ are the global classifier and bias term *w.r.t.* the g-th feature.

Let $\mathcal{N}_g^i = \{x_g^i, x_g^{i_1}, ..., .x_g^{i_k}\}$ where $x_g^{i_1}, ..., .x_g^{i_k}$ are the k-nearest neighbors of x_g^i according to the g-th feature. $X_g^i = [x_g^i, x_g^{i_1}, ..., .x_g^{i_k}] \in \mathbb{R}^{d_g \times (k+1)}$. The objective function of MLHR can be rewritten as:

$$\begin{aligned} \min_{\substack{F, f_g, w_g^i, \\ W_g, b_g^i, B_g}} & \sum_{g=1}^{v} \sum_{i=1}^{n} \left(\left\| (X_g^i)^T w_g^i + 1_{k+1} b_g^i - f_g^i \right\|_F^2 + \lambda \left\| w_g^i \right\|_F^2 \right) \\ & + \mu_1 \sum_{g=1}^{v} \left(\left\| (X_g)^T W_g + 1_n B_g - f_g \right\|_F^2 + \gamma \|W_g\|_F^2 \right) \\ & + \mu_2 \sum_{g=1}^{v} \|F - f_g\|_F^2 + Tr\left((F-Y)^T U (F-Y) \right), \end{aligned} \tag{9}$$

where $f_g^i = [f_{gi}, f_{gi_1}, ...f_{gi_k}]^T$ is the predicted local label matrix of the data in $\mathcal{N}_g^i$ according to the g-th feature.

By setting the derivative of (9) *w.r.t.* w_g^i and b_g^i to be zero, we have

$$b_g^i = \frac{1}{k+1} \left((f_g^i)^T 1_{k+1} - (w_g^i)^T X_g^i 1_{k+1} \right), \tag{10}$$

$$w_g^i = (X_g^i H_{k+1} (X_g^i)^T + \lambda I)^{-1} X_g^i H_{k+1} f_g^i, \tag{11}$$

where $H_{k+1} = I - \frac{1}{k+1}1_{k+1}1_{k+1}^T$ is the local centering matrix. Similarly, by setting the derivative of (9) *w.r.t.* W_g and B_g to be zero, we have

$$B_g = \frac{1}{n}\left(f_g^T 1_n - W_g^T X_g 1_n\right), \tag{12}$$

$$W_g = (X_g H_n X_g^T + \gamma I)^{-1} X_g H_n f_g, \tag{13}$$

Substituting b_g^i, w_g^i, B_g and W_g in (9) by (10), (11), (12) and (13) respectively, we arrive at

$$\begin{aligned}\min_{F,f_g} \sum_{g=1}^{v}\sum_{i=1}^{n} Tr\left((f_g^i)^T L_g^i f_g^i\right) + \mu_1 \sum_{g=1}^{v} Tr(f_g^T A_g f_g) \\ +\mu_2 \sum_{g=1}^{v} \|F - f_g\|_F^2 + Tr\left((F-Y)^T U (F-Y)\right),\end{aligned} \tag{14}$$

where

$$L_g^i = H_{k+1} - H_{k+1}(X_g^i)^T (X_g^i H_{k+1} (X_g^i)^T + \lambda I)^{-1} X_g^i H_{k+1} \tag{15}$$

and

$$A_g = H_n - H_n X_g^T (X_g H_n X_g^T + \gamma I)^{-1} X_g H_n. \tag{16}$$

For the ease of representation, we define the selection matrix $S_g^p \in \mathbb{R}^{n\times(k+1)}$ as follows.

$$(S_g^p)_{ij} = \begin{cases} 1 & \text{if } x_g^i \text{ is the } j\text{-th element in } \mathcal{N}_g^p; \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

Recall that all the elements in $\mathcal{N}_g^i = \{x_g^i, x_g^{i_1}, ..., x_g^{i_k}\}$ are selected from the whole training set according to the distance derived from the g-th feature. Therefore, we have

$$f_g^i = (S_g^i)^T f_g. \tag{18}$$

Then we have

$$\begin{aligned}\sum_{i=1}^{n} Tr\left((f_g^i)^T L_g^i f_g^i\right) &= \sum_{i=1}^{n} Tr\left(f_g^T S_g^i L_g^i (S_g^i)^T f_g\right) \\ &= Tr\left(f_g^T \sum_{i=1}^{n}\left(S_g^i L_g^i (S_g^i)^T\right) f_g\right)\end{aligned}$$

