Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Multiple hypergraph ranking for video concept detection*

Ya-hong HAN, Jian SHAO‡, Fei WU, Bao-gang WEI

(*Department of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*)

E-mail: yahong@zju.edu.cn; jshao@zju.edu.cn; wufei@cs.zju.edu.cn; wbg@zju.edu.cn

Received July 25, 2009; Revision accepted Sept. 30, 2009; Crosschecked May 11, 2010

Abstract: This paper tackles the problem of video concept detection using the multi-modality fusion method. Motivated by multi-view learning algorithms, multi-modality features of videos can be represented by multiple graphs. And the graph-based semi-supervised learning methods can be extended to multiple graphs to predict the semantic labels for unlabeled video data. However, traditional graphs represent only homogeneous pairwise linking relations, and therefore the high-order correlations inherent in videos, such as high-order visual similarities, are ignored. In this paper we represent heterogeneous features by multiple hypergraphs and then the high-order correlated samples can be associated with hyperedges. Furthermore, the multi-hypergraph ranking (MHR) algorithm is proposed by defining Markov random walk on each hypergraph and then forming the mixture Markov chains so as to perform transductive learning in multiple hypergraphs. In experiments on the TRECVID dataset, a triple-hypergraph consisting of visual, textual features and multiple labeled tags is constructed to predict concept labels for unlabeled video shots by the MHR framework. Experimental results show that our approach is effective.

Key words: Multiple hypergraph ranking, Video concept detection, Multi-view learning, Multiple labeled tags, Clustering
doi:10.1631/jzus.C0910453 **Document code:** A **CLC number:** TP391

1 Introduction

With the explosion of digital video data, managing and retrieving the videos becomes more and more challenging. Recently, many studies have been done to tackle the problem of video concept detection by machine learning methods, which include supervised, semi-supervised, and graph-based transductive learning methods. For example, Yanagawa *et al*. (2007) trained three support vector machine (SVM) classifiers individually over each of three low-level feature spaces on training data and adopted the late fusion strategy to combine the concept detection results for test data. Moreover, by computing pairwise visual similarities, an adjacency graph on video data could be constructed. And then the graph-based transductive learning methods, such as local and global consistency (LGC) (Zhou *et al*., 2004b) and Gaussian random field with harmonic function (GRF) (Zhu *et al*., 2003) algorithms, can be used to detect the video concept (Tang *et al*., 2007; Wang *et al*., 2007a). These methods exploit only the low-level features of video data. The performance improvement is bounded by the 'semantic gap' between low-level features and high-level concepts. Video data involve multi-modality information, such as visual and textual information. Many research efforts (Tong *et al*., 2005; Liu J *et al*., 2007; Wang *et al*., 2007b; Zhang *et al*., 2007; Hoi and Lyu, 2008; Liu Y *et al*., 2008; Tan *et al*., 2008; Weng and Chuang, 2008; Yang *et al*., 2008) have been made to circumvent the 'semantic gap' by a multi-modality fusion approach. To use the multi-modality features effectively in video semantic analysis, this paper proposes a novel hypergraph-based multi-modality fusion approach. We tackle mainly three problems: multi-modality representation, multi-modality fusion strategy, and multiple labeled tags.

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 60603096 and 60673088), the National High-Tech Research and Development Program (863) of China (No. 2006AA010107), and the Program for Changjiang Scholars and Innovative Research Team in University of China (No. IRT0652)

1. Multi-modality representation. Traditional concatenated vector methods for multi-modality fusion always cause the problem of 'curse of dimensionality' (Liu Y *et al*., 2008). Moreover, correlations among different modalities may be lost by such simple concatenated representation. So Liu Y *et al*. (2008) represented image, audio, and text modalities in video shots by the 3rd-order tensor called 'tensor shot'. Different from the tensor model, more studies (Liu J *et al*., 2007; Wang *et al*., 2007b; Hoi and Lyu, 2008) represent multi-modality of videos by multi-graph, and then the generalized multi-graph learning algorithms (Tong *et al*., 2005; Wang *et al*., 2007b) are used to improve the detection precision. Usually, the graph model takes video shots as vertex, and the weights of pairwise linking are similarities calculated from corresponding feature spaces. However, these graph representation methods consider only the homogeneous pairwise linking of shots and therefore ignore the high-order relations that are inherent in that representation (Tan *et al*., 2008). For example, visual similarities among key video frames are not always pairwise links. Temporal continuous key frames and the so-called near-duplicate key frames (NDKs) always involve more than two visually similar shots (Zhao *et al*., 2007). Moreover, a textual feature from automatic speech recognition (ASR) video transcription is associated with several shots in one story or news item. Instead of having edges between pairs of vertices, hypergraphs have edges that connect sets of two or more vertices, called 'hyperedges' (Tan *et al*., 2008). To use the PageRank algorithm, the hypergraph model in Tan *et al*. (2008) is converted to a simple graph by star expansion. Therefore the method in Tan *et al*. (2008) is not a hypergraph-native method and cannot be used to integrate other modalities. To represent the high-order correlations, we represent each modality of videos by a hypergraph and conduct the multi-modality fusion on multiple hypergraphs.

2. Multi-modality fusion strategy. Two schemes are usually adopted (Tong *et al*., 2005) to fuse multiple modalities. First, fusion is implemented at a low-level feature layer, which can be classified as an early fusion scheme (Wang *et al*., 2007b). Second, fusion is performed at an output level, e.g., the SVM classifier fusion method (Yanagawa *et al*., 2007), which can be classified as a late fusion scheme (Wang *et al*., 2007b). As introduced in Wang *et al*. (2007b), multi-graph fusion can be taken as a middle fusion scheme. In order to extend the transductive learning methods to multi-graph, different graph fusion methods have been proposed, for example, linear and sequential fusion (Tong *et al*., 2005) or convex combination of discrete Laplacians for each graph (Wang *et al*., 2007b). However, the underlying principle of these linear or convex combination methods is not clear (Tong *et al*., 2005; Zhou and Burges, 2007). Moreover, learning from multi-graph can be taken as multi-view learning (Bickel and Scheffer, 2004; Virginia, 2005; Zhou and Burges, 2007; Long *et al*., 2008), in which the same instances may have multiple representations from different graph spaces (Long *et al*., 2008). Thus, we propose a novel multi-hypergraph learning framework for multi-modality fusion to detect the video concept. Motivated by the multi-graph learning algorithms proposed in Zhou and Burges (2007), we first define Markov random walk in each hypergraph and then form a mixture of Markov chains on multiple hypergraphs. Labels are propagated from labeled shots to unlabeled shots in such a mixture of Markov chains, resulting in a stable and optimal state in each hypergraph.

