Effective Multi-modality Fusion Framework For Cross-media Topic Detection

Lingyang Chu, Yanyan Zhang, Guorong Li*, Shuhui Wang,
Weigang Zhang and Qingming Huang

***Abstract*—Due to the prevalence of "We-Media", everybody quickly publishes and receives information in various forms anywhere and anytime through the Internet. The rich cross-media information carried by the multi-modal data in multiple media has a wide audience, deeply reflects the social realities and brings about much greater social impact than any single media information. Therefore, automatically detecting topics from cross-media is of great benefit for the organizations (i.e., advertising agencies, governments) that care about the social opinions. However, cross-media topic detection is challenging from following aspects: 1) the multi-modal data from different media often involve distinct characteristics; 2) topics are presented in an arbitrary manner among the noisy web data. In this paper, we propose a multi-modality fusion framework and a topic recovery approach to effectively detect topics from cross-media data. The multi-modality fusion framework flexibly incorporates the heterogeneous multi-modal data into a Multi-Modality Graph (MMG), which takes full advantage from the rich cross-media information to effectively detect topic candidates. The topic recovery (TR) approach solidly improves the entirety and purity of detected topics by: 1) merging the topic candidates that are highly relevant themes of the same real topic; 2) filtering out the less-relevant noise data in the merged topic candidates. Extensive experiments on both single-media and cross-media data sets demonstrate the promising flexibility and effectiveness of our method in detecting topics from cross media.**

***Index Terms*—We-media, topic detection, cross-media, multi-modality, fusion, topic recovery**

I. Introduction

THE rapid promotion of Web2.0 technology facilitates massive mutual interaction between real-world individuals and web contents, pushing our everyday life into the bright new era of "We-Media". One remarkable characteristic of "We-Media" is that web contents are no-longer static; instead, they interact with many real-world individuals, who frequently create diversified web contents. Individuals concerned on the same real-world event tend to create web data with similar contents. As more and more such data are created, they would gradually embody the opinions of the individuals, thus a topic with focused theme naturally emerges on the Web. These topics provide reliable insights into the public opinions, which are highly valuable to commercial companies, governments and any organization that cares about the focuses of the entire society. Therefore, effectively detecting the topics has become a novel, beneficial and urgent problem in the era of "We-Media".

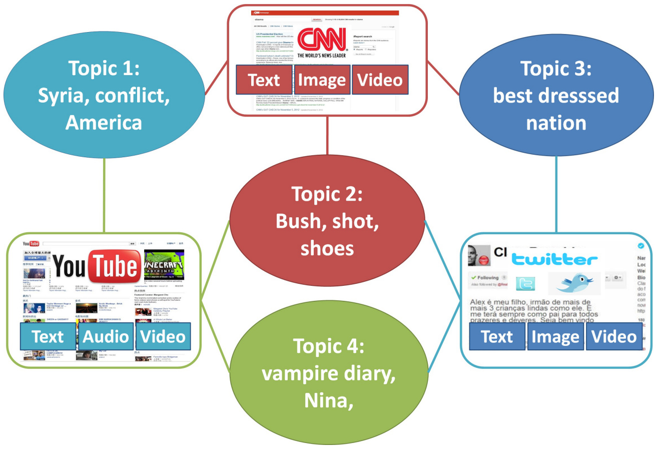


Fig. 1. Topics naturally exist in the multi-modal data from different media, such as news website (CNN), video sharing website (Youtube) and micro-blog (Twitter). The most common data modalities consist of text, image and video, where the dominant data modality of different media varies a lot.

"We-Media" refers to "we are media", which indicates that every single citizen can be a journalist to create multimedia data on the Web [1]. Different citizen-journalists have varied favorite media in collecting, reporting, analyzing and disseminating news and information. For example, Tweeter fans prefer tweeting news with short text and casually taken pictures; Youtube masters share news by short videos and roughly edited films; professional journalists often publish uniform news articles with long text and carefully selected images. As a result, huge amount of information is delivered by the multi-modal data from diversified types of media, making most topics simultaneously exist in multiple media (see Fig. 1). Compared with the limited intrinsic information capacity of a single media, the complementary cross-media information delivered by multiple media is much richer, has a broader audience group and reflects real-world events from more aspects. Therefore, it is rational and beneficial to comprehensively detect topics with the multi-modal data from multiple media,

Manuscript received in March 3, 2014, accepted in July 29, 2014.

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, in part by National Natural Science Foundation of China: 61025011, 61202322, 61332016, 61390511, 61303160 and 61303153, in part by 863 program of China: 2014AA015202, and in part by China Postdoctoral Science Foundation: 2013M530739 and 2012M520436.

* is the corresponding author. L. Chu, S. Wang and Q. Huang are with the Key Lab of Intelli. Info. Process., Inst. of Comput. Tech., CAS, Beijing 100190, China (e-mail: lychu@jdl.ac.cn, shwang@jdl.ac.cn). Y. Zhang, G. Li and Q. Huang are with the University of Chinese Academy of Sciences (CAS), Beijing 100080, China (e-mail:yyzhang@jdl.ac.cn, grli@jdl.ac.cn, qmhuang@jdl.ac.cn). W. Zhang is with the School of Computer Science and Technology, Harbin Institute of Technology,Harbin 150001, China (e-mail:wgzhang@jdl.ac.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2347551, IEEE Transactions on Circuits and Systems for Video Technology

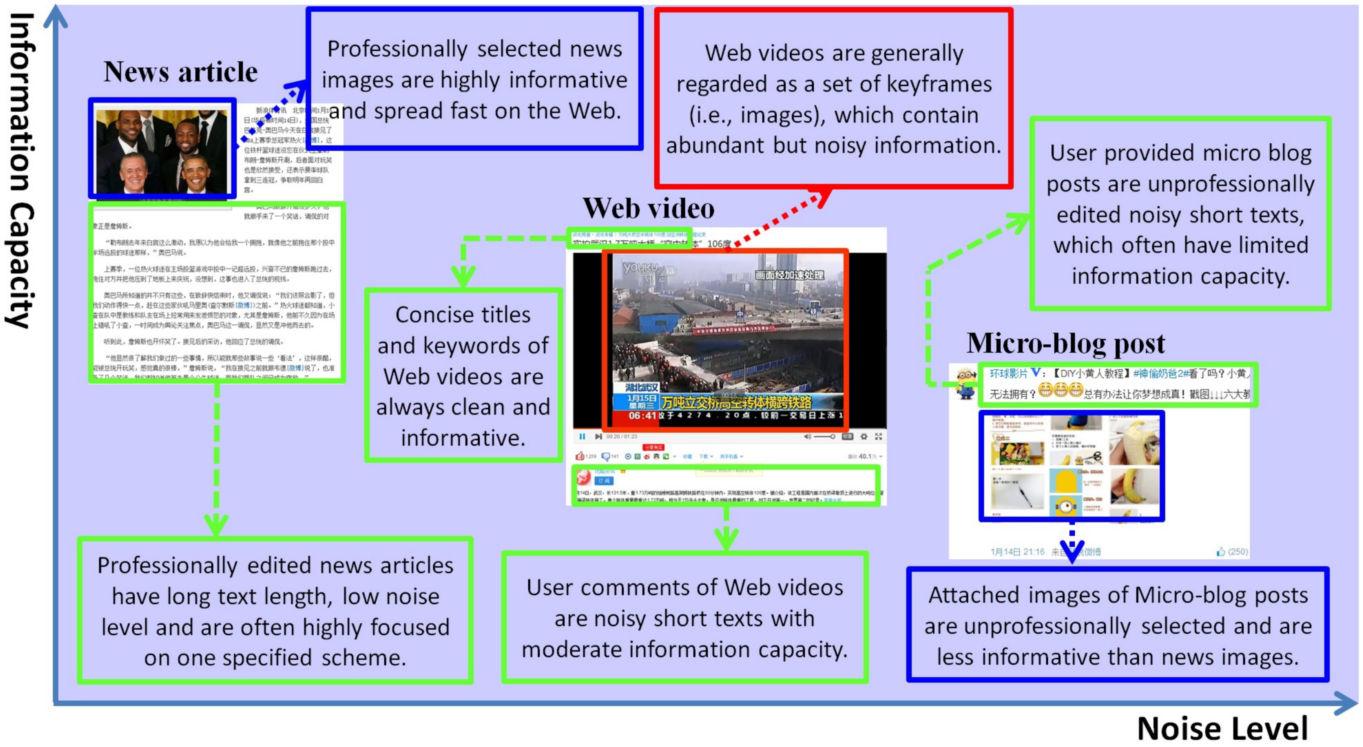

Information Capacity
Noise Level
News article
Professionally selected news images are highly informative and spread fast on the Web.
Web videos are generally regarded as a set of keyframes (i.e., images), which contain abundant but noisy information.
User provided micro blog posts are unprofessionally edited noisy short texts, which often have limited information capacity.
Web video
Concise titles and keywords of Web videos are always clean and informative.
Micro-blog post
Professionally edited news articles have long text length, low noise level and are often highly focused on one specified scheme.
User comments of Web videos are noisy short texts with moderate information capacity.
Attached images of Micro-blog posts are unprofessionally selected and are less informative than news images.