Denote $L_g = \sum_{i=1}^{n}\left(S_g^i L_g^i (S_g^i)^T\right)$. Then (14) can be written as

$$\begin{aligned}\min_{F,f_g} \sum_{g=1}^{v} Tr\left(f_g^T L_g f_g\right) + \mu_1 \sum_{g=1}^{v} Tr(f_g^T A_g f_g) + \mu_2 \sum_{g=1}^{v} \|F - f_g\|_F^2 \\ +Tr\left((F-Y)^T U (F-Y)\right),\end{aligned}$$

which is equivalent to the following

$$\begin{aligned}\min_{F,f_g} \sum_{g=1}^{v} Tr\left(f_g^T (L_g + \mu_1 A_g) f_g\right) + \mu_2 \sum_{g=1}^{v} \|F - f_g\|_F^2 \\ +Tr\left((F-Y)^T U (F-Y)\right).\end{aligned} \tag{19}$$

Let us define

$$Q_g = \begin{bmatrix} L_g^1 & & 0 \\ & \cdots & \\ 0 & & L_g^n \end{bmatrix} \tag{20}$$

and

$$S_g = [S_g^1, ..., S_g^n]. \tag{21}$$

Then we have

$$L_g = S_g Q_g S_g^T. \tag{22}$$

Note that

$$\begin{aligned}L_g^i &= H_{k+1} - H_{k+1}(X_g^i)^T (X_g^i H_{k+1} (X_g^i)^T + \lambda I)^{-1} X_g^i H_{k+1} \\ &= H_{k+1}((X_g^i)^T X_g^i + \lambda I)^{-1} H_{k+1}.\end{aligned}$$

It is easy to prove that the objective function shown in (19) is convex. By setting the derivative of (19) *w.r.t* f_g to be zero, we have

$$\begin{aligned}&2\left(L_g + \mu_1 A_g\right) f_g - 2\mu_2 (F - f_g) = 0 \\ &\Rightarrow f_g = \mu_2 (L_g + \mu_1 A_g + \mu_2 I)^{-1} F\end{aligned} \tag{23}$$

By setting the derivative of (19) *w.r.t* F to be zero, we have

$$\sum_{g=1}^{v} \mu_2 (F - f_g) + U(F - Y) = 0 \tag{24}$$

Substituting f_g in (24) by (23), we have

$$\sum_{g=1}^{v} \mu_2 (F - \mu_2 (L_g + \mu_1 A_g + \mu_2 I)^{-1} F) + U(F - Y) = 0$$

$$\Rightarrow F = \left(v\mu_2 I + U - \mu_2^2 \sum_{g=1}^{v} (L_g + \mu_1 A_g + \mu_2 I)^{-1}\right)^{-1} UY. \tag{25}$$

In this way, we have obtained the optimal solution of F, f_g, W_g and B_g of the proposed MLHR algorithm. The detailed approach of MLHR is summarized in Algorithm 1 as follows.

Algorithm 1: The MLHR algorithm.

```
for g = 1 to v do
    for i = 1 to n do
        Compute L_g^i according to (15);
    Compute L_g according to (22);
    Compute A_g according to (16);
Compute F according to (25);
for g = 1 to v do
    Compute f_g according to (23);
    Compute B_g according to (12);
    Compute W_g according to (13);
```

For a testing datum represented by v features $x_1^t, ..., x_v^t$, its predicted label vector F_t can be computed by

$$F_t = \sum\nolimits_{g=1}^{v}\left(W_g^T x_g^t + B_g\right)/v. \tag{26}$$

Next, we briefly discuss the difference between MLHR and some other semi-supervised learning algorithms. In recent years, several transductive classification algorithms have been proposed in [35], [38] and applied to different applications for multimedia content analysis. Compared with other algorithms, the main advantage of MLHR is that it is able to deal with the data which are outside the training set, without rerunning the training processing. Considering that a large amount of multimedia data are generated in every single day, MLHR is more suitable for real world applications.

Apart from the aforementioned algorithms, there are some other graph-based semi-supervised learning algorithms, which are able to predict the labels of the data outside the training set, e.g., Manifold Regularization (MR) [3] and Flexible Manifold Embedding (FME) [25]. MLHR mainly differs from MR and FME in two aspects. First, although the multimedia data are represented by multiple independent

features in many cases, both MR and FME simply concatenate them to produce a high dimensional vector as input. MLHR is more capable of exploiting multiple features because the structural information of individual feature is preserved and considered. Second, both MR and FME compute the affinity matrix according to (1) directly. The limitation of this type of approach is that the affinity matrix A is completely based on feature similarities [29]. Besides, the performance is usually sensitive to the parameter σ in (1). Differently, our algorithm employs a group of local classifiers to exploit the manifold structure and we advocate that a statistical approach is more capable to exploit manifold structure than directly computing the pairwise distances [35].