3. Multiple labeled tags. Videos and images are intrinsically and visually polysemous; thus, they may be annotated by multiple labeled tags. Fig. 1, for example, represents the key frames of corresponding shots selected from the TRECVID dataset. There are apparently a lot of interesting objects/regions in each image so that one may have a lot to comment on the image. Multiple labeled tags, such as fire balloon, sky, car, person, road, and desert, may be associated with Fig. 1a. Similarly, tags such as tree, person, flowers, building, and sky may be associated with Fig. 1b. Furthermore, people may deduce some concepts of scenes or events from the whole images, which may not apparently occur in the images. It is suitable that the concept 'outdoor' is associated with these two

(a) (b)

Fig. 1 Examples of visually polysemous (ellipses denote visual objects)

Multiple labeled tags such as fire balloon, sky, car, person, road, and desert may be associated with (a); tags such as tree, person, flowers, building, and sky may be associated with (b)

images simultaneously. As shown by statistics in Qi *et al*. (2007), the TRECVID dataset has a multi-labeling property. Moreover, as shown in Fig. 1, there are strong correlative relations among tags. For example, outdoor often co-occurs with sky. Recently, multi-label image/video understanding has attracted many researchers (Zhang and Zhou, 2008; Zha *et al*., 2009), and many studies (Qi *et al*., 2007; Wang *et al*., 2008; Weng and Chuang, 2008) have exploited the multi-label correlations for video semantic analysis. In this paper, we represent the multi-labeled tags associated with video shots as a hypergraph. The idea of 'hypergraph for multi-labeled tags' is shown in Fig. 2c. Concepts from $\boldsymbol{c}_1$ to $\boldsymbol{c}_4$ are snow, outdoor, office, and person. Video shots from $\boldsymbol{s}_1$ to $\boldsymbol{s}_8$ can be represented by a row vector respectively, in which value 1 denotes that the corresponding shot is labeled by concept tags at that position, and 0 otherwise. For example, $\boldsymbol{s}_1$ is labeled by tags of snow, outdoor, and person simultaneously. Each tag $\boldsymbol{c}_i$ can be represented by a hyperedge, which is a column vector. So each concept tag may associate multiple shots; e.g., person, denoted by $\boldsymbol{c}_4$, associates six shots. By hypergraph, therefore, homogeneous correlations (i.e., shot-shot or concept-concept) and heterogeneous correlations (i.e., shot-concept) are integrated into one view seamlessly.

Fig. 2 shows the triple-hypergraph representation. The same set of video shots is represented by triple views: visual-feature hypergraph, textual-feature hypergraph, and hypergraph for multi-labeled tags. The visual- and textual-feature hypergraphs are constructed using a novel clustering-based hypergraph construction (CBHC) algorithm proposed in this paper. As introduced above, the representation of a hypergraph for multiple labeled tags is straightforward. As a result, we combine visual, textual features and multi-labeled tags in such a triple-hypergraph. By defining a mixture of Markov chains on the multi-hypergraph, we propose a multiple hypergraph ranking (MHR) algorithm, and use the hypergraph-based multi-modality fusion framework for video concept detection. Empirical studies for MHR and CBHC algorithms are also presented.

2 Related works

2.1 Graph learning for video annotation

Graph-based semi-supervised learning algorithms, such as LGC (Zhou *et al*., 2004a) and GRF (Zhu *et al*., 2003), have been well studied and applied into video annotation and concept detection (Tang *et al*., 2007; Wang *et al*., 2007a). The basic idea is to propagate concept labels from labeled data to unlabeled data in a graph that is constructed for video shots. Traditionally, similarity of two samples is estimated according to the Euclidean distance between them. Tang *et al*. (2007) and Wang *et al*. (2007a) considered the structural assumption in similarity measures by combining the local distribution of differences. However, since these methods take into account only the visual features of videos, as discussed above, their performance of video semantic analysis is inevitably bounded.

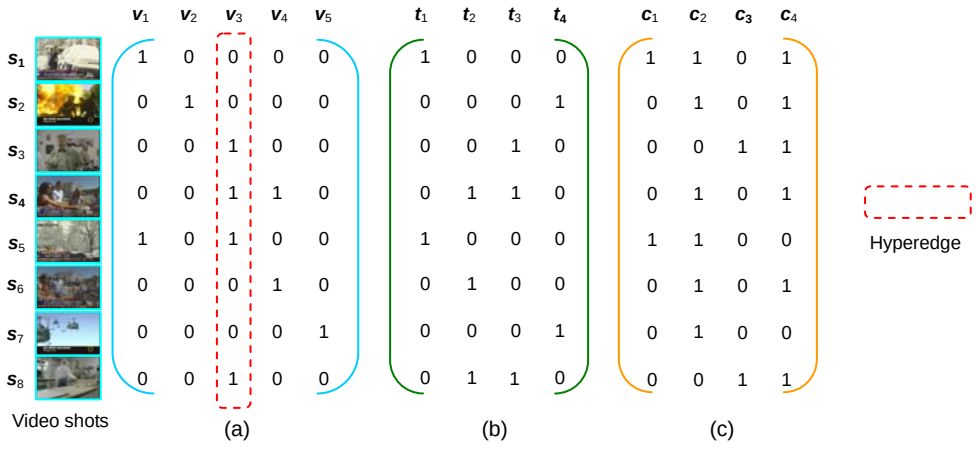

Fig. 2 Example of triple-hypergraph representation for video shots

(a) Visual-feature hypergraph; (b) Textual-feature hypergraph; (c) Hypergraph for multi-labeled tags. Value 1 denotes that the corresponding shot is labeled by the concept tag or associated with the visual/textual-feature hyperedge, and 0 otherwise

Furthermore, if each modality is represented by a graph, we can fuse multi-modality features of videos on a multi-graph model. Many efforts have been made to extend the graph-based semi-supervised learning algorithms over multi-graph and to develop the multi-graph learning algorithm for fusion (Tong *et al*., 2005; Liu *et al*., 2007; Wang *et al*., 2007b; Hoi and Lyu, 2008). However, all above methods consider only the pairwise similarities and ignore the high-order relations.

2.2 Multi-view learning and random walk on multi-graph

Multi-view learning has been well explored recently in both machine learning and data mining fields (Bickel and Scheffer, 2004; Virginia, 2005; Zhou and Burges, 2007; Long *et al*., 2008). In multi-modality fusion, the same instances may have multiple representations from different feature spaces or graph spaces (Long *et al*., 2008). Therefore the transductive multi-graph learning methods, e.g., convexly combining discrete Laplacians for each graph (Wang *et al*., 2007b), can be taken as a multi-view learning approach.

Different from the graph Laplacians combining methodology, Zhou and Burges (2007) defined a mixture Markov model on multiple graphs to formulate a transductive learning algorithm on multiple graphs. The basic intuition is to form random walk and label propagation on multiple graphs so as to define a mixture of Markov chains. Starting from a vertex in one graph, the random walker may stay at such a vertex according to a certain stationary distribution, or jump to other adjacency vertexes in the same graph by a certain transition probability, or even jump to the arbitrary vertex in other different graphs. From the view of a multi-graph normalized cut, the motivation of multi-graph transductive learning is to find a cut that is optimal on each graph (Zhou and Burges, 2007). Zhou and Burges (2007) made such a multi-graph normalized cut problem a real-valued optimization problem and proposed the transductive learning algorithm with multiple graphs.

2.3 Multi-label learning

Multi-label learning resolves the problem where each training instance is associated with multiple class labels. Zhang and Zhou (2008) proposed a maximum margin method for multi-instance multi-label learning (M^3MIML) algorithm. To tackle the semi-supervised learning with multiple labels, Zha *et al*. (2009) proposed a graph-based multi-label learning framework, which is the extension of the LGC and GRF methods. Since video shots are associated with multiple labeled tags, semantic video analysis can be performed by multi-label learning methods. Qi *et al*. (2007) modeled multi-labeling by Gibbs random fields, and the proposed correlative multi-labeling (CML) approach gained performance improvement in video concept detection. Similarly, Wang *et al*. (2008) presented a discrete hidden Markov random field (dHMRF) model. To improve the accuracy of semantic video indexing, Weng and Chuang (2008) proposed a multi-cue fusion approach to exploit both multi-labeling correlations and temporal dependency among shots. Sun *et al*. (2008) employed hypergraphs to capture the correlation information among different labels for improving classification performance.

3 Multiple hypergraph ranking framework

3.1 Problem formulation and definition

Predicting the label for unlabeled video shots from the labeled training set can be considered a transductive classification problem. We cast the video concept detection as a classification problem, and view the ranking as an extreme case of classification, in which only positive examples are available (Zhou *et al*., 2004a).