Fig. 2. The complex multi-modal data from different media (i.e., news articles, web videos and micro-blog posts) vary significantly in data structure, data modality, information capacity and noise level. Such diversification of data properties is a big challenge to comprehensively utilize the complementary cross-media information for accurate topic detection.

which is exactly the objective of cross-media topic detection.

A common solution for cross-media topic detection is to detect dense clusters of the multi-modal data with high intra cluster similarity [2]–[6]. Such dense clusters are most likely to be topic candidates, since the data of the same topic always contain similar contents [2]–[4], [6]. However, the effectiveness and robustness of cross-media topic detection are greatly challenged by the *data complexity* and the *topic diversity* as follow:

- *Data complexity:* The complexity of multi-modal data mainly attributes to the significantly varied data characteristics, such as information capacity, noise level, data modality and data structure (illustrated in Fig. 2). As a result, the rich cross-media information is difficult to be comprehensively utilized, since the data of different modalities are incomparable and most multi-modal data from multiple media are poorly structured [7].
- *Topic diversity:* Most topics consist of dense clusters of multi-modal data with high intra cluster similarity, however, the granularity, theme distribution and noise level of different topics (i.e., dense clusters) diversify a lot (illustrated in Fig. 3). Such significant *topic diversity* makes accurate topic detection a very difficult problem, since most traditional clustering approaches [8]–[10], such as k-means [8] and spectral clustering [10], cannot effectively detect the unknown number of diversified dense clusters from highly noisy multi-modal data.

In this paper, we propose an effective cross-media topic detection method to achieve impressive detection performance on two widely used data sets. The challenges of *data complexity* and *topic diversity* are effectively dealt with by the following techniques:

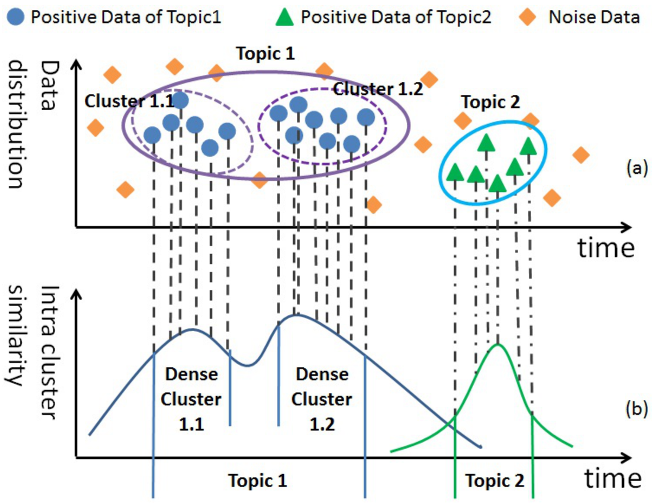


Fig. 3. Typical illustration of *topic diversity*, where Topic 1 differs a lot from Topic 2 in granularity and data distribution. The x-axes in (a) and (b) show the creation time of cross-media data. (a) shows the data distribution and (b) shows the intra cluster similarities of the dense clusters (i.e., topic candidates).

- To tackle the *data complexity* problem, a flexible multi-modality graph fusion framework is carefully designed to fuse the complex multi-modal data from different media into a multi-modality graph (MMG), where the multi-modal data of the same topic naturally form one or more dense clusters. Each graph node of MMG represents a data and the edge weight is measured by the time-embedded Jaccard similarity [11]. Due to the additivity of the Jaccard similarity, MMG is extremely flexible to incorporate different modalities by simply adding the corresponding single modality graphs together. As a result, MMG effectively leverages the complementary cross-media information to achieve significant improvement of topic detection performance. For systematic simplicity, we mainly focus on fusing the most common data modalities of text and video, which cover most of the effective multi-modal data on the Web.

- To tackle the *topic diversity* problem, a time-decay coefficient and a topic recovery approach are proposed to effectively deal with the diversified granularity and complex theme distribution of real topics. The time decay coefficient models the influence of time on the topic-similarity between multi-modal data, which makes the data from different topics more distinguishable according to their disparate time stamps. Besides, the soft quantization of time coefficient largely maintains the continuity of the time attribute, which enables our method to effectively detect topics with various granularities and prevents the unexpected segmentation of real topics caused by hard splitting the timeline [2], [3]. The topic recovery (TR) approach first joints the relevant topic candidates that are generated by multiple themes of the same real topic, then effectively filters out the noise data by their relevance with the jointed topic. This strengthens the performance of our method in detecting topics with complex theme distributions and high noise level.

II. Related Work

The Topic Detection and Tracking (TDT) task [5], [6], [12]–[19] originates from the DARPA sponsored research program [20], which aims to automatically detect and organize topics from traditional text media. However, under the impetus of the "We-Media" era, the major media form has changed from the single-modal text data to the more informative multi-modal data. This produces the need for cross-media topic detection, which comprehensively utilizes the rich multi-modal data for better topic detection performance.

A. Multi-modality Fusion

The key idea of multi-modality fusion is to effectively leverage the rich multi-modal data by robustly fusing them together. The effectiveness of multi-modality fusion has been extensively demonstrated in various tasks [21]–[23], such as topic detection [2]–[4], [7], [21], [23], [24], multimedia event detection [25]–[27], video story summarization [28]–[30], and video annotation [31].

In the topic detection task, Shao *et al.* [7] proposed a Star-structured K-partite Graph (SKG) to integrate multi-modality features for web video topic detection. However, SKG requires every data to have the same number and same type of features, which limits their generation capacity to fully use the rich non-uniform cross-media data on the Web. Cao *et al.* [2] introduced a salient trajectory method to build the tag and visual information into a topic evolution link graph for topic detection. Chen *et al.* [3] fused the dense bursty tag groups with near duplicate keyframes to detect web video topics. However, both Cao's and Chen's methods strongly rely on video tags; this limits their effectiveness in fully leveraging the rich information from various text data on the Web, such as news articles, micro blogs and user comments. Zhang *et al.* [4] proposed the Multi-Modality Graph (MMG), which effectively fuses the multi-modal information in various data forms and is able to detect topics in both single and multiple media by finding the dense subgraphs of MMG. However, MMG's performance is limited in detecting the complex real topics with multiple topic themes and high noise level, since each theme of the same complex topic would be detected as an individual topic with noise, which divides the complex real topics apart and degenerates the topic detection performance.

In the other tasks, Lan *et al.* [26] proposed a double fusion scheme to combine the former feature fusion with the latter output fusion for multimedia event detection. Tong *et al.* [32] proposed a graph-based semi-supervised learning algorithm to fuse multi-modal data by both linear and sequential fusion schemes, where the data from each modality is represented as an independent graph. Wang *et al.* [31] extended Tongs method [32] to the OMG-SSL approach that fuses multiple graphs for video annotation and person identification. OMG-SSL embeds different visual features and temporal information into a set of single modality graphs, which are further fused with the optimal fusion weights learned by a semi-supervised learning algorithm. Fu *et al.* [30] proposed a multi-graph fusion method for multi-view video summarization, where the video shots and the corresponding spatial-temporal relations are fused by a hyper graph. Zhang *et al.* [33] proposed a graph-based query specific fusion approach for image retrieval, where multiple results returned by different image retrieval methods are fused and re-ranked to achieve better retrieval performance. The fusion method of Zhang's work [33] is focused on image data, however, the advanced image retrieval performance reveals the potential effectiveness of Jaccard similarity in processing heterogeneous information. This also provides us a fundamental tool to fuse the more complex cross-media multi-modal data for topic detection. All these methods are deeply customized for their own tasks; however, the key idea of fusing multi-modal data by graphs inspires us a lot.

B. Topic Model

Probabilistic topic models are extensively used to discover the underlying structure of data. Though the "latent topic" in topic models is not exactly the same as the "real topic" in the task of topic detection, they are potentially relevant with each other from the perspective of semantic. As a result, many researchers have successfully applied topic models to detect topics from text data. Chou *et al.* [13] proposed the Incremental Probabilistic Latent Semantic Indexing (IPLSI) algorithm that captures the story line of events by maintaining the continuity of the latent semantics. He *et al.* [5] incorporated time information into a temporal Discriminative Probabilistic Model (DPM) to strengthen the topic detection performance.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2347551, IEEE Transactions on Circuits and Systems for Video Technology

Pan *et al.* [16] combined the Latent Dirichlet Allocation (LDA) [34] model with the temporal and spatial clustering for topic detection. They also extended the Spatial Latent Dirichlet Allocation [35] method to detect topics from news document collections. AlSumait *et al.* proposed an Online-LDA method to dynamically extract the thematic patterns from text streams for the identification and tracking of emerging topics. All these works focus on detecting topics from the single modal text data, where the rich information from other modalities are not used. However, their achievements in text-oriented topic detection have demonstrated the effectiveness of topic models in processing text data. These inspires us to use the Latent Dirichlet Allocation (LDA) [34] for text feature extraction.