IV. Experiments

In this section, we test the proposed framework in terms of video concept annotation and action recognition.

A. Video concept annotation

In this subsection, we test the performance of the proposed algorithm in video concept annotation. The Trecvid 2005 video corpus consisting of 160 hours news video is used in our experiment [1]. It contains 61,901 key-frames, which are from 137 news videos recorded from 13 different programs in English, Arabic and Chinese, which are segmented into 49,532 shots and 61,901 sub-shorts. For each sub-shorts, we extract one key-frame. We annotate 36 concepts which have more than 100 key-frames associated to them.

We compare our algorithm with two representative multiple feature learning algorithms, SVM-2K [6] and CCA [11] (followed by SVM and Least Square regression, which are denoted as CCA-SVM and CCA-LS, respectively). To show the advantage of MLHR over the existing semi-supervised learning algorithms, we report the results from Manifold Regularized Least Square Regression (MRLS) [3]. Besides, we compare MLHR with the multi-label classification algorithm Shared-subspace Learning for Multilabel Classification (SLMC) [12]. Three types of visual features are extracted and then normalized to represent the key-frames. The first feature is 225-D block-wise LAB-based Color Moments (CM) extracted over 55 fixed grid partitions. The second feature is 500-D bag of visual words based on Scale Invariant Feature Transform (SIFT) descriptors [20]. We also use 144-D color correlogram in Hue Saturation Value (HSV) color space to represent the videos. SVM-2K and CCA are two-view learning algorithms which are designed to deal with only two types of features. Therefore, while more features can be used in our algorithm, we only use two features in our algorithm to compare the different algorithms. Specifically, we report the results of using CM and SIFT as well as the results of using HSV and SIFT. We concatenate different features as the input of MRLS and SLMC. We also report the results from MLHR when only one type of feature is used. We denote the results of using only one feature type as CM, HSV, and SIFT respectively.

To show the effectiveness of semi-supervised learning, we label a small amount of key-frames for training while most of the training data are unlabeled. Specifically, we have sampled 10,000 key-frames from the training set indicated by [32] as the training data. Among the 10,000 training data, we have labeled 100 key-frames for each concept, which is much less than in [32] where over 40,000 key-frames are labeled for training. As reported in [35], the performance is not sensitive to the local regularization parameter λ. We did not tune this parameter and fix it as 1. For the other parameters in MLHR, including μ_1, μ_2 and γ, we tune them from $\{10^{-6}, 10^{-3}, 10^{0}, 10^{3}, 10^{6}\}$ and report the best result. For SVM-2K, CCA, MRLS, HSV and CM, we tune all the parameters from the same range. All of the unlabeled key-frames are used as the testing data. Evaluation metric is Average Precision (AP).

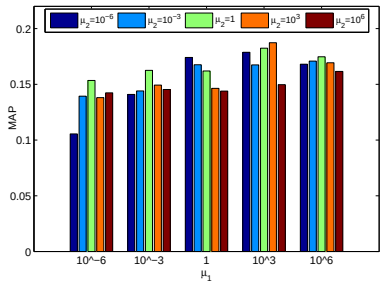
1) The parameter sensitivity study: First, we test the performance variation of the MLHR algorithm *w.r.t* the three parameters γ, μ_1 and μ_2 when CM and SIFT features are used to represent the videos. We average the APs over all of the 36 concepts to compute the Mean Average Precision (MAP). In this experiment, we fix one of the three parameters and report the MAP while the other two parameters are changing. The results of using CM and SIFT features are shown in Figure 1. MLHR gains the best performance when $\gamma = 1$, $\mu_1 = 10^3$ and $\mu_2 = 10^3$ for this dataset. From Figure 1(b) and Figure 1(c), we can see that MLHR is comparatively less sensitive to the parameter γ when it is smaller than 1. Generally speaking, in order to obtain better performance, μ_2 should not be smaller than 1. This implies that the mismatch between F and $f_g|_{g=1}^{v}$ (i.e., $\sum_{i=1}^{v} \|F - f_g\|_F^2$) incurs heavier penalty than other terms. As for the regularization parameter μ_1, we observe from Figure 1(c) that if μ_2 is fixed, the performance of MLHR is not very sensitive. Yet, we would emphasize that the optimal parameters for MLHR are data dependent.