The triple hypergraphs G_v, G_t, and G_c denote the visual-feature hypergraph, textual-feature hypergraph, and hypergraph for multi-labeled tags, respectively. Let V denote a set of video shots and S the labeled subset of video shots V. For a certain class label, our goal is to predict the label of the remaining video shots in the unlabeled subset S^c, which is the complement of S. Let f: $V\rightarrow S$ denote the classification function. Define a function y with $y(\boldsymbol{v})=1$ for positive shots and -1 for negative items if $\boldsymbol{v}\in S$, and 0 if $\boldsymbol{v}\in S^c$. Then we can solve the classification function f by a regularization framework as

$$\min_{f\in\mathbb{R}^{|V|}}[\Omega(f)+\gamma\,\|f-y\|^2], \tag{1}$$

where $\gamma>0$ is the parameter specifying the tradeoff between the first smoothness constraint and the second fitness constraint terms. In our framework, functional $\Omega(f)$ imposes smoothness on values of f over the mixture of the Markov chains defined on the triple hypergraphs G_v, G_t, and G_c.

In the following, our main task is to develop the solution to the optimization problem (1). We firstly introduce preliminaries of the hypergraph. By combining random walk on different hypergraphs, we present the multi-hypergraph ranking algorithm.

3.2 Random walk on hypergraph

Let $G=(V, E, w)$ denote a weighted hypergraph, where V denotes a finite set of nodes, and E denotes a family of subsets e of V such that $\cup_{e\in E}=V$. We call V the vertex set and E the hyperedge set. A positive number $w(e)$ associated with each hyperedge e is called the weight of hyperedge e. Note that, $w(e)=1$ for an unweighted hypergraph. For a vertex $\boldsymbol{v}\in V$, its degree is defined by $d(\boldsymbol{v})=\sum_{e\in E|\boldsymbol{v}\in e} w(e)$; for a hyperedge $e\in E$, its degree is defined as $\delta(e)=|e|$. A hypergraph G can be represented by a $|V|\times|E|$ matrix $\boldsymbol{H}$ with entries $h(\boldsymbol{v}, e)=1$ if $\boldsymbol{v}\in e$ and 0 otherwise. Then $d(\boldsymbol{v})=\sum_{e\in E} w(e)h(\boldsymbol{v},e)$, and $\delta(e)=\sum_{\boldsymbol{v}\in V} h(\boldsymbol{v},e)$. Let $\boldsymbol{D}_v$ and $\boldsymbol{D}_e$ denote the diagonal matrices containing the vertex and hyperedge degrees, respectively, and $\boldsymbol{W}$ the diagonal matrix containing the weights of hyperedges. Then the adjacency matrix $\boldsymbol{A}$ of hypergraph G is defined as $\boldsymbol{A}=\boldsymbol{H}\boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}-\boldsymbol{D}_v$ (Zhou *et al*., 2007).

Random walk on hypergraph was firstly introduced to give random walk explanation for the normalized hypergraph cut (Zhou *et al*., 2007). For a vertex subset $S\subset V$, let S^c denote the complement of S. A cut of hypergraph G is a partition of V into two parts S and S^c. The hypergraph boundary ∂S of S is defined as

$$\partial S=\{e\in E\mid e\cap S\neq\varnothing,\ e\cap S^c\neq\varnothing\}.$$

∂S can be taken as a hyperedge set consisting of hyperedges that are cut. The volume of S is defined as $\mathrm{vol}(S)=\sum_{\boldsymbol{v}\in S} d(\boldsymbol{v})$. Similarly, the volume of ∂S is defined by (Zhou *et al*., 2007)

$$\mathrm{vol}(\partial S)=\sum_{e\in\partial S} w(e)\frac{|e\cap S|\cdot|e\cap S^c|}{\delta(e)},$$

where $|a|$ denotes the number of elements in set a. Clearly, we have $\mathrm{vol}(\partial S)=\mathrm{vol}(\partial S^c)$. From above definitions, the normalized hypergraph cut can be formulated as follows (Zhou *et al*., 2007):

$$\min_{S\subset V,\ S\neq\varnothing}\left[c(S)=\frac{\mathrm{vol}(\partial S)}{\mathrm{vol}(S)\cdot\mathrm{vol}(S^c)}\right]. \tag{2}$$

Random walk on hypergraph generalizes the random walk transition rule defined on simple graphs as follows (Zhou *et al*., 2007): Given the current position $\boldsymbol{u}\in V$, first choose a hyperedge e over all hyperedges incident with $\boldsymbol{u}$ with a probability proportional to $w(e)$, and then choose a vertex $\boldsymbol{v}\in e$ uniformly at random. Let $\boldsymbol{P}$ denote the transition probability matrix of this hypergraph random walk. Then each entry $p(\boldsymbol{u},\boldsymbol{v})$ of $\boldsymbol{P}$ is defined as (Zhou *et al*., 2007)

$$p(\boldsymbol{u},\boldsymbol{v})=\sum_{e\in E} w(e)\frac{h(\boldsymbol{u},e)}{d(\boldsymbol{u})}\frac{h(\boldsymbol{v},e)}{\delta(e)}, \tag{3}$$

while the stationary distribution of the random walk is

$$\pi(\boldsymbol{v})=\frac{d(\boldsymbol{v})}{\mathrm{vol}(V)}. \tag{4}$$

For random walk on hypergraph, we can derive that

$$\sum_{\boldsymbol{u}\in V}\pi(\boldsymbol{u})p(\boldsymbol{u},\boldsymbol{v})=\pi(\boldsymbol{v}). \tag{5}$$

From the stationary distribution, we have

$$\frac{\mathrm{vol}(S)}{\mathrm{vol}(V)}=\sum_{\boldsymbol{v}\in V}\pi(\boldsymbol{v}), \tag{6}$$

which is the probability of the random walk occupying some vertex in S. Moreover, the probability of a jump of the random walk from S to S^c is formulated as follows:

$$\frac{\mathrm{vol}(\partial S)}{\mathrm{vol}(V)}=\sum_{\boldsymbol{u}\in S}\sum_{\boldsymbol{v}\in S^c}\pi(\boldsymbol{u})p(\boldsymbol{u},\boldsymbol{v}). \tag{7}$$

Therefore, a normalized hypergraph cut can be interpreted as a cut under the stationary distribution. While the probability of transition from one cluster to another is as small as possible, the probability of remaining in the same cluster is as large as possible.

3.3 Mixture Markov chain for multi-hypergraph

For convenience, we take two hypergraphs for example, which can be straightforwardly generalized to multiple hypergraphs.

Assume two hypergraphs $G_i=(V, E_i, w_i)$ (i=1, 2) share the same set of vertices while having heterogeneous edges and weights. Suppose S is a nonempty subset of V. We define the multi-volume of S as

$$\mathrm{mvol}(S)=\alpha\,\mathrm{vol}_1(S)+(1-\alpha)\,\mathrm{vol}_2(S), \tag{8}$$

and the multi-volume of ∂S as

$$\mathrm{mvol}(\partial S)=\alpha\,\mathrm{vol}_1(\partial S)+(1-\alpha)\,\mathrm{vol}_2(\partial S), \tag{9}$$

where α is a parameter in [0, 1]. Then the multiple-hypergraph cut problem can be formulated as

$$\min_{S\subset V,\, S\neq\varnothing}\left[c(S)=\frac{\mathrm{mvol}(\partial S)}{\mathrm{mvol}(S)\cdot\mathrm{mvol}(S^{\mathrm{c}})}\right]. \tag{10}$$

Clearly, when α=0 or 1, expression (10) reduces to the cut for a single hypergraph as expression (2).