Besides, topic models are widely used in other multimedia applications as well. Wang *et al.* [36] extended the supervised Latent Dirichlet Allocation (sLDA) to a multi-class sLDA for image classification and annotation, where the relationship between the image class labels and the image annotations is discovered by a predictive latent space. Fu *et al.* [37] proposed a multi-modal latent attribute topic model to describe user-defined and latent attributes in a semi-latent attribute space; they also proposed a scalable probabilistic topic model to learn semi-latent attributes from sparse and incomplete labels.

C. Near Duplicate Keyframe

Near duplicate keyframes (NDK) are defined as keyframes that are similar with each other in spite of the variations of viewpoint, motion, lighting and acquisition time [28]. According to the studies of Wu *et al.* [38] on video sharing web sites, the top ranked results retrieved by the same topic always contain many near duplicate videos. Xie *et al.* [39] also claimed that more than 50% of the news videos about the same real-world event contain frequently remixed and reposted near duplicate contents. The primary cause of such phenomenons is that the videos about the same topic always use near duplicate shots to convey similar information about the same real world events. Therefore, typical near duplicate video contents, such as keyframes and shots, are effective visual features to evaluate the similarity between web videos.

Many excellent works [38], [40]–[43] have been proposed to effectively detect near duplicate keyframe (NDK), which is widely used in the task of topic detection [3], [7], [23], [29], [44]. Hsu *et al.* [44]utilized NDK and other multi-modal features to track known topics across broadcasting news videos. Shao *et al.* [7] integrated NDK with other multi-modality features by a Star-structured K-partite Graph (SKG) to improve the topic detection performance. Wu *et al.* [23] presented a weighted bipartite graph to discover topic-related stories, where pairs of NDK are used as visual feature nodes to serve as visual constraints. T. Chen *et al.* [3] fused the dense bursty tag groups with NDK to efficiently detect topics from web videos. Wang *et al.* [29] introduced a event driven web video summarization approach, which localizes surrouding tags into associated video shots and identify a set of keyshots by near-duplicate keyframe detection. All these works have well demonstrated the effectiveness of NDK in topic detection, thus NDK is also adopted as a basic visual feature to measure the similarity between videos in our topic detection framework.

D. Time Attribute

Time is a natural attribute of most topics and how to effectively make use of the time attribute is an open problem widely studied by many researchers. He *et al.* [5] incorporated time information into a temporal Discriminative Probabilistic Model (DPM) to detect topics from text documents. DPM represents each document by discriminative words with time stamp and computes the posterior probability of topics when the corresponding document and time stamp are given. Cao *et al.* [2] and Chen *et al.* [3] split the time line into fragments and joint the related fragments by the temporal consistency of topics. However, hard splitting the originally continuous time line inevitably divides many topics apart, which limits the robustness of the topic detection system and degenerates the detection quality of the topics with long life time.

The time attribute is also important for many other tasks that process sequential data with temporal consistency. Wang *et al.* [31] utilized the temporal consistency of video data to improve the video annotation performance by exploring temporal consistency in a temporal-graph. Fu *et al.* [30] used the temporal consistency of videos for multi-view video summarization, which calculates the temporal similarity of video shots according to the time stamp of the corresponding keyframes. Speakman *et al.* [45] developed a dynamic pattern detection method, which allows the detected pattern of nodes to change with the incorporation of temporal consistency constraints. The achievements of these works further demonstrate the effectiveness of time attribute in the corresponding tasks.

In our topic detection framework, we propose a time decay coefficient to model the continuous influence of the time attribute on evolving topics and smoothly fuse such temporal information into the carefully designed multi-modality graph (MMG). This enables us to control the granularity of topics to detect and strengthens the topic detection performance by maintaining the continuity of the time attribute.

III. Proposed Approach

In this section, we introduce the cross-media topic detection framework (Fig. 4) based on multi-modality graph (MMG), which effectively utilizes the complementary multi-modal information by merging single modality graphs constructed from different modalities. MMG also smoothly embeds the time information into the edge weights with a carefully designed time decay coefficient. Without loss of generation, we mainly focus on the major modalities of text (denoted by T) and video (denoted by V), which are the most typical modalities widely used by many topic detection methods [2], [3], [7], [28]. A multi-modal data is denoted by $d_i = (d_i^T; d_i^V)$, where d_i^T and d_i^V represent the text and video modalities of the i-th data d_i, respectively. Since the data modalities involved by a certain medium generally do not cover all potential modalities, there are a large proportion of incomplete multi-modal data with missing data modalities. For such incomplete data, either d_i^T or d_i^V would be set as null.

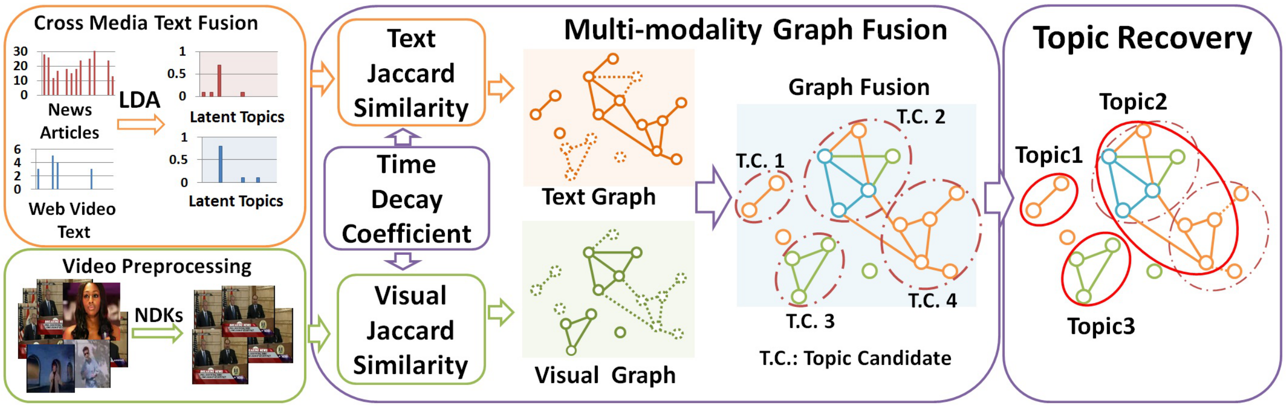


Fig. 4. The flowchart of the proposed Multi-modality graph fusion method for cross-media topic detection. In text graph and visual graph, the virtual nodes plotted in dotted lines represent the missing data of the corresponding modality. The virtual edges plotted in dotted lines represents the virtual connection of missing data. The multi-modality graph (MMG) is obtained by directly adding the single modality graphs together. The dense subgraphs of MMG are regarded as topic candidates (T.C.), which are further recovered by the topic recovery (TR) approach to detect the real topics.

A. The Text Graph

LDA-based Text Feature. Cross-media text data refers to the highly imbalanced text data from different media, which vary significantly in text length, noise level, and information capacity. For example, the professionally edited official news articles generally have long text length, contain few noise, and deliver rich information of a highly focused theme. On the contrary, the user provided text data, such as web video annotations, image tags and tweets, often consist of a few sentences (or keywords) with high noise level and convey limited information of relatively less focused theme.

Traditional text features, such as *tf-idf* histogram and keywords group, are unreliable to comprehensively measure the similarity between cross-media text data, since their descriptive power changes dramatically with the text characteristics. To describe such imbalanced text data more accurately and robustly, we propose to extract text feature by the Latent Dirichlet Allocation (LDA) [34], which is able to learn descriptive and robust latent topics from noisy and non-uniform text data on various media. LDA is a widely used topic model in natural language processing. It assumes that each document is a mixture of a number of topics and each word is generated from one of the document's topics. By modeling the proportion of topics as a Dirichlet distribution, LDA represents each topic as a multinomial distribution over words and infers the latent topic distribution of a given document from the corresponding distribution of word counts. Such latent topics are potentially related with the real-world topics, therefore we employ LDA to learn the latent topics from all the text data (e.g., news articles and video surrounding texts) and use the latent topic distribution as the topic-sensitive text feature. In this way, the text data d_i^T from any media can be uniformly represented by a L_2 normalized distribution of the latent topics:

$$d_i^T = [p_{i1}, p_{i2}, \cdots, p_{ic}, \cdots, p_{iC}] \tag{1}$$

where p_{ic} is the normalized probability of d_i^T over the *c-th* latent topic and C is the total number of latent topics for LDA. The feature-level similarity between two text features d_i^T and d_j^T is measured by the cosine similarity:

$$Sim_{ij}^T = cosine(d_i^T, d_j^T) = \frac{d_i^T \cdot d_j^T}{\|d_i^T\| \cdot \|d_j^T\|} \tag{2}$$

where $\|\cdot\|$ represents the L_2-norm, $d_i^T \cdot d_j^T$ is the inner product between d_i^T and d_j^T.

The descriptive power of the latent topic distribution (i.e., d_i^T in Eqn. 1) is more balanced for different media than traditional text features, since the basic descriptive power of latent topics is robust to the text length, noise level and information capacity of single text data, and all cross-media text data are treated equally when learning the latent topics with the LDA. The latent topic distribution also gains a stronger descriptive power in the scenario of cross-media topic detection, since the latent topics learnt by the LDA are highly related to the real-world topics from the perspective of semantic. Therefore, using the normalized distribution of latent topics as text feature is able to describe the real-world topics of text data more accurately. The latent topics can also be learnt from the text data of single media. However, comprehensively learning them from cross-media text data is a natural text fusion process, which further enhances their descriptive power by utilizing the complementary cross-media information.