2) Performance comparison of different algorithms: Next, we compare the MLHR algorithm proposed in this paper with the other algorithms. The results of using CM and SIFT features are shown in Table I. The results of using HSV and SIFT features are shown in Table II. Because CM outperforms HSV, we additionally report the APs of each concept in Figure 2 when the videos are represented by CM and SIFT. From Table I, Table II and Figure 2, we have the following observations.

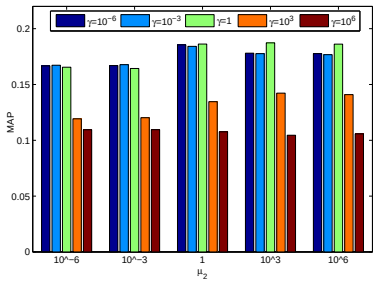
- First, we observe from Table I that MLHR gains the highest MAP over the 36 concepts. More specifically, the MAP of MLHR is 0.1872, which outperforms MRLS by about 10% comparatively. Table II has similar results.
- The performance of MLHR is more stable and it always gains good performance for different concepts. In summary, MLHR gains the best performance or the second best performance for 33 out of the 36 concepts.
- The semi-supervised algorithms (MRLS and MLHR) outperform the supervised ones (SLMC, CCA-SVM, CCA-LS, SVM-2K), indicating that it is beneficial to utilize unlabeled data for multimedia semantics understanding for this dataset, especially when the number of labeled data is not large.
- The accuracy could be limited when a single feature type is used. When we utilize multiple features, the performance is improved, even if one of the features is a weak one. For example, as a single feature, CM outperforms SIFT but the combination of the two gains much better performance than using CM only. The experiment results shown in Figure 2 validate that the MLHR algorithm proposed in this paper is capable to utilize multiple feature types for multimedia analysis.
- Both MRLS and MLHR exploit two features of the labeled and unlabeled data for video concept annotation. Yet, MLHR proposed in this paper outperforms MRLS significantly mainly due to the following two reasons. First, different from MRLS, MLHR does not directly compute the pairwise distance to exploit the data distribution. Second, the individual structural information of each feature type is preserved in MLHR. In contrast, MRLS simply concatenates the two features, and thus the individual structural information of a single feature may be lost.

B. Action Recognition

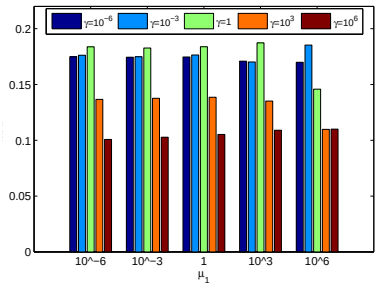
In this subsection, we test the proposed algorithm in terms of action recognition using the CareMedia dataset collected by Carnegie Mellon University. The CareMedia dataset was collected to provide useful statistics to help doctors' diagnosis and patients' heath status



(a) The performance variation *w.r.t.* μ_1 and μ_2 when γ is fixed as 1.



(b) The performance variation *w.r.t.* μ_2 and γ when μ_1 is fixed as 10^3.



(c) The performance variation *w.r.t.* μ_1 and γ when μ_2 is fixed as 10^3.

Fig. 1. Performance variation of different parameter settings for video concept annotation using CM and SIFT features.

TABLE I
MAP OF THE 36 CONCEPTS OF VIDEO CONCEPT ANNOTATION USING CM AND SIFT FEATURES.

	SLMC	CCA-SVM	CCA-LS	SVM-2K	MRLS	CM	SIFT	MLHR
MAP	0.1531	0.1589	0.1361	0.1298	0.1701	0.1528	0.1238	0.1872

TABLE II
MAP OF THE 36 CONCEPTS OF VIDEO CONCEPT ANNOTATION USING HSV AND SIFT FEATURES.

	SLMC	CCA-SVM	CCA-LS	SVM-2K	MRLS	HSV	SIFT	MLHR
MAP	0.1430	0.1462	0.1225	0.1211	0.1627	0.1001	0.1238	0.1803

assessment. 15 geriatric patients' activities in public spaces were recorded in a nursing home by fielding an array of video cameras with patient, caregiver, and institution approval. Twenty three cameras and microphones were mounted in fixed locations, designed to be as unobtrusive as possible.