Now we define the mixture Markov chain on hypergraphs G_1 and G_2. First, we define the transition probabilities as follows:

$$p^m(\boldsymbol{u},\boldsymbol{v})=\beta_1(\boldsymbol{u})\,p_1(\boldsymbol{u},\boldsymbol{v})+\beta_2(\boldsymbol{u})\,p_2(\boldsymbol{u},\boldsymbol{v}), \tag{11}$$

where $p_i(\boldsymbol{u},\boldsymbol{v})$ (i=1, 2) are defined as Eq. (3). Functions $\beta_i(\boldsymbol{u})$ (i=1, 2) are given by

$$\beta_1(\boldsymbol{u})=\frac{\alpha\pi_1(\boldsymbol{u})}{\alpha\pi_1(\boldsymbol{u})+(1-\alpha)\pi_2(\boldsymbol{u})}, \tag{12}$$

and

$$\beta_2(\boldsymbol{u})=\frac{(1-\alpha)\pi_2(\boldsymbol{u})}{\alpha\pi_1(\boldsymbol{u})+(1-\alpha)\pi_2(\boldsymbol{u})}. \tag{13}$$

It is easy to check that $\beta_1(\boldsymbol{u})+\beta_2(\boldsymbol{u})=1$ and $\beta_i\geqslant 0$.

Then we define the stationary distribution $\pi^m(\boldsymbol{v})$ of the mixture Markov model as

$$\pi^m(\boldsymbol{v})=\alpha\pi_1(\boldsymbol{v})+(1-\alpha)\pi_2(\boldsymbol{v}), \tag{14}$$

where $\pi_i(\boldsymbol{v})$ (i=1, 2) are defined as Eq. (4).

As discussed in Zhou and Burges (2007), by Eqs. (12) and (13) we can derive that, β_1 and β_2 vary from vertex to vertex rather than being a constant. Therefore the mixture of random walk transition probabilities $p^m(\boldsymbol{u},\boldsymbol{v})$ is not simply a linear combination of the transition probability matrices on each hypergraph.

3.4 Multiple hypergraph ranking algorithm

Given a set of hypergraphs $G_i=(V, E_i, w_i)$ (i=1, 2, …, k) with a common vertex set V, and a subset of vertices $S\subset V$ labeled as 1 or −1, our goal of using multiple hypergraphs for classification is to predict the labels of the remaining unlabeled vertices in S^{c}. Let f: $V\to S$ denote the classification function. Define a function y with $y(\boldsymbol{v})$=1 or −1 if $\boldsymbol{v}\in S$, and 0 if $\boldsymbol{v}\in S^{\mathrm{c}}$. Then we can transform problem (1) as the following optimization problem:

$$\min_{f\in\mathbb{R}^{|V|}}\left[\sum_{\boldsymbol{u},\boldsymbol{v}\in V}\pi^m(\boldsymbol{u})\,p^m(\boldsymbol{u},\boldsymbol{v})(f(\boldsymbol{u})-f(\boldsymbol{v}))^2+\gamma\sum_{\boldsymbol{v}\in V}\pi^m(\boldsymbol{v})(f(\boldsymbol{v})-y(\boldsymbol{v}))^2\right], \tag{15}$$

where the parameter γ>0. Note that the first term in expression (15) is a smoothness constraint, which forces function f to change as slowly as possible on densely connected subgraphs, and the second term is a fitness constraint, which forces function f to fit the given labels as well as possible. The tradeoff between these two requirements is measured by parameter γ.

Hence, multiple hypergraph ranking can be formalized as: given a set of labeled vertices $S\subset V$, rank the remaining unlabeled vertices in S^{c} with respect to their relevance to the given vertex. Define y with $y(\boldsymbol{v})$=1 or 0 if $\boldsymbol{v}\in S$, and then rank the unlabeled vertices $\boldsymbol{u}\in S^{\mathrm{c}}$ according to $f(\boldsymbol{u})$.

To solve the optimization problem (15), we differentiate the objective function in (15) with respect to f and obtain

$$\left(\gamma\boldsymbol{\Pi}+\boldsymbol{\Pi}-\frac{\boldsymbol{\Pi P}+\boldsymbol{P}^{\mathrm{T}}\boldsymbol{\Pi}}{2}\right)f=\gamma\boldsymbol{\Pi y}, \tag{16}$$

where $\boldsymbol{P}$ denotes the matrix with the elements $p^m(\boldsymbol{u},\boldsymbol{v})$, and $\boldsymbol{\Pi}$ the diagonal matrix with diagonal elements $\pi^m(\boldsymbol{u})$. Let $\lambda=1/(1+\gamma)$. Then we have $0<\lambda<1$, and the linear system Eq. (16) can be rewritten as

$$\left(\boldsymbol{\Pi}-\lambda\frac{\boldsymbol{\Pi P}+\boldsymbol{P}^{\mathrm{T}}\boldsymbol{\Pi}}{2}\right)f=(1-\lambda)\boldsymbol{\Pi y}. \tag{17}$$

Define

$$\boldsymbol{M}=\boldsymbol{\Pi}-\lambda\frac{\boldsymbol{\Pi P}+\boldsymbol{P}^{\mathrm{T}}\boldsymbol{\Pi}}{2},$$

and then the linear system Eq. (16) has the closed-form solution

$$f=(1-\lambda)\boldsymbol{M}^{-1}\boldsymbol{\Pi} y. \tag{18}$$

As discussed in Zhou and Burges (2007) and Zhou *et al*. (2007), linear system Eq. (18) is positive definite, and we can avoid computing the inverse but instead use a fast solver such as that introduced in Spielman and Teng (2003) and Zhou *et al*. (2004a). We summarize the MHR algorithm in Algorithm 1.

Algorithm 1 Multiple hypergraph ranking (MHR)

Input: Given k hypergraphs $G_i=(V, E_i, w_i)$, $1\leq i\leq k$, assume the vertices in a subset $S\subset V$ have been labeled as 1 or 0. G is represented by matrix $\boldsymbol{H}$.

Output: Ranking for the unlabeled vertices in S^{c}.

Step 1: For each hypergraph G_i, associate it with a unique random walk distribution. Compute the transition probabilities p_i and the stationary distribution π_i using Eqs. (3) and (4) respectively.

Step 2: Define a mixture of random walk for G_i. Compute the unique stationary distribution for the mixture of random walk by

$$\pi^m(\boldsymbol{v})=\sum_{i\leq k}\alpha_i\pi_i(\boldsymbol{v}).$$

Mixture of transition probabilities is computed by

$$p^m(\boldsymbol{u},\boldsymbol{v})=\sum_{i\leq k}\beta_i(\boldsymbol{u})\cdot p_i(\boldsymbol{u},\boldsymbol{v}),$$

where $\sum_{j\leq k}\alpha_j=1$ and $\alpha_j\geqslant 0$.

Step 3: Define a function y on V with $y(\boldsymbol{v})=1$ if $\boldsymbol{v}\in S$ and $\boldsymbol{v}$ is positive, and 0 otherwise. Solve the linear system $\boldsymbol{Mf}=(1-\lambda)\boldsymbol{\Pi y}$.

In fact, there are k parameters in algorithm MHR: $\alpha_1, \alpha_2, \ldots, \alpha_{k-2}, 1-\sum_{i=1}^{k-2}\alpha_i$, and λ. Moreover, when $k=1$ and $\alpha_1=1$, MHR is simplified to rank vertices for a single hypergraph.