Text Graph Construction. The text graph $G^T = (\{n_i^T\}, \{w_{ij}^T\})$ is the single modality graph constructed from the text data, where n_i^T represents the i-th graph node and w_{ij}^T is the edge weights between n_i^T and n_j^T. Each node n_i^T corresponds to the i-th text data (i.e., a text document) d_i^T, which is the textual part of multi-modal data d_i. The edge weight w_{ij}^T represents the fusion-level similarity between two text documents d_i^T and d_j^T, which is evaluated by the Jaccard similarity [11]:

$$w_{ij}^T = \frac{|N_i^T(k) \cap N_j^T(k)|}{|N_i^T(k) \cup N_j^T(k)|} \tag{3}$$

where $N_i^T(k)$ and $N_j^T(k)$ are the k-nearest neighbor sets of text data d_i^T and d_j^T, respectively. $|\cdot|$ is the L_0-norm that evaluates the number of elements in the nearest neighbor sets.

The multi-modal data without textual information are represented as a virtual node (Fig. 4) in the text graph and the relative edge weights are set to zero. This makes the proposed fusion framework more robust in handling the large proportion of incomplete cross-modal data with missing text data.

B. The Visual Graph

Near Duplicate Keyframes. Near-duplicate keyframes (NDK) has been widely used in large-scale news video topic detection and tracking. In our work, we detect NDK in a similar way as L. Xie *et al.* [39] and T. Chen *et al.* [3], which consists of the following steps:

1) Segment each video into video shots according to the difference of color histogram and extract the first frame (FF), the middle frame (MF) and the last frame (LF) for each shot. If the color histogram of MF is similar with either one of FF or LF, we select the MF as the keyframe. Otherwise, FF, MF and LF are all selected as keyframes.
2) Index all keyframes by the FLANN method [46], where the visual feature of color correlogram is adopted to represent each keyframe.
3) Use each keyframe as a query to search the FLANN index for the neighboring candidates, where the neighboring keyframes within a certain L_2 distance to the query keyframe are selected as the near duplicate keyframes (NDK).

Since the likelihood of two videos being about the same topic is positively correlated with the number of near duplicate keyframes between them, we directly measure the feature-level similarity between two videos by their number of near duplicate keyframes:

$$Sim_{ij}^{V} = \#\ NDK\ between\ d_i^V\ and\ d_j^V \tag{4}$$

where d_i^V and d_j^V are the video modality (i.e., video data) of the multi-modal data d_i and d_j, respectively.

Visual Graph Construction. The visual graph $G^V = (\{n_i^V\}, \{w_{ij}^V\})$ is constructed in a similar way with the text graph G^T, where n_i^V represents the i-th graph node and w_{ij}^V is the edge weight between n_i^V and n_j^V. Each node n_i^V corresponds to the i-th visual data (i.e., a web video) d_i^V, which is the visual part of the i-th multi-modal data d_i. The edge weight w_{ij}^V represents the fusion-level similarity between two web videos d_i^V and d_j^V, which is evaluated by the Jaccard similarity [11]:

$$w_{ij}^V = \frac{|N_i^V(k) \cap N_j^V(k)|}{|N_i^V(k) \cup N_j^V(k)|} \tag{5}$$

where $N_i^V(k)$ and $N_j^V(k)$ are the k-nearest neighbor sets of web videos d_i^V and d_j^V, respectively. $|\cdot|$ is the L_0-norm that evaluates the number of elements in the nearest neighbor sets.

Each multi-modal data without visual information is represented as a virtual node (see Fig. 4) in the visual graph and the relative edge weights are all set to zero. This makes the proposed fusion framework more robust in dealing with the large proportion of incomplete cross-modal data with missing visual data.

C. The Multi-modality Graph

Time Decay Coefficient. The time decay coefficient is proposed to smoothly measure the temporal similarity of two data by the interval between their upload times. In this way, the data from different topics can be more distinguishable according to their disparate time stamps. Given two multi-modal data d_i and d_j, the time decay coefficient is obtained by:

$$\alpha_{ij} = e^{-\beta\left(\lfloor \frac{|t_i - t_j|}{\delta} \rfloor\right)^2} \tag{6}$$

where β is a positive scale parameter to control the rate of decay, δ is a small fixed unit time factor and t_i, t_j are the upload times of d_i and d_j. Note that, $\lfloor\cdot\rfloor$ denotes the round down operation.

Apparently, when the time interval $|t_i - t_j|$ increases, the time-decay coefficient decreases exponentially, which further indicates that d_i and d_j are less likely to be about the same topic. As a result, the time decay coefficient is able to properly model the influence of time on evolving topics, where the probability that two data are about the same topic drops exponentially with their time interval. Meanwhile, the soft quantization strategy of the proposed time decay coefficient largely maintains the continuity of the time attribute. This enables our topic detection system to effectively detect topics with various granularities and prevents the unexpected segmentation of real topics caused by hard splitting the timeline [2], [3].

Multi-modality Graph Fusion. After obtaining the text graph $G^T = (\{n_i^T\}, \{w_{ij}^T\})$ and the visual graph $G^V = (\{n_i^V\}, \{w_{ij}^V\})$, we fuse them into the Multi-Modality Graph $G = (\{n_i\}, \{w_{ij}\})$, where the node set is obtained by:

$$\{n_i\} = \{n_i^T\} \cup \{n_i^V\} \tag{7}$$

and the edge weight w_{ij} is obtained by:

$$w_{ij} = \alpha_{ij}(w_{ij}^T + w_{ij}^V) \tag{8}$$

In this way, the single-modality nodes n_i^T and n_i^V, which correspond to the text and video modalities of the same multi-modal data d_i, are fused into one Multi-Modality Graph (MMG) node n_i. Other single-modality nodes, which correspond to the incomplete data missing either text or visual modalities, are directly transformed to MMG nodes without fusion. All nodes in MMG are treated equally.

For the fusion of edge weights w_{ij}^T and w_{ij}^V, although the feature-level similarity of text data is not directly comparable with the feature-level similarity of video data, the corresponding fusion-level Jaccard similarities are comparable, since both of them reflect the consistency of two k-nearest neighborhood sets. Considering that there is no prior about the relative importance of each modality, a proper solution is to treat all modalities equally by simply summing up the edge weights (see Eqn. 8). The time information is also embedded into the fused edge weights of MMG to make the graph nodes (i.e., multi-modal data) from different topics more distinguishable according to their time stamps. Apparently, this fusion framework of MMG is flexible enough to robustly incorporate the multi-modal data in different media.

D. Topic Candidate Detection and Topic Recovery

Topic Candidate Detection. The edge weights $\{w_{ij}\}$ of the Multi-Modality Graph (MMG) jointly evaluate both the upload time similarities and the content similarities of multi-modal data. Since data about the same theme of a topic are generally similar with each other in both content and upload time, the corresponding MMG nodes would be strongly connected with each other and naturally form a dense subgraph. Such dense subgraph is a topic-sensitive pattern, which is robust to noise and can be be effectively detected by pair-wise clustering methods. Therefore, we can transform the cross-media topic detection problem into a dense subgraph detection problem on MMG, where each dense subgraph of MMG is regarded as a meaningful *topic candidate*.

The dense subgraph seeking problem is well studied in previous literatures [47]–[50]. An arbitrary subgraph of MMG is represented by a *probabilistic cluster* $x \in \triangle^m$, where $\triangle^m = \{x \mid x \in R^m, x \geq 0, \sum_{i=1}^m x_i = 1\}$ is the standard simplex and m is the total number of graph nodes in MMG. In fact, x is a unit indicator vector, which maps a cluster of graph nodes to the standard simplex $\triangle^m$. The i-th component of x is denoted by x_i, which is the probability that the subgraph x contains the MMG node n_i. Therefore, the index set of all the non-zero components of x (i.e., $\{i \mid x_i > 0, i \in [1,m]\}$) identifies the set of all nodes contained by the subgraph x. Particularly, $x_i = 0$ means that n_i is not contained by subgraph x, and the subgraph containing a single node n_i can be represented by $x = e^i$, where e^i is the i-th column of the identity matrix.

Let W be the affinity matrix that stores the edge weights of MMG (i.e., $W(i,j) = w_{ij}$), then the average connection strength between all nodes of subgraph x can be measured by:

$$g(x) = x^T W x = \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} x_i x_j \tag{9}$$

According to Pavan *et al.* [48], each dense subgraph of MMG uniquely corresponds to a local maximum point x^* of $g(x)$, which identifies a topic candidate and can be easily obtained by solving the standard quadratic optimization problem (StQP) as follow:

$$x^* = \max_{x} \ g(x), \quad s.t. \ x \in \triangle^m \tag{10}$$

There are many well designed mathematical tools to solve the StQP problem in Eqn. 10, such as graph shift [47], dominant set method [48], hierarchical dominant set method [49] and infection-immunization dynamics [50]. We adopt the graph shift method [47] for its advanced efficiency.