We defined 19 different human actions that caregivers would be interested in. These can be categorized into two types. The first type is concerned with patients' movement activities and the second type is more concerned with patients' detailed behaviors. We test the performance by annotating the following 6 concepts of the first type: *Walking Through*, *Walking To Standing Point*, *Standing Up*, *Sitting Down*, *Object Placed On Table* and *Object Removed From Table*. The first feature used in this experiment is 1000 dimension Bag-of-Words STIP feature [15]. Besides, we also propose to incorporate optical flow into feature design and use MoSIFT feature [4] as a second feature to represent videos. Similarly, 1000 dimension Bag-of-Words MoSIFT feature is used.

In this experiment, we use a subset recorded by a particular camera in the dining room. After removing the clips which are too short to be captured by the feature extracting code share by [15], there are 1796 video sequences left. We use 1000 data as training set and the other 796 data as testing set. 10 positive samples per concept are labeled as positive examples for training. The same as video concept annotation, we compare our algorithm with SLMC, CCA-SVM, CCA-LS, SVM-2K, MRLS. We also report the results from single feature MoSIFT and STIP using our algorithm. For the parameters of different algorithms, we tune them from $\{10^{-6}, 10^{-3}, 10^{0}, 10^{3}, 10^{6}\}$ and report the best results. Again, AP is used as the evaluation metric.

Table III and Figure 3 show the experiment results. We can see that our algorithm outperforms all of the competitors. When compared with the two well-known multi-view learning algorithms CCA (followed by SVM) and SVM-2K, the relative performance gains are 14.8% and 10.0%, respectively. Compared with the non-linear model, the linear model has many advantages, such as simplicity and efficiency. However, CCA-SVM greatly outperforms CCA-LS, indicating that the non-linear model is more capable of dealing with the BoW feature for action recognition. Yet, our algorithm, which utilizes a linear model for recognition, gains the best performance, which further validates the advantage of our method. Note that it is easy to extend our algorithm to a kernel method and we omit the discussion on this due to the space limit. We observe that when using MLHR with only one type of feature, MoSIFT greatly outperforms STIP. If we combine the two features, the performance can be further improved. These observations indicate that 1.) it is beneficial to incorporate optical flow for action recognition; 2.) multiple feature fusion helps in action recognition even though one of the two features is a weak one.

C. Discussions

Using video concept annotation and action recognition as examples, we have compared our algorithm to different algorithms. Our experiments show that multiple feature fusion usually yields better performance than a single feature. As a single feature, the MAPs of CM and SIFT are 0.1528 and 0.1238, respectively. However, if we combine the two, the MAP increases to 0.1872. Employing the complementary information for multimedia analysis tends to lead to an optimal solution in the global view. If we analyze feature fusion under a regularization framework, the two different features regularize each other, by which over-fitting can be reduced. Our MLHR algorithm and the second best algorithm MRLS are both semi-supervised learning algorithms. This indicates that simultaneously utilizing both the labeled and the unlabeled data for multimedia analysis does help. Compared to the second best algorithm MRLS, the improvement of our algorithm is 10.05% and 15.53% in video concept annotation and action recognition, respectively. Generally speaking, there are two main reasons why the performance of our algorithm is further improved. First, in our algorithm the structural information of each type of feature is preserved and considered during the training. Second, our algorithm employs a group of local regression models to exploit the manifold structure of the data. The