For the mixture Markov chain, in the remainder of this paper, each hypergraph is treated equally, i.e., $\alpha_i=1/k$, $i=1, 2, \ldots, k$. And for parameter λ in MHR, we choose $\lambda=0.5$.

4 Multi-modality of videos as multiple hypergraphs

In this section, we first describe the clustering-based hypergraph construction algorithm for the visual- and textual-feature hypergraphs, and then present the steps for the construction of the triple-hypergraph in Fig. 2.

4.1 Learning high-order correlations by clustering

As discussed previously, the key difference between a graph and a hypergraph is that the hyperedge in a hypergraph can represent high-order correlations of data items. If we ignore the weight of the hyperedge, such as the hypergraph representation in Fig. 2, each hyperedge can be taken as a cluster of data items associated by a corresponding hyperedge simultaneously. Therefore, in what follows, we learn the high-order correlations among video shots by clustering on visual- and textual-feature vectors, which are extracted from video shots using feature extraction methods.

Clusters as hyperedges. Let $G=(V, E)$ denote an unweighted hypergraph. Each hyperedge $\boldsymbol{e}_j\in E$ ($j=1, 2, \ldots, |E|$) can be represented by a column vector $\boldsymbol{e}_j$ of length $|V|$, where $\boldsymbol{e}_j(i)=1$ ($i=1, 2, \ldots, |V|$) denotes that the ith vertex $\boldsymbol{v}_i\in V$ is associated by hyperedge $\boldsymbol{e}_j$, and 0 otherwise. If $\boldsymbol{e}_j(i)=0$ and $i=1, 2, \ldots, |V|$, then hyperedge $\boldsymbol{e}_j$ is deleted. Under this definition, each hyperedge $\boldsymbol{e}\in E$ associates at least one vertex. Assume data items set V is clustered into cluster set C by a clustering algorithm. Let $\boldsymbol{c}_j$ ($j=1, 2, \ldots, |C|$) denote the cluster indicator (column) vector of length $|V|$ for cluster $\boldsymbol{c}_j\in C$, where $\boldsymbol{c}_j(i)=1$ ($i=1, 2, \ldots, |V|$) denotes that the ith data item $\boldsymbol{v}_i\in V$ belongs to cluster $\boldsymbol{c}_j$, and 0 otherwise. Intuitively, vectors $\boldsymbol{e}_j$ and $\boldsymbol{c}_j$ are equivalent when $|C|=|E|$. Therefore, under this formulation, we can represent each hyperedge as a cluster or each cluster as a hyperedge. Hypergraph $G=(V, E)$ can be constructed by concatenating the column vector $\boldsymbol{c}_j$ ($j=1, 2, \ldots, |E|$). Thus, $|V|\times|E|$ matrix $\boldsymbol{H}$ can be represented by $\boldsymbol{H}=[\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_{|E|}]$.

Hypergraph construction algorithm. To learn the high-order correlations among video shots under visual- and textual-feature spaces respectively, we first cluster video shots into clusters and then represent the clusters by hyperedges, by which we construct visual- and textual-feature hypergraphs, respectively. Before presenting the hypergraph construction algorithm, we explore four related problems as follows:

1. Single-item cluster. Assume only one data item $\boldsymbol{v}_i \in V$ is clustered into cluster $\boldsymbol{c}_j \in C$. We have $\boldsymbol{c}_j(i)=1$ and $\boldsymbol{c}_j(l)=0$, where $l=1, 2, \ldots, |V|$ and $l \neq i$. Note that $\boldsymbol{c}_j$ can be easily represented by hyperedge $\boldsymbol{e}_j$, since the hyperedge can represent a unary relation.

2. Hard clustering vs. soft clustering. By the hard clustering algorithm, if data item $\boldsymbol{v}_i$ belongs to cluster $\boldsymbol{c}_j$, it cannot belong to any other cluster, while by soft clustering, $\boldsymbol{v}_i$ may belong to more than one cluster. Equivalence between a cluster and a hyperedge should be formulated under soft clustering setting. As shown in Fig. 2, a shot may be associated by more than one hyperedge.

3. Clustering performance. As we know, no clustering algorithm can obtain good results for all types of data collections by performance metrics of clustering precision or clustering mutual information. To tackle this problem and effectively learn the high-order correlations among video shots, we combine multiple clustering results by multiple clustering algorithms and use the combined clustering results to construct the hypergraph.

4. Number of clusters. Automatic determination of the number of clusters is a challenge for many clustering algorithms, such as *k*-means and spectral co-clustering (Dhillon, 2001). The affinity propagation (AP) clustering algorithm was recently proposed by Frey and Dueck (2007). One of the key differences between AP clustering and other clustering methods is that AP does not take the target number of clusters as an input parameter. In this work, we take the number of clusters output by AP clustering as a reference and then determine the number of clusters for *k*-means and spectral co-clustering heuristically.

Therefore, in this work, we combine clustering results from three algorithms, i.e., *k*-means, spectral co-clustering, and AP clustering, to construct visual- and textual-feature hypergraphs. We summarize the proposed CBHC algorithm in Algorithm 2.

Algorithm 2 Clustering-based hypergraph construction (CBHC)

Input: Clustering results C^1, C^2, and C^3 for data collection V by three clustering algorithms, i.e., *k*-means, spectral co-clustering, and AP clustering, respectively, and corresponding cluster numbers k_1, k_2, and k_3.

Output: $|V|\times|E|$ matrix $\boldsymbol{H}$ for hypergraph G.

Step 1: for each $\boldsymbol{c}_j \in C^1$ ($j=1, 2, \ldots, |k_1|$)

$\boldsymbol{H}=[\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_{|k_1|}]$.

Step 2: for $i=2$ to 3

for each $\boldsymbol{c}_j \in C^i$ ($j=1, 2, \ldots, |k_i|$)

for each $\boldsymbol{c}_s$ in $\boldsymbol{H}$ satisfying $\boldsymbol{c}_j \neq \boldsymbol{c}_s$

concatenate $\boldsymbol{c}_j$ to $\boldsymbol{H}$.

4.2 Visual-feature hypergraph construction

Step 1: visual-feature extraction.

We extract one key frame within each shot, and then extract visual features of the corresponding shot from the key frames. We use three different types of image features (Yanagawa *et al*., 2007): edged direction histogram (EDH, 73 dimensions), Gabor (GBR, 48 dimensions), and grid color moment (GCM, 225 dimensions). As a result, we concatenate and normalize all three features into one 346-dimensional visual feature vector.

Step 2: Construct the visual-feature hypergraph by applying the CBHC algorithm in such a visual feature space as obtained in Step 1.

4.3 Textual-feature hypergraph construction

Step 1: textual-feature extraction.

We use the automatic speech recognition (ASR) transcript as the source text in video. Due to the asynchrony between the words and visual stream in video, we use a sliding-window and define the text associated with a shot as the text occurring within a corresponding shot. Since ASR associates with a number of shots within a video story or news item, we choose a much larger size of sliding-window heuristically. After performing stemming by Porter's algorithm and removing stop words, we represent the text of each shot by the tf-idf features. Finally, we use latent semantic analysis (LSA) (Dumais *et al*., 1998) to reduce the text dimension.

Step 2: Construct the textual-feature hypergraph by applying the CBHC algorithm in such a textual feature space as obtained in Step 1.

4.4 Multi-labeled tags as hypergraph

As introduced in Section 1, the representation of multi-labeled tags as a hypergraph is straightforward. As shown by the 'hypergraph for multi-labeled tags' in Fig. 2c, value 1 in hyperedge $\boldsymbol{c}_i$ denotes that corresponding video shot is labeled by concept $\boldsymbol{c}_i$, and 0 otherwise.