Topic Recovery. The quality of the topic candidates detected by seeking dense subgraphs on the Multi-Modality Graph (MMG) are often challenged by the large proportion of real topics with complex theme distribution and high noise level. Such topics generally involve multiple themes, which are often detected as multiple independent dense subgraphs (i.e., topic candidates) by the graph shift method [47]. This divides many real topics apart and lowers the integrity of the detected topics. Besides, the confusing noise data also decreases the purity of detected topics. As a result, we propose the topic recovery (TR) approach, which increases the integrity of topics by merging the relevant topic candidates together and improves the topic purity by filtering out the less relevant noise data.

Algorithm 1: The merging step of topic recovery

```
Input: The set of all detected topic candidates
       $X = \{x^*\}$.
Output: The set of merged topic candidates $Z = \{z^*\}$.
Set $Z = \emptyset$.
repeat
    Select the most significant topic $x^* \in X$ with the
    maximum value of $g(x^*)$ (see Eqn. 9).
    Find all topic candidates $Y = \{y^*\} \subset X$ satisfying
    $Rel(x^*, y^*) > \theta$ (see Eqn. 11), where $\theta$ is the
    integrity threshold.
    Obtain the merged topic candidate $z^*$ by merging $x^*$
    with all topic candidates in $Y$ (see Eqn. 13).
    Remove $x^*$ and $Y$ from $X$ and add $z^*$ into $Z$.
until $X = \emptyset$.
return $Z = \{z^*\}$
```

Algorithm 2: The filtering step of topic recovery

```
Input: A merged topic candidate $z^* \in Z$.
Output: The multi-modal data set $T$ of the final topic.
Set $T = \emptyset$.
repeat
    Select $d_i$ with the largest component $z_i^*$ of $z^*$.
    Compute $L = \{d_j \mid w_{ij} > \eta, d_j \in T\}$, where $\eta$ is the
    purity threshold and $w_{ij} \in W$ measures the relevance
    level between multi-modal data $d_i$ and $d_j$.
    if $|L|/|T| \geq 0.5$ then
        Add $d_i$ into $T$ and set $z_i^*$ to zero.
    else
        Break.
    end if
until $\sum_{i=1}^m z_i^* = 0$
return $T$
```

The *merging step* of TR joints the relevant topic candidates that corresponds to multiple themes of the same topic. The rationality of this step lies in the fact that the relevance level between most themes of the same topic are much higher than the themes from different topics. Therefore, the topic candidates that are actually the themes of the same real topic can be effectively identified according to their high relevance level. We propose to measure the relevance between two topic candidates (i.e., dense subgraphs) x^* and y^* by:

$$Rel(x^*, y^*) = (x^*)^T W y^* = \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} x_i^* y_j^* \tag{11}$$

which is the average connection strength between the graph nodes (i.e., multi-modal data) of the corresponding dense subgraphs (i.e., topic candidates) x^* and y^*. If the relevance level between x^* and y^* are higher than a fixed entirety threshold θ (i.e., $Rel(x^*, y^*) > \theta$), they are regarded as the themes of the same real topic, hence they are merged together by the weighted average as follow:

$$z^* = \frac{x^* \cdot |x^*| + y^* \cdot |y^*|}{|x^*| + |y^*|}. \tag{12}$$

where $|\cdot|$ represents the L_0-norm, $|x^*|$ ($|y^*|$) represents the number of graph nodes contained by the dense subgraph x^* ($|y^*|$) and $z^* \in \triangle^m$ is the merged topic candidate. The weighed average in Eqn. 12 can be easily extended to merge x^* with a set of relevant topic candidates $Y = \{y^*\}$ by:

$$z^* = \frac{x^* \cdot |x^*| + \sum_i (Y_i \cdot |Y_i|)}{|x^*| + \sum_i |Y_i|} \tag{13}$$

where Y_i represents the i-th topic candidate y^* in Y.

Details about the merging step of topic recovery are summarized in Algorithm 1, where all highly relevant topic candidates in $X = \{x^*\}$ are merged together to strengthen the integrity of merged topic candidates in $Z = \{z^*\}$. Note that, each merged topic candidate $z^* \in Z$ corresponds to a merged subgraph in the multi-modality graph.

The *filtering step* of TR effectively filters out the noise data that are less relevant with the merged topic candidate $z^* \in Z$. Since the positive data are generally more relevant with the merged topic candidate than the noise data, we can easily filter out the noise data by their low relevance with the topic candidate.

Given a merged topic candidate $z^* \in Z$, Algorithm 2 iteratively grows the final topic T by selecting the multi-modal data d_i, whose relevance level with more than 50% of the multi-modal data in current T are larger than a fixed purity threshold η. Note that, the relevance level between multi-modal data d_i and d_j is directly measured by the affinity value $w_{ij} \in A$, which is also the edge weight between graph nodes n_i and n_j in MMG.

IV. Experiment

To demonstrate the effectiveness of the proposed multi-modality graph with topic recovery (MMG+TR), we compare the topic detection performances with the salient trajectory method (ST) [2], the tag group method (TG) [3] and the multi-modality graph (MMG) [4] without topic recovery. For both MMG and MMG+TR, the NDKs are extracted by the method proposed in [39] and the latent topics are generated by the "topic modeling toolbox" published by L. Thomas *et al.* [51]. All experiments are conducted on a common PC with Core i-5 CPU and 12 GB memory.

A. Data Sets and Evaluation Criteria

Data Sets. The topic detection performances of compared methods are analyzed on two standard data sets: the core data set of MCG-WEBV [52] and YKS [4]. The core data set of MCG-WEBV is published by J. Cao *et al.* [52] in 2009. It is a widely used single-media data set for web video topic detection [2]–[4], [7]. The data set is built with every day's most viewed videos and their surrounding texts (e.g. titles, tags, descriptions) on "www.youtube.com" from December 2008 to February 2009. It contains 3,660 web videos and 73 manually annotated ground truth topics, and the average topic-duration is 42.2 days. For notational compactness, we refer to the "core data set of MCG-WEBV" as MCG-WEBV throughout the paper. The YKS data set is a cross-media data set for topic detection published by Y. Zhang *et al.* [4] in 2013. YKS contains 2,131 web videos from "www.youku.com" (YouKu) and 7,325 news articles from "www.sina.com.cn" (Sina). All the data of YKS are crawled from May 2012 to June 2012. The ground truth contains 20 pure web video topics, 225 pure news article topics and 73 hybrid topics, which involve both the two media of web video sharing site (i.e., YouKu) and official news website (i.e., Sina). The average topic-duration is 13.0 days.

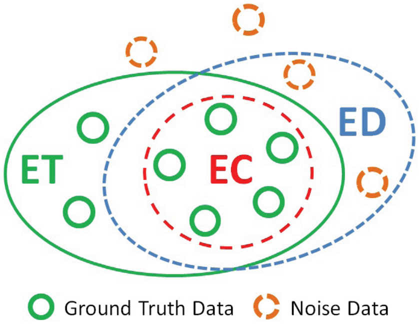


Fig. 5. An illustration of ET, ED and EC. Each multi-modal data is represented as a node. ED is the data set of a detected topic, which is marked by the blue dashed ellipse. ET is the data set of ground truth topic (in the green ellipse) that matches the best with the detected topic ED. EC is the set of correctly detected data (in the red dashed circle), which is the intersection of ED and ET.

Evaluation Criteria. We adopt the same evaluation method with J. Cao *et al.* [2] and T. Chen *et al.* [3], where the standard evaluation criteria, such as precision, recall and F-Measure, are used to evaluate the topic detection performance. The precision and recall are defined as:

$$Precision = \frac{|EC|}{|ED|} \quad Recall = \frac{|EC|}{|ET|} \tag{14}$$

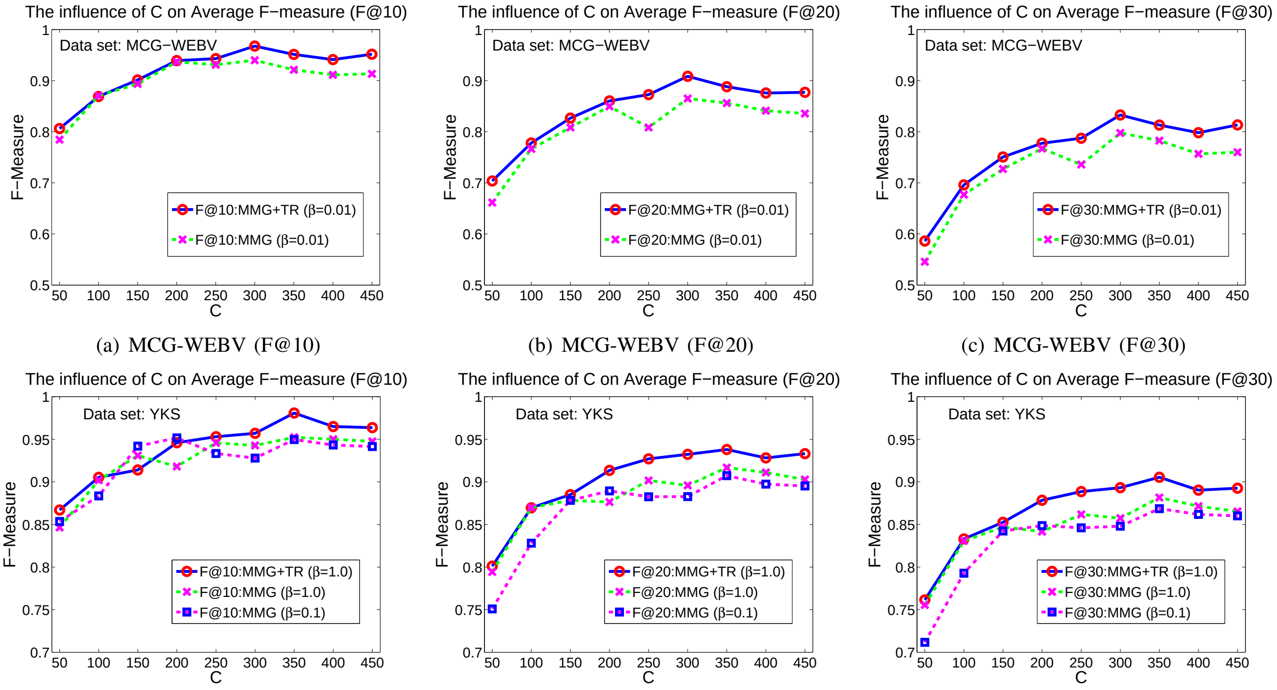
where $|ED|$ is the number of multi-modal data in the detected topic, $|ET|$ is the number of data in the ground truth topic best matched with the detected topic, and $|EC|$ is the number of correctly detected data (see Fig. 5). After obtaining the precision and recall for each topic, we calculate the F-Measure by:

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{15}$$

which is a comprehensive evaluation on both precision and recall.

Similar with Cao *et al.* [2], we sort all detected topics by their F-measure performance and evaluate the overall performance of a topic detection approach by the average of precision, recall and F-measure on the top-N detected topics ($N = [10, 20, 30]$). Specifically, the average precision, recall and F-measure performances for the Top-N detected topics are denoted as "P@N", "R@N" and "F@N", respectively. Additionally, we also adopt the CP measurement proposed by T. Chen *et al.* [3] to evaluate the percentage of correctly

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2347551, IEEE Transactions on Circuits and Systems for Video Technology



(a) MCG-WEBV (F@10) (b) MCG-WEBV (F@20) (c) MCG-WEBV (F@30)

(d) YKS (F@10) (e) YKS (F@20) (f) YKS (F@30)

Fig. 6. The topic detection performances of MMG and MMG+TR on two standard data sets of MCG-WEBV and YKS. The TR parameters are set as $[\theta = 0.65, \eta = 0.5]$ for MCG-WEBV and $[\theta = 0.45, \eta = 0.6]$ for YKS. (a)∼(c) show the average F-measure performances of F@10, F@20 and F@30 on MCG-WEBV, respectively. (d)∼(f) show the average F-measure performances of F@10, F@20 and F@30 on YKS, respectively.

detected topics. CP is defined as:

$$CP = \frac{NCDT}{NDT} = \frac{\#\ Correctly\ Detected\ Topics}{\#\ Detected\ Topics} \qquad (16)$$

where NDT is the total number of detected topics, and NCDT is the number of correctly detected topics. A detected topic is regarded as correct if its F-Measure is bigger than 0.5.

B. Parameter Analysis

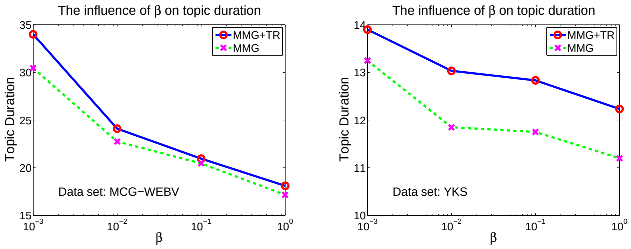
In this section, we analyze how each parameter affects the topic detection performances of both MMG and MMG+TR. The number of nearest neighbors k (see Eqn. 3 and Eqn. 5) and the unit time factor δ (see Eqn. 6) do not have much influences on the final result, thus they are fixed as $k = 30$, $\delta = 3$ for both MMG and MMG+TR. All experimental results are obtained with optimal parameter settings by default.

The number of latent topics C. The latent topics are learnt by the Topic Modeling Toolbox published by L. Griffiths *et al.* [51] with default settings, where the number of iterations is set to 1000. For each data set, we use all the documents of the text collections to learn the latent topics of LDA. Specifically, for the MCG-WEBV data set, we use the surrounding texts of all 3,660 videos to learn the latent topics; for the YKS data set, we use the surrounding texts of all 2,131 videos and the 7,325 news documents. The average document sizes of MCG-WEBV and YKS are 33 and 429, respectively.

The number of latent topics C decides the descriptive power of the text feature (Eqn. 1), which affects the edge weights of the multi-modality graph and influences the topic detection performances of both MMG and MMG+TR. Fig. 6 shows the influence of C on the average F-Measure of MMG and MMG+TR. On both MCG-WEBV and YKS, the F-measure performances (i.e., F@10,F@20 and F@30) first increase with the growth of C due to the increasing descriptive power of the text feature. However, due to the intrinsic descriptive power bottleneck of the latent topics, the descriptive power of text feature cannot increase infinitely with the growth of C. Therefore, the F-measure performances stabilize at an optimal level when C becomes large. Besides, we can also see from Fig. 6 that MMG+TR generally outperforms MMG on both data sets, which demonstrates the effectiveness of the topic recovery approach. According to the experimental results, the optimal value of C for both MMG and MMG+TR are set as $C = 300$ and $C = 350$ for MCG-WEBV and YKS, respectively.

The scale parameter β for time decay coefficient. The scale parameter β controls the decay rate of the time decay coefficient (Eqn. 6). This parameter is designed to control the interested granularity of the topics to be detected. A small β strengthens the edge weights between MMG nodes, which increases the chance of the positive nodes from large granularity topics to form a dense subgraph. On the contrary, a big β weakens the MMG edge weights, which favors the detection performances of small granularity topics. This phenomenon is shown in Fig. 7, where the average granularity (i.e., topic duration) of top-20 detected topics decreases with the growth of β on both MCG-WEBV and YKS.

Different values of β focus MMG+TR on detecting the topics with different ranges of granularity. From this perspective, each value of β is optimal for its own focused range of topic granularity. However, from the perspective of overall topic detection performance, there is a trade-off between the topics



(a) MCG-WEBV (b) YKS

Fig. 7. The influence of β on the duration of topics detected by MMG and MMG+TR. (a) shows the results on MCG-WEBV. (b) shows the results on YKS. The x-axis is plotted in log scale. Since the proposed topic recovery (TR) approach merges multiple topic candidates together, the average duration of the topics detected by MMG+TR is relatively larger than MMG.

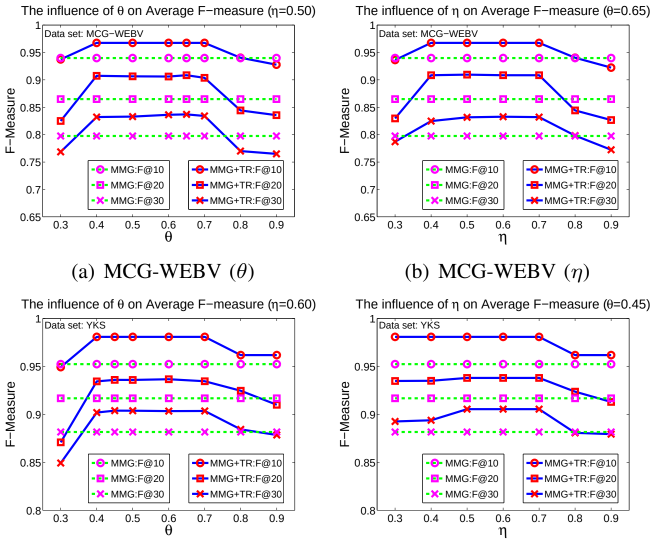
TABLE I
THE INFLUENCE OF β ON THE F@10 PERFORMANCES OF MMG AND MMG+TR. FOR MCG-WEBV: $[C = 300, \theta = 0.65, \eta = 0.5]$. FOR YKS: $[C = 350, \theta = 0.45, \eta = 0.6]$.

β	YKS		MCG-WEBV	
	MMG	MMG+TR	MMG	MMG+TR
0.001	0.9046	0.9265	0.9302	0.9508
0.01	0.9258	0.9373	**0.9398**	**0.9675**
0.1	0.9497	0.9766	0.9165	0.9294
1.0	**0.9511**	**0.9808**	0.8179	0.8434

with large and small granularity. Table I shows the influence of β on F@10 of MMG and MMG+TR on MCG-WEBV and YKS. For both MMG and MMG+TR, the optimal value of β are set as $\beta = 0.01$ and $\beta = 1.0$ on MCG-WEBV and YKS, respectively. Besides, we can also see that the optimal value of β on MCG-WEBV is smaller than the optimal β on YKS. This is because that the average topic granularity of MCG-WEBV is larger than YKS, which requires a relatively smaller β to focus on detecting the topics with large granularity.