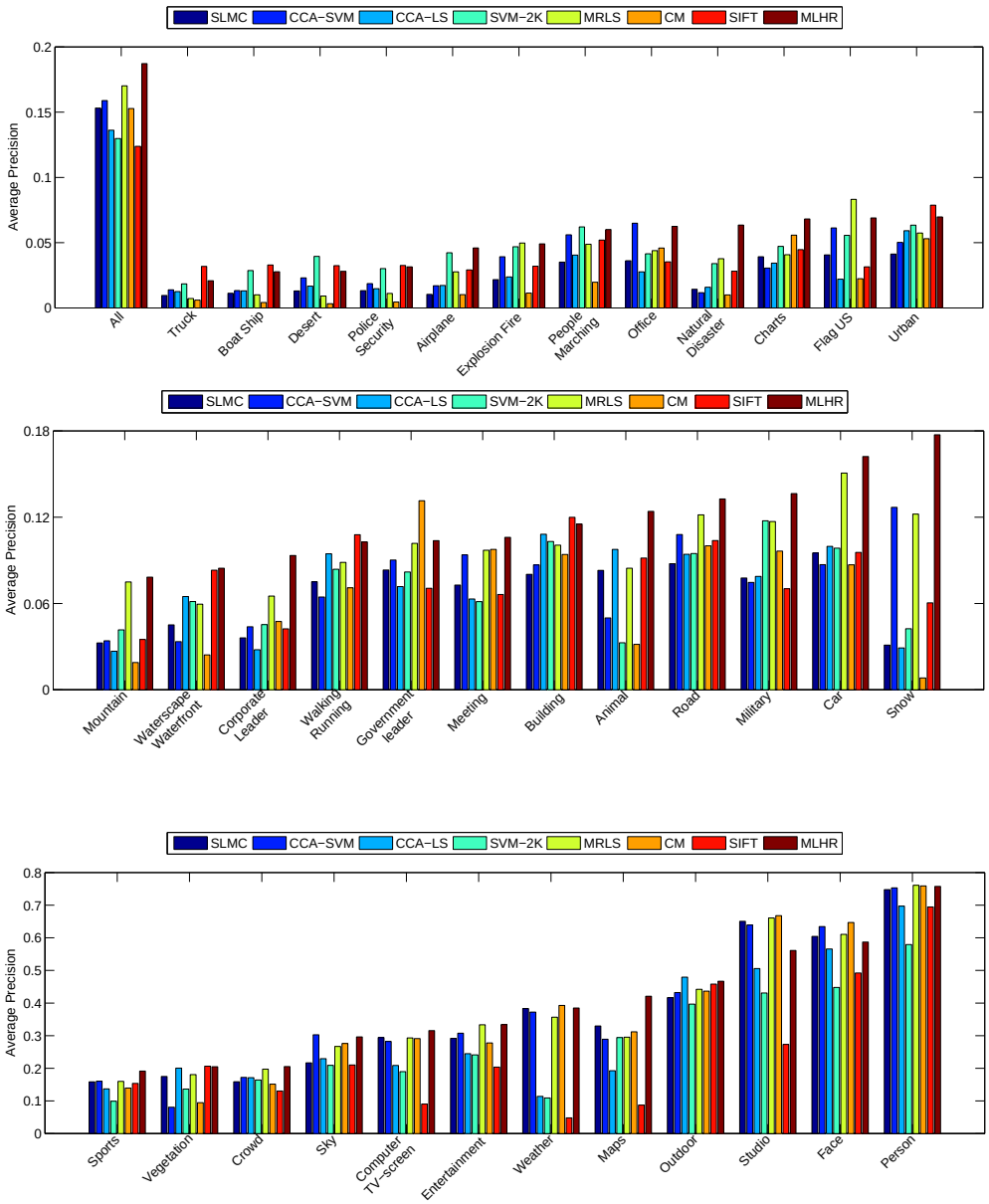


Fig. 2. A comparison of different algorithms for video concept annotation. This figure shows the APs corresponding to each concept of different algorithms. 'All' indicates the MAP of all the 36 concepts. CM and SIFT are used to represent the videos.

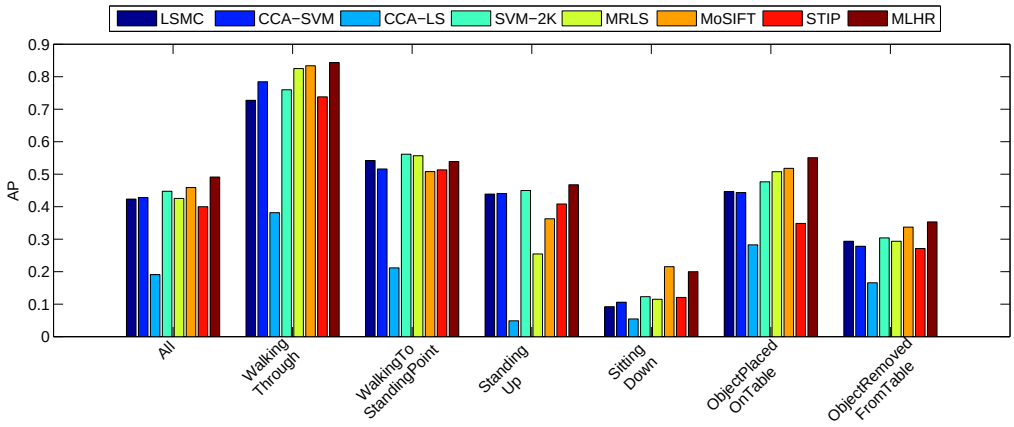


Fig. 3. A comparison of different algorithms for geriatric patient action recognition. This figure shows the APs corresponding to each concept of different algorithms. 'All' indicates the MAP of all the 6 concepts.