5 Experiments

5.1 Dataset and multi-labeled tags

To evaluate the proposed MHR for video concept detection, we conducted experiments on the benchmark video corpus of TRECVID 2005. We chose English news videos in the development set as our dataset, and took the annotations of concepts defined in LSCOM (Naphade *et al*., 2006) as ground truth for the video concept detection. Let 1 denote the presence of each concept annotated for the video shots, and 0 otherwise. As a result, there were about 29 000 shots in our dataset, and we divided the dataset into three partitions: the training set (70%), the validation set (10%), and the testing set (20%).

Furthermore, in order to evaluate the effectiveness of combining multi-labeled tags into the multi-modality fusion framework, we need concepts that appear frequently in the dataset. Therefore, we also chose concept annotations of Columbia374 released by Columbia University (Yanagawa *et al*., 2007) and selected 35 concepts for our detection task, in which 17 concepts belong to the 39 concepts annotated in TRECVID 2005 and 18 other concepts annotated in Columbia374.

5.2 Evaluation metrics

We used average precision (AP) and mean average precision (MAP) to evaluate the performance of the video concept detection by our MHR method.

AP is defined as (Liu *et al*., 2008)

$$\mathrm{AP}=\frac{1}{R}\sum_{k=1}^{S}\frac{R_k}{k}\cdot I_k,$$

where S is the size of the test set, R is the number of relevant shots returned, R_k is the number of relevant shots in the top-k shots detected, and I_k=1 if the shot ranked at the kth position is relevant and 0 otherwise.

MAP is obtained by averaging the AP over all the 35 concepts.

5.3 Baseline and constructed hypergraphs

To evaluate the effectiveness of the proposed multi-modality fusion framework, we compared the MHR algorithm with the graph-based algorithms LGC and GRF, which have already been successfully applied in image and video content analysis due to their high effectiveness and efficiency (He *et al*., 2004; Yuan *et al*., 2006; Tang *et al*., 2007; Hoi and Lyu, 2008). LGC and GRF are conducted on a graph, where the vertices are labeled and unlabeled samples, and the edges reflect the similarities between sample pairs. Denote by $\boldsymbol{A}$ an affinity matrix with A_{ij} indicating the similarity between the ith and jth samples. Given two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, their similarity is estimated by

$$A_{ij}=\begin{cases}\exp\left(-\dfrac{d(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sigma}\right), & \text{if } i\neq j,\\ 0, \quad \text{else},\end{cases}$$

where $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ denotes a distance metric between samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and σ a positive radius parameter. In this experiment, we conducted LGC and GRF on visual and textual features of video shots for comparison. For visual features, we adopted Euclidean distance as the distance metric, i.e.,

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j)=\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2,$$

while for textual features, we adopted the cosine similarity, i.e.,

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j)=\cos(\boldsymbol{x}_i,\boldsymbol{x}_j)=\frac{\boldsymbol{x}_i\cdot\boldsymbol{x}_j}{|\boldsymbol{x}_i|\cdot|\boldsymbol{x}_j|},$$

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are the corresponding term vectors.

Visual-feature hypergraph $\boldsymbol{H}_{\mathrm{v}}$ and textual-feature hypergraph $\boldsymbol{H}_{\mathrm{t}}$ were constructed using the proposed CBHC algorithm. In this experiment, the numbers of hyperedges for $\boldsymbol{H}_{\mathrm{v}}$ and $\boldsymbol{H}_{\mathrm{t}}$ were 322 and 361, respectively. Hypergraph $\boldsymbol{H}_{\mathrm{c}}$ for multi-labeled tags was constructed according to the methods proposed above. Since there are 159 concepts of Columbia374 in our dataset, including the selected 35 concepts, hypergraph $\boldsymbol{H}_{\mathrm{c}}$ was constructed to have 159 hyperedges.

Experiment 1 (Evaluation of basic MHR semantic concept detection) We first compare our MHR method with the LGC algorithm. We conducted LGC on visual and textual features respectively. The results are shown in Fig. 3, where LGC-t denotes the LGC conducted on textual features, LGC-v denotes the LGC conducted on visual features, and MHR-tri denotes the MHR algorithm with triple hypergraphs $\boldsymbol{H}_t$, $\boldsymbol{H}_v$, and $\boldsymbol{H}_c$ as the input. To further investigate the effect of combining multiple labeled tags in the detection task, we also output the AP values from MHR with only $\boldsymbol{H}_t$ and $\boldsymbol{H}_v$ as the input, which is denoted by MHR-bi.

Considering AP values for each of the 35 concepts (Fig. 3), detailed analysis is given as follows:

1. Improvement of AP by our MHR algorithm was consistent for most of the 35 concepts. The improvement ratio of AP was salient for the concepts with a sparse distribution in our dataset; for example, concepts such as Computers, Flags, and Sunny were labeled in less than 200 shots. However, the improvement ratio for these concepts was very salient by MHR methods.

2. Comparing MHR-tri with MHR-bi we derive that the performance improvement through combining multiple labeled tags into our MHR framework was not so consistent for all the 35 concepts. The reason is that, the label distribution for each concept in the development and test sets is not uniform. More labeled tags are extracted from training data when there are more shots in the training set labeled by a corresponding concept. Therefore, the concept detection performance for this concept may be better.

Furthermore, the overall evaluation results, i.e., MAPs, from LGC, GRF, and our MHR algorithms are shown in Fig. 4, where GRF-t and GRF-v denote the GRF conducted on textual and visual features, respectively. Our MHR method performed better than the LGC and GRF methods, with a MAP improvement of 31.3% and 25.1% when comparing MHR-tri with LGC-v and GRF-v respectively (Fig. 4). Moreover, multi-modality fusion of only visual and textual features in our MHR framework performed even better than the LGC and GRF approaches, with a MAP improvement of 27.5% and 21.6% when comparing MHR-bi with LGC-v and GRF-v respectively, which also validates the effectiveness of the proposed CBHC approach. Furthermore, combining multiple labeled tags into our MHR framework gains a 3% improvement of MAP values, when comparing MHR-tri with MHR-bi.

Experiment 2 (Comparison of MHR with other multi-modality fusion methods) We compare MHR with other multi-modality fusion methods for video concept detection, i.e., middle fusion and late fusion.

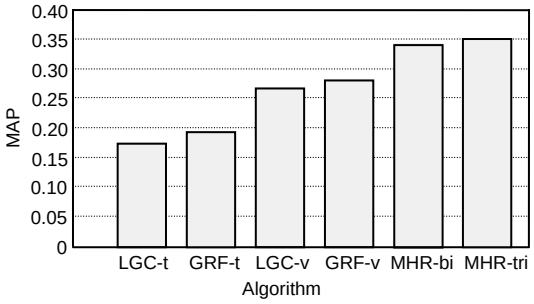

Fig. 4 MAPs from LGC-t, GRF-t, LGC-v, GRF-v, MHR-bi, and MHR-tri

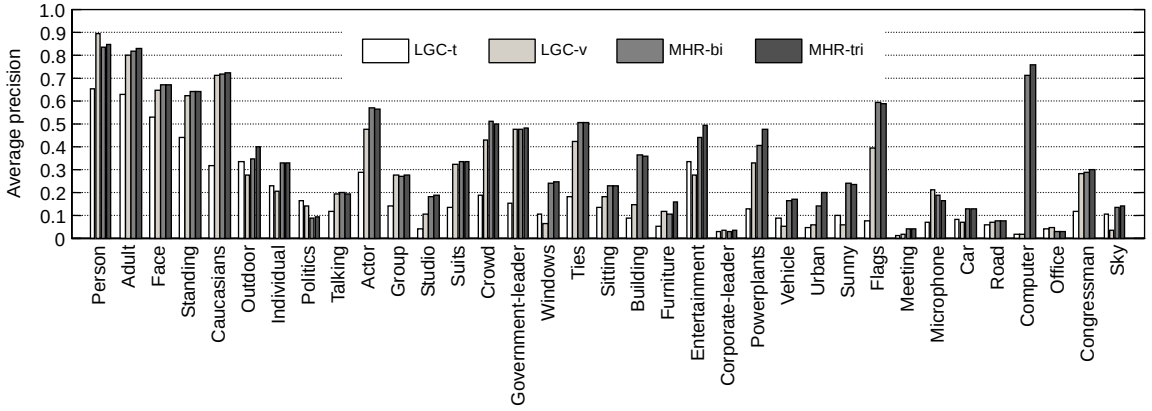

Fig. 3 Average precision for the 35 concepts from LGC-t, LGC-v, MHR-bi, and MHR-tri