The integrity parameter θ. The integrity parameter θ is the linkage threshold between two highly relevant topic candidates detected by MMG. Since MMG may divide different themes of a complex real topic as multiple topic candidates, linking such topic candidates increases the integrity of the detected topics, which improves the topic detection performance. On the other hand, the less relevant topic candidates from different real topics are not linked together, since their relevance level is generally much lower than the highly-relevant topic candidates. As shown in Fig. 8 (a) and Fig. 8 (c), the average F-measure performances on both MCG-WEBV and YKS are all stable at the optimal level for a wide range of θ; this is a natural result due to the large relevance-level margin between the highly relevant topic candidates and the less relevant ones. Such property of θ provides us a safe window that strengthens the robustness of the proposed topic detection method. As a result, the integrity parameter θ is optimally set as $\theta = 0.65$ and $\theta = 0.45$ for MCG-WEBV and YKS, respectively.

The purity parameter η. The purity parameter η is used to filter out the noise multi-modal data that are less relevant with the merged topic candidates. Filtering such noise data increases the purity of detected topics and improves the topic detection performance. As shown in Fig. 8 (b) and Fig. 8 (d), the average F-measure performances on both MCG-WEBV and YKS are stable at the optimal level for a wide range of η.



(a) MCG-WEBV (θ) (b) MCG-WEBV (η)

(c) YKS (θ) (d) YKS (η)

Fig. 8. The influence of θ and η on the average F-measure performance of MMG+TR. (a) and (b) show the results on MCG-WEBV with $C = 300$ and $\beta = 0.01$. (c) and (d) show the results on YKS with $C = 350$ and $\beta = 1.0$. The optimal F-measure performances of MMG are plotted as green dashed lines in each figure for clear comparison.

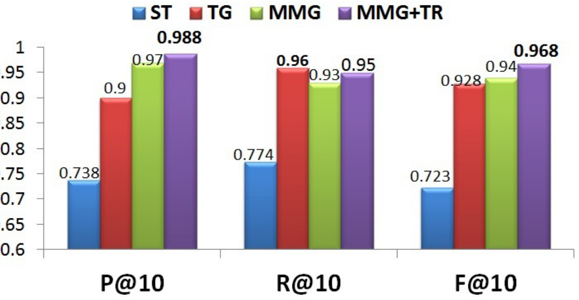


Fig. 9. The average precision, recall and F-Measure of the top-10 detected topics on MCG-WEBV.

This is a reasonable result, since the positive data that belongs to a real topic generally have a much higher relevance level with the merged topic candidate than the noise data, and the relevance-level margin is large. According to the experimental result, the purity parameter η is optimally set as $\eta = 0.5$ and $\eta = 0.6$ for MCG-WEBV and YKS, respectively.

C. Experiments on MCG-WEBV

In this section, we compare the topic detection performances between the salient trajectory method (ST) [2], the tag group method (TG) [3], MMG [4] and MMG+TR. The single-media data set MCG-WEBV is used to evaluate the topic detection performances, where all methods are compared with their own optimal parameters. For both MMG and MMG+TR, the parameters are optimally set as $[C = 300, \beta = 0.01]$ and the TR parameters are set as $[\theta = 0.65, \eta = 0.5]$.

Fig. 9 shows the comparison results of P@10, R@10 and F@10. We can see that the F@10 performances of both TG and MMG outperform ST, while MMG+TR achieves the best F-measure performance. The experimental results in Table II

TABLE II
THE TOPIC DETECTION PERFORMANCES OF MMG+TR AND MMG ON MCG-WEBV.

Evaluation	P@20	R@20	F@20	P@30	R@30	F@30
MMG+TR	0.9381	0.8898	**0.9085**	0.8835	0.8120	**0.8327**
MMG	0.9079	0.8410	0.8650	0.8969	0.7462	0.7975

TABLE III
THE COMPARISON RESULTS OF CP ON MCG-WEBV.

Data set	Method	NDT	NCDT	CP
MCG-WEBV	TG	83	31	37.35%
	MMG	57	38	66.67%
	MMG+TR	50	**45**	**90.00%**

further demonstrate the effectiveness of the topic recovery approach, where MMG+TR generally outperforms MMG.

Table III shows the experimental results under the evaluation criteria of CP. As it is shown, the CP performances of MMG+TR and MMG significantly outperform TG, where both MMG+TR and MMG detect more topics than TG. This is attributed to the robustness and topic-sensitive property of the naturally formed dense subgraphs in the multi-modality graph. Besides, since TR merges the topic candidates detected by MMG, the NDT of MMG+TR is smaller than MMG. However, TR effectively increases entirety and purity of the detected topics, which improves the quality of detected topics and increases the NCDT of MMG+TR over MMG. In sum, we can conclude from Table III that TR effectively improves the CP performance of MMG by merging the highly relevant topic candidates and filtering out the noise data.

D. Experiments on YKS

In this section, we analyze the topic detection performances of TG, MMG and MMG+TR on the cross-media data set of YKS. The source code of TG is kindly provided by T. Chen [3]. Since the video topic detection approach TG is only able to process the web video data in YKS, we fairly compare with it by running MMG and MMG+TR on exactly the same web video data set of YKS, which is referred to as "YKS-V". The corresponding topic detection performances of MMG and MMG+TR on YKS-V are referred to as MMG-V and MMG+TR-V. Additionally, the topic detection performances of MMG and MMG+TR are further compared using the full cross-media data set of YKS. All methods are compared with their own optimal parameters:

1) On the web video data set YKS-V:
   - TG : $\theta_1 = 0.2$; $\beta_1 = -0.25$ and $\eta_1 = 0.43$, which are the native parameters of TG [3].
   - MMG-V: $\beta = 1$, $C = 200$.
   - MMG+TR-V: $\beta = 0.1$, $C = 250$, $\theta = 0.35$, $\eta = 0.55$.
2) On the full cross-media data set YKS:
   - MMG: $\beta = 1.0$, $C = 350$.
   - MMG+TR: $\beta = 1.0$, $C = 350$, $\theta = 0.45$, $\eta = 0.6$.

Fig. 10 shows the topic detection performances of TG, MMG-V and MMG+TR-V. We can see that both MMG-V and MMG+TR-V achieve better detection performances than TG, which demonstrates the effectiveness of the proposed graph fusion framework. Furthermore, MMG+TR-V significantly outperforms MMG-V from all aspects, which demonstrates the effectiveness of the topic recovery (TR) approach. Table IV shows the experimental results on the full cross-media data set of YKS, which further demonstrate the effectiveness of TR in improving the topic detection performance.

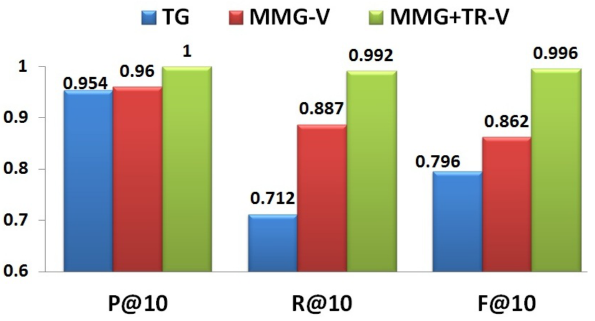


Fig. 10. The average precision, recall and F-Measure of the top-10 detected topics on YKS-V.

TABLE IV
THE TOPIC DETECTION PERFORMANCES OF MMG+TR AND MMG ON YKS.

Method	MMG+TR	MMG
P@10	1.0000	0.9757
R@10	0.9639	0.9288
F@10	**0.9808**	0.9511
P@20	0.9747	0.9455
R@20	0.9126	0.8914
F@20	**0.9380**	0.9168
P@30	0.9603	0.9301
R@30	0.8704	0.8580
F@30	**0.9054**	0.8816

Table V shows the CP performances of all compared methods on YKS-V and YKS, where both MMG and MMG+TR significantly outperform TG. Similar with the experimental results on MCG-WEBV (see Table.III), the CP performances on both YKS-V and YKS further demonstrate the effectiveness of TR in improving the topic detection performances.

Table VI shows details about the detected topics, where the correctly detected topics are divided into three groups according to their modalities. The "Articles" refers to the topics consisting of pure news articles; the "Videos" refers to the topics only containing web videos; the "Hybrids" are the topics that consist of both news articles and web videos. "\" means the methods are not suitable to detect this type of topics. Apparently, MMG+TR-V and MMG-V detect more topics than TG in both the groups of "Videos" and "Hybrids", which proves the effectiveness of our method in dealing with the multi-modal data from the single media of web video. Moreover, MMG+TR and MMG not only detect more video-related topics than MMG+TR-V and MMG-V, but also detect many pure News article topics; this shows the significant performance enhancement brought by the complementary cross-media information and the effectiveness of the proposed multi-modality graph fusion framework in dealing with the incomplete multi-modal data from different media.

TABLE V
THE COMPARISON RESULTS OF CP ON YKS-V AND YKS.

Data set	Method	NDT	NCDT	CP
YKS-V	TG	58	11	18.97%
	MMG-V	57	23	40.35%
	MMG+TR-V	48	**30**	**62.50%**
YKS	MMG	276	117	42.39%
	MMG+TR	233	**124**	**53.21%**

TABLE VI
THE COMPONENTS OF CORRECTLY DETECTED TOPICS ON YKS-V AND YKS.