TABLE III
MAP OF THE 6 CONCEPTS OF ACTION RECOGNITION USING DIFFERENT ALGORITHMS.

	SLMC	CCA-SVM	CCA-LS	SVM-2K	MRLS	MoSIFT	STIP	MLHR
MAP	0.4245	0.4281	0.1918	0.4469	0.4255	0.4582	0.4002	0.4916

statistical approach is more capable to uncover the manifold structure than directly computing the pairwise distances [36].

V. CONCLUSION

In this paper, we have proposed a new multiple feature learning algorithm MLHR for multimedia content analysis. Compared to the existing related algorithms, MLHR has two major advantages. First, instead of computing the affinity matrix directly according to data features, MLHR employs a statistical approach to exploit the structural information of both the labeled and unlabeled data. Such approach is more capable to leverage the unlabeled data for semi-supervised learning. Second, the structural information of each feature type is preserved in MLHR during the training phase, resulting in more stable learning results. Although the semi-supervised multiple feature fusion algorithm has shown great potential for effective multimedia analysis, exploiting the unlabeled data might have negative effects if the manifold assumption does not hold. Also, it may not always be the case that including more features is beneficial for multimedia analysis. The future work includes designing: 1) algorithms that are able to predict if the unlabeled data will contribute; 2) methods that decide which of the multiple features should be combined.

REFERENCES

[1] TRECVID Retrieval Evaluation. [Online]. Available: http://trecvid.nist.gov/.

[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*, 2005, 1-17.

[3] M. Belkin, P. Niyogi and V. Sindhwani. Manifold Regularization: A Geometric Framework for Learning from Labled and Unlabeled Examples. *Journal of Machine Learning Research*, 12:2399-2434, 2006.

[4] M. Chen and A. Hauptmann. MoSIFT: Recognizing Human Actions in Surveillance Videos, *Tech. Rep.* Carnegie Mellon University, 2009.

[5] I. Cohen, F. Cozman, N. Sebe, M. Cirelo and T. Huang. Semisupervised Learning of Classifiers: Theory, Algorithms, and Their Application to Human-computerInteraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12):1553-1566, 2004.

[6] J. Farquhar, D. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak Two view learning: SVM-2K, theory and practice, In *Proc. NIPS*, 355-362, 2006.

[7] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *Proc. ICCV*, 2009.

[8] K. Goh, E. Chang and B. Li. Using One-Class and Two-Class SVMs for Multiclass Image Annotation. *IEEE Transactions on Knowledge and Data Engineering*, 12:2399-2434, 2006.

[9] A. Hauptmann, Lessons for the future from a decade of informedia video analysis research. In *Proc. ACM Int. Conf. Image and Video Retrieval*, 1-10, 2005.

[10] A. Hauptmann, R. Yan, W. Lin, M. Christel, and H. Wactlar. Can High-Level Concepts Fill the Semantic Gap in Video Retrieval? A Case Study With Broadcast News. *IEEE Transactions on Multimedia*, 9(5):958-966, 2007.

[11] H. Hotelling. Relations between Two Sets of Variates. *Biometrika*, 28(3-4):321-377, 1936.

[12] S. Ji, L. Tang, S. Yu, and J. Ye. A Shared-subspace Learning Framework for Multi-label Classification. *ACM Transactions on Knowledge Discovery from Data*, 2(1):8:1-8:29, 2010.

[13] T. Kim, J. Kittler and R. Cipolla. Discriminative Learning and Recognition of Image Set Classes Using Canonical Correlations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1005-1018, 2007.

[14] Z. Lan, L. Bao, S. Yu, W. Liu, and A. Hauptmann. Double Fusion for Multimedia Event Detection. In *Proc. Multimedia Modeling*,2012.

[15] I. Laptev and T. Lindeberg. Space-time Interest Points. In *Proc. ICCV*, 2003.

[16] M. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based Multimedia Information Retrieval: State-of-the-art and Challenges. *ACM Transactions on Multimedia Computing, Communication, and Applications*, 2(1): 1-19, 2006.

[17] H. Li, Y. Shi, M. Chen, A. Hauptmann, Z. Xiong. Hybrid active learning for cross-domain video concept detection. In *Proc. ACM MM* 1003-1006, 2010.