LGC-t: LGC conducted on textual features; LGC-v: LGC conducted on visual features; MHR-tri: MHR algorithm with triple hypergraphs $\boldsymbol{H}_t$, $\boldsymbol{H}_v$, and $\boldsymbol{H}_c$ as the input; MHR-bi: MHR with only $\boldsymbol{H}_t$ and $\boldsymbol{H}_v$ as the input

1. Middle fusion. We chose the multi-graph fusion method by convexly combining discrete Laplacians for each graph (Wang *et al*., 2007b). Let $\boldsymbol{L}_{\mathrm{v}}$ and $\boldsymbol{L}_{\mathrm{t}}$ denote the graph Laplacians for visual and textual affinity graphs respectively. The Laplacians $\boldsymbol{L}$ for the two graphs are defined as

$$\boldsymbol{L}=\lambda\cdot\boldsymbol{L}_{\mathrm{t}}+(1-\lambda)\cdot\boldsymbol{L}_{\mathrm{v}}, \tag{19}$$

where λ is a parameter in (0, 1). Embedding the $\boldsymbol{L}$ defined in Eq. (19) into LGC and GRF algorithms respectively, we denote the middle fusion of LGC and GRF by LGC M-F and GRF M-F respectively.

2. Late fusion. We chose the linear combination of the outputs from LGC and GRF respectively on visual and textual affinity graphs as the late fusion methods. Let $\boldsymbol{I}_{\mathrm{v}}$ and $\boldsymbol{I}_{\mathrm{t}}$ denote the detection output for visual and textual affinity graphs respectively. The final output $\boldsymbol{I}$ is defined as

$$\boldsymbol{I}=\lambda\cdot\boldsymbol{I}_{\mathrm{t}}+(1-\lambda)\cdot\boldsymbol{I}_{\mathrm{v}}, \tag{20}$$

where λ is a parameter in (0, 1). We denote the late fusion of LGC and GRF by LGC L-F and GRF L-F respectively.

Fig. 5 plots the MAPs from middle and late fusion for different values of λ. The fusion results from GRF were better than those from LGC. A better performance was obtained by assigning the visual cue with a higher weight.

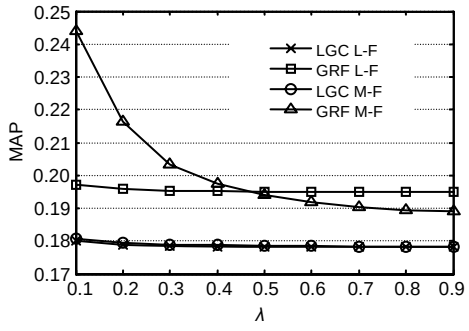

Fig. 5 MAPs from the late fusion and middle fusion methods for different values of λ

The overall comparison results (λ=0.1) (Fig. 6) show that the performance of the MHR fusion method proposed in this paper is evidently optimal.

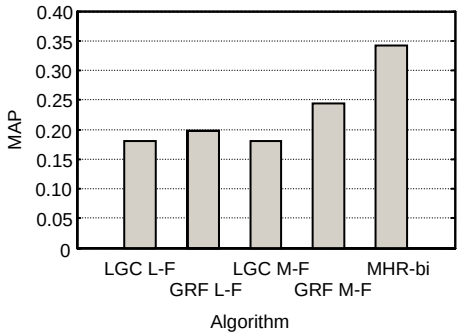

Fig. 6 MAPs from the late fusion, middle fusion methods and MHR-bi (λ=0.1)

Experiment 3 (Evaluation of the CBHC methods for hypergraph construction) Effectiveness of the construction for visual- and textual-feature hypergraphs is significant for the MHR modality-fusion framework. Association of video shots in a hyperedge provides correlations among corresponding video shots. Too many inaccurate and missing correlations may mislead the mixture Markov random walk on multiple hypergraphs so as to pull down the detection performance.

To investigate the effect of the CBHC method, we took the constructed hypergraphs $\boldsymbol{H}_{\mathrm{v}}$ and $\boldsymbol{H}_{\mathrm{t}}$ as a single input for the MHR algorithm respectively. In Fig. 7, MHR-t denotes the MHR with textual-feature hypergraph $\boldsymbol{H}_{\mathrm{t}}$ as the input, and MHR-v denotes the MHR with visual-feature hypergraph $\boldsymbol{H}_{\mathrm{v}}$ as the input. Comparing MAPs of MHR-t and MHR-v with LGC-t, LGC-v, GRF-t, and GRF-v in Fig. 4 respectively on the original dataset, we derive that, performance of MHR-t and MHR-v is better than the LGC and GRF counterpart, which further validates the effectiveness of the CBHC method. Furthermore, we conducted

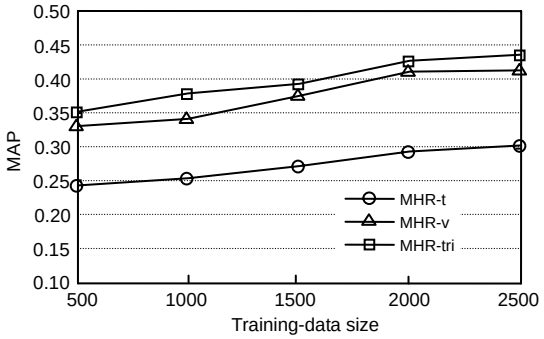

Fig. 7 MAPs from MHR-t, MHR-v, and MHR-tri by adding more training data

experiments on MHR-t and MHR-v by gradually enlarging the training set. Experimental results were also encouraging (Fig. 7). Especially, performance of MHR-v was close to that of MHR-tri, which is corresponding to the results shown in Figs. 3 and 4, in that the detection results by visual-features were better than the results by textual features.

6 Conclusions

In this paper, we propose a multiple hypergraph ranking algorithm to integrate multiple hypergraphs into a multi-modality fusion framework. With this algorithm the mixture Markov chain defined in multiple hypergraphs can produce optimal ranking scores in each hypergraph. Motivated by the multi-view learning algorithms, we represent multi-modality in videos as multiple hypergraphs. We consider visual, textual features and multiple-labeled tags and construct hypergraphs for visual and textual features by a proposed cluster-based hypergraph construction method. Experimental results show that our approach achieves performance improvement in video semantic analysis. In the future, we will integrate other modalities, e.g., temporal features, into our MHR framework, and take up such a challenge as properly associating weights with hyperedges.