[18] G. Li, S. Hoi and K. Chang Two-View Transductive Support Vector Machines. In *Proc. SDM*, 235-244, 2010.

[19] C. Lin, B. Tseng, and J. Smith. Video Collaborative Annotation Forum: Establishing Ground-Truth Labels on Large Multimedia Datasets. In *Proc. TRECVID Workshop*, 2003.

[20] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2(60):91-110, 2004.

[21] H. Ma, J. Zhu, M. Lyu, and I. King. Bridging the Semantic Gap Between Image Contents and Tags. *IEEE Transactions on Multimedia*, 12(5):462-47, 2010.

[22] Z. Ma, F. Nie, Y. Yang, J. Uijlings, N. Sebe, A. Hauptmann. Discriminating Joint Feature Analysis for Multimedia Data Understanding. *IEEE Transactions on Multimedia*, DOI=10.1109/TMM.2012.2199293.

[23] Z. Ma, Y. Yang, Y. Cai, N. Sebe, A. Hauptmann. Knowledge Adaptation for Ad Hoc Multimedia Event Detection with Few Examplars. In *Proc. ACM MM*, 2012.

[24] P. Mylonas, E. Spyrou, Y. Avrithis and S. Kollias. Using Visual Context and Region Semantics for High-level Concept Detection. *IEEE Transactions on Multimedia*, 11(2):229-243, 2009.

[25] F. Nie, D. Xu, I. Tsang, and C. Zhang. Flexible Manifold Embedding: A Framework for Semi-Supervised and Unsupervised Dimension Reduction. *IEEE Transactions on Image Processing*, 19(7):1921-1932, 2010.

[26] C. Snoek, M. Worring and A. Smeulders Early versus Late Fusion in Semantic Video Analysis. In *Proc. ACM MM*, 399-402, 2005.

[27] J. Song, Y. Yang, Z. Huang, H. Shen and R. Hong. Multiple Feature Hashing for Real-time Large Scale Near-duplicate Video Retrieval. In *Proc. ACM MM*, 423-432, 2011.

[28] A. Vinokourov, J. Shawe-taylor and N. Cristianini. Inferring a Semantic Representation of Text via Cross-Language Correlation Analysis. In *Proc. NIPS*, 339-342, 2003.

[29] M. Wang, X. Hua, R. Hong, J. Tang, and R. Song. Beyond Distance Measurement: Constructing Neighborhood Similarity for Video Annotation. *IEEE Transactions on Multimedia*, 11(3):465-476, 2009.

[30] R. Yan and A. Hauptmann. The combination limit in multimedia retrieval. In *Proc. ACM MM*, 339-342, 2003.

[31] R. Yan and A. Hauptmann. Probabilistic Latent Query Analysis for Combining Multiple Retrieval Sources. In *Proc. ACM SIGIR*, 324-331, 2006.

[32] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia Universitys Baseline detectors for 374 LSCOM Semantic Visual Concepts. *Tech. Rep.* Columbia University, 2007.

[33] Y. Yang, Y. Zhuang, F. Wu and Y. Pan. Harmonizing Hierarchical Manifolds for Multimedia Document Semantics Understanding and Cross-Media Retrieval. *IEEE Transactions on Multimedia*, 10(3):437-446, 2008.

[34] Y. Yang, Y. Zhuang, D. Xu, Y. Pan, D. Tao and S. Maybank. Retrieval Based Interactive Cartoon Synthesis via Unsupervised Bi-Distance Metric Learning. In *Proc. ACM MM*, 311-320, 2009.

[35] Y. Yang, D. Xu, F. Nie, J. Luo and Y. Zhuang. Ranking with Local Regression and Global Alignment for Cross Media Retrieval. In *Proc. ACM MM*, 175-184, 2009.

[36] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang and Y. Pan. A Multimedia Retrieval Framework based on Semi-Supervised Ranking and Relevance Feedback. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 34(4): 723-742, 2012.

[37] Z. Zha, M. Wang, Y. Zheng, Y. Yang, R. Hong and T. Chua. Interactive Video Indexing with Statistical Active Learning. *IEEE Transactions on Multimedia*, 14(1):17-27, 2012.

[38] D. Zhou, O. Bousquet, T. Navin Lal, J. Weston and B. Schölkopf. Learning with Local and Global Consistency. In *Proc. NIPS*, 321-328, 2004.

[39] X. Zhu. Semi-Supervised Literature Survey. *Tech. Rep.* University of Wisconsin Madison, 2006.