References

Bickel, S., Scheffer, T., 2004. Multi-View Clustering. Proc. 4th IEEE Int. Conf. on Data Mining, p.19-26. [doi:10.1109/ICDM.2004.10095]

Dhillon, I.S., 2001. Co-clustering Documents and Words Using Bipartite Spectral Graph Partitioning. Proc. 7th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.269-274. [doi:10.1145/502512.502550]

Dumais, S.T., Furnas, G.W., Landauer, T.K., 1998. Using Latent Semantic Analysis to Improve Access to Textual Information. Proc. SIGCHI Conf. on Human Factors in Computing Systems, p.281-285.

Frey, B.J., Dueck, D., 2007. Clustering by passing messages between data points. *Science*, **315**(5814):972-976. [doi:10.1126/science.1136800]

He, J., Li, M., Zhang, H.J., Tong, H.H., Zhang, C.S., 2004. Manifold-Ranking Based Image Retrieval. Proc. 12th Annual ACM Int. Conf. on Multimedia, p.9-16. [doi:10.1145/1027527.1027531]

Hoi, S.C.H., Lyu, M.R., 2008. A multimodal and multilevel ranking scheme for large-scale video retrieval. *IEEE Trans. Multimedia*, **10**(4):607-619. [doi:10.1109/TMM.2008.921735]

Liu, J., Lai, W., Hua, X., Huang, Y., Li, S., 2007. Video Search Re-ranking via Multi-Graph Propagation. Proc. 15th Annual ACM Int. Conf. on Multimedia, p.208-217. [doi:10.1145/1291233.1291279]

Liu, Y., Wu, F., Zhuang, Y., Xiao, J., 2008. Active Post-Refined Multimodality Video Semantic Concept Detection with Tensor Representation. Proc. 16th Annual ACM Int. Conf. on Multimedia, p.91-100. [doi:10.1145/1459359.1459372]

Long, B., Yu, P.S., Zhang, Z.F., 2008. A General Model for Multiple View Unsupervised Learning. Proc. SIAM Int. Conf. on Data Mining, p.822-833.

Naphade, M., Smith, J.R., Tesic, J., Chang, S.F., Hsu, W., Kennedy, L., Hauptmann, A., Curtis, J., 2006. Large-scale concept ontology for multimedia. *IEEE Multimedia*, **13**(3):86-91. [doi:10.1109/MMUL.2006.63]

Qi, G., Hua, X.S., Rui, Y., Tang, J., Mei, T., Zhang, H.J., 2007. Correlative Multi-Label Video Annotation. Proc. 15th Annual ACM Int. Conf. on Multimedia, p.17-26. [doi:10.1145/1291233.1291245]

Spielman, D.A., Teng, S.H., 2003. Solving Sparse, Symmetric, Diagonally-Dominant Linear Systems in Time $O(m^{1.31})$. 44th Annual IEEE Symp. on Foundations of Computer Science, p.416-427. [doi:10.1109/SFCS.2003.1238215]

Sun, L., Ji, S., Ye, J., 2008. Hypergraph Spectral Learning for Multi-Label Classification. Proc. 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.668-676. [doi:10.1145/1401890.1401971]

Tan, H., Ngo, C., Wu, X., 2008. Modeling Video Hyperlinks with Hypergraph for Web Video Reranking. Proc. 16th Annual ACM Int. Conf. on Multimedia, p.659-662. [doi:10.1145/1459359.1459453]

Tang, J., Hua, X.S., Qi, G., Wang, M., Mei, T., Wu, X., 2007. Structure-Sensitive Manifold Ranking for Video Concept Detection. Proc. 15th Annual ACM Int. Conf. on Multimedia, p.852-861. [doi:10.1145/1291233.1291430]

Tong, H., He, J., Li, M., Zhang, C., Ma, W.Y., 2005. Graph Based Multi-Modality Learning. Proc. 13th Annual ACM Int. Conf. on Multimedia, p.862-871. [doi:10.1145/1101149.1101337]

Virginia, R.S., 2005. Spectral Clustering with Two Views. Proc. 22nd Int. Conf. on Machine Learning, p.20-27.

Wang, J., Zhao, Y., Wu, X., Hua, X., 2008. Transductive Multi-Label Learning for Video Concept Detection. Proc. 1st Annual ACM Int. Conf. on Multimedia Information Retrieval, p.298-304. [doi:10.1145/1460096.1460145]

Wang, M., Mei, T., Yuan, X., Song, Y., Dai, L., 2007a. Video Annotation by Graph-Based Learning with Neighborhood Similarity. Proc. 15th Annual ACM Int. Conf. on Multimedia, p.325-328. [doi:10.1145/1291233.1291303]

Wang, M., Hua, X.S., Yuan, X., Song, Y., Dai, L., 2007b. Optimizing Multi-Graph Learning: Towards a Unified Video Annotation Scheme. Proc. 15th Annual ACM Int. Conf. on Multimedia, p.862-871. [doi:10.1145/1291233.1291431]

Weng, M., Chuang, Y., 2008. Multi-Cue Fusion for Semantic Video Indexing. Proc. 16th Annual ACM Int. Conf. on

Multimedia, p.71-80. [doi:10.1145/1459359.1459370]

Yanagawa, A., Chang, S.F., Kennedy, L., Hsu, W., 2007. Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts. ADVENT Technical Report No. 222-2006-8, Columbia University, New York.

Yang, Y., Zhuang, Y., Wu, F., Pan, Y., 2008. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Trans. Multimedia*, **10**(3):437-446. [doi:10.1109/TMM.2008.917359]

Yuan, X., Hua, X.S., Wang, M., Wu, X., 2006. Manifold-Ranking Based Video Concept Detection on Large Database and Feature Pool. Proc. 14th Annual ACM Int. Conf. on Multimedia, p.623-626. [doi:10.1145/1180639.1180768]

Zha, Z., Mei, T., Wang, J., Wang, Z., Hua, X., 2009. Graph-based semi-supervised learning with multiple labels. *J. Vis. Commun. Image Represent.*, **20**(2):97-103. [doi:10.1016/j.jvcir.2008.11.009]

Zhang, H., Zhuang, Y., Wu, F., 2007. Cross-Modal Correlation Learning for Clustering on Image-Audio Dataset. Proc. 15th Annual ACM Int. Conf. on Multimedia, p. 273-276. [doi:10.1145/1291233.1291290]

Zhang, M., Zhou, Z., 2008. M^3MIML: a Maximum Margin Method for Multi-Instance Multi-Label Learning. Proc. 8th IEEE Int. Conf. on Data Mining, p.688-697. [doi:10.1109/ICDM.2008.27]

Zhao, W., Ngo, C., Tan, H., Wu, X., 2007. Near-duplicate keyframe identification with interest point marching and pattern learning. *IEEE Trans. Multimedia*, **9**(5):1037-1048. [doi:10.1109/TMM.2007.898928]

Zhou, D., Burges, C.J.C., 2007. Spectral Clustering and Transductive Learning with Multiple Views. Proc. 24th Int. Conf. on Machine Learning, p.1159-1166. [doi:10.1145/1273496.1273642]

Zhou, D., Weston, J., Gretton, A., Bousquet, O., Schölkopf, B., 2004a. Ranking on Data Manifolds. Advances in Neural Information Processing Systems 16, p.169-176.

Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B., 2004b. Learning with Local and Global Consistency. Advances in Neural Information Processing Systems 16, p.321-328.

Zhou, D., Huang, J., Schölkopf, B., 2007. Learning with Hypergraphs Clustering, Classification, and Embedding. Advances in Neural Information Processing Systems 19, p.1601-1608.

Zhu, X., Ghahramani, Z., Lafferty, J., 2003. Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. Proc. 20th Int. Conf. on Machine Learning, p.912-919. [doi:10.1109/18.850663]