Data set	Method	Articles	Videos	Hybrids	Total
YKS-V	TG	\	7	4	11
	MMG-V	\	13	10	23
	MMG+TR-V	\	**14**	**16**	**30**
YKS	MMG	69	**18**	**30**	117
	MMG+TR	**93**	7	24	**124**

E. Typical Results

To provide a more intuitive demonstration of the detected topics, we presents some examples of the detected topics with representative keyframes and key words. The representative keyframes are selected as the near duplicated keyframes (NDK) with the largest shared times. The representative key words are extracted by voting the latent topics (learnt by LDA [34]) with the text features of all data in the detected topic. The significant keywords from the top-3 most voted latent topics are then selected as the representative key words. Considering that MCG-WEBV is a public data set, we present the detected topics on it. Table VII lists the top-12 topics detected by MMG+TR on MCG-WEBV with their IDs and manually labeled titles [52]. We also sample 3 typical examples from the top-12 detected topics and present the corresponding representative key words and keyframes in TableVIII. As it is shown, the extracted keyframes and key words are highly relevant to the corresponding topics.

V. CONCLUSION

We present a cross-media topic detection system with three key techniques to robustly detect topics from the multi-modal data in multiple media. The multi-modality graph is proposed to efficiently fuse the heterogenous cross-media data together and optimize the usage of rich multi-modal information. The time decay coefficient models the time characteristics of data and makes it controllable to detect topics with different time span. The topic recovery approach merges the falsely segmented topics and filters out the noise data to improve the entirety and purity of detected topic candidates. As demonstrated by extensive experiments, the fusion of cross media data leads to significant information gain, which largely improves the topic detection performance. The proposed cross-media detection approach is robust in detecting topics from the multi-modal data in different media. In our future work, we will extend our method to modeling the evolution trend of topics under streaming data.

ACKNOWLEDGEMENT

The authors would like to thank Tianlong Chen for the support of code and his kindness help.

TABLE VII
TOP-12 HOT TOPICS AMONG THE 45 CORRECTLY DETECTED TOPICS IN MCG-WEBV

Topic ID	Description
28	Hajime no Ippo New Challenger, a cartoon.
51	WWE SmackDown, a popular TV shows held by World Wrestling Federation.
50	Mozart's Greatest Hits.
41	South Africa trip videos.
53	BBC Nature's Great Events.
3	New characteristic in google earth 5.0.
22	Amazing car accident.
16	The videos about Madonna's Concert in brazil.
39	Application videos for the best job in the world, island caretaker on Australia's Great Barrier Reef.
19	HotForWords: one popular channel on YouTube hosted by Marina Orlova for discussing the origins of words.
11	Videos about the dream boxing match.
1	Bush was attacked by shoes in press conference in Iraq

REFERENCES

[1] D. Gillmor, "We the media - grassroots journalism by the people, for the people," pp. 1–301, 2006. 1

[2] J. Cao, C. Ngo, Y. Zhang, and J. Li, "Tracking web video topics: Discovery, visualization, and monitoring," *IEEE TCSVT*, vol. 21, no. 12, pp. 1835–1846, 2011. 2, 3, 4, 6, 8, 10

[3] T. Chen, C. Liu, and Q. Huang, "An effective multi-clue fusion approach for web video topic detection," in *ACM Multimedia*, 2012, pp. 781–784. 2, 3, 4, 6, 8, 10, 11

[4] Y. Zhang, G. Li, L. Chu, S. Wang, W. Zhang, and Q. Huang, "Cross-media topic detection: A multi-modality fusion framework," in *ICME*, 2013, pp. 1–6. 2, 3, 8, 10

[5] Q. He, K. Chang, E.-P. Lim, and A. Banerjee, "Keep it simple with time: A reexamination of probabilistic topic detection models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1795–1808, 2010. 2, 3, 4

[6] K. Chen, L. Luesukprasert, and S. cho Timothy Chou, "Hot topic extraction based on timeline analysis and multidimensional sentence modeling," *IEEE TKDE*, vol. 19, no. 8, pp. 1016–1025, 2007. 2, 3

[7] J. Shao, S. Ma, W. Lu, and Y. Zhuang, "A unified framework for web video topic discovery and visualization," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 410–419, 2012. 2, 3, 4, 8

[8] S. P. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–136, 1982. 2

[9] S. Liu, Y. Liu, L. Ni, J. Fan, and M. Li, "Towards mobility-based clustering," in *ACM SIGKDD*, 2010. 2

[10] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007. 2

[11] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, pp. 37–50, 1912. 3, 5, 6

[12] J. Allan, "Topic detection and tracking: event-based information organization," in *The Kluwer international series on information retrieval*, 2002. 3

[13] T. C. Chou and M. C. Chen, "Using incremental plsi for threshold-resilient online event analysis," *IEEE TKDE*, vol. 20, pp. 289–299, 2008. 3

[14] H. Anaya-Sánchez, A. Pons-Porrata, and R. B. Llavori, "A document clustering algorithm for discovering and describing topics," *PRL*, vol. 31, no. 6, pp. 502–510, 2010. 3

[15] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: an exploration of temporal text mining," in *ACM SIGKDD*, 2005, pp. 198–207. 3

[16] C. Pan and P. Mitra, "Event detection with spatial latent dirichlet allocation," in *JCDL*, 2011, pp. 349–358. 3, 4

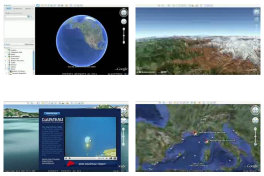
[17] Y. Yang, J. G. Carbonell, R. D. Brown, T. Pierce, B. T. Archibald, and X. Liu, "Learning approaches for detecting and tracking news events," *IEEE Intelligent Systems*, vol. 14, no. 4, pp. 32–43, Jul. 1999. 3

[18] Y. Zhai and M. Shah, "Tracking news stories across different sources," in *ACM Multimedia*, 2005, pp. 2–10. 3

[19] X. Wan and J. Xiao, "Graph-based multi-modality learning for topic-focused multi-document summarization," in *IJCAI*, 2009, pp. 1586–1591. 3

[20] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study final report," *Proceedings of the*

TABLE VIII
THREE TYPICAL EXAMPLES OF DETECTED TOPICS

Topic ID	Topic Title	Top 12 Automatically Generated Keywords	Representative Keyframes	#Word in Topic	#Word in Intersection
1	Bush was attacked by shoes during press conference in Iraq	Bush, Iraq, shoes, George, throw, president, journalist, conference, press, Baghdad, insult, Dodge		284	51
3	New characteristic in Google earth 5.0.	Google, earth, ocean, explore, Googleocean, historical, imagery, touring, googleearth, layer, changes, aral		38	14
16	The videos about Madonna's Concert in Brazil	Rio, janeiro, brasil, beat, maracan, candy, brazil, globo, copacabana, rede, palace, falls		73	48

DARPA Broadcast News Transcription and Understanding Workshop, pp. 194–218, Feb. 1998. 3

[21] W. H. Hsu and S. Chang, "Topic tracking across broadcast news videos with visual duplicates and semantic concepts," in *ICIP*, 2006, pp. 141–144. 3

[22] S. Wang, S. Jiang, Q. Huang, and Q. Tian, "Multi-feature metric learning with knowledge transfer among semantics and social tagging," in *CVPR*, 2012, pp. 2240–2247. 3

[23] X. Wu, C. Ngo, and A. G. Hauptmann, "Multimodal news story clustering with pairwise visual near-duplicate constraint." 2008, pp. 188–199. 3, 4

[24] L. Liu, L. Sun, Y. Rui, Y. Shi, and S. Yang, "Web video topic discovery and tracking via bipartite graph reinforcement model," in *WWW*, 2008, pp. 1009–1018. 3

[25] Y. Jiang, X. Zeng, G. Ye, D. Ellis, S. Chang, S. Bhattacharya, and M. Shah, "Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching," in *TRECVID*, 2010. 3

[26] Z. Lan, L. Bao, S.-I. Yu, W. Liu, and A. G. Hauptmann, "Double fusion for multimedia event detection," in *MMM*, 2012, pp. 173–185. 3

[27] Y. Fu, T. M. Hospedales, T. Xiang, D. Ellis, and S. Gong, "Queen mary university of london trecvid-2013 multimedia event detection (med) system report authors," in *TRECVID*, 2013. 3

[28] R. Hong, J. Tang, H. Tan, C. Ngo, S. Yan, and T. Chua, "Beyond search: Event-driven summarization for web videos," *TOMCCAP*, vol. 7, no. 4, p. 35, 2011. 3, 4

[29] M. Wang, R. Hong, G. Li, Z. Zha, S. Yan, and T. Chua, "Event driven web video summarization by tag localization and key-shot identification," *IEEE Trans. on Multimedia*, vol. 14, no. 4, pp. 975–985, 2012. 3, 4

[30] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z. Zhou, "Multi-view video summarization," *IEEE Trans. on Multimedia*, vol. 12, no. 7, pp. 717–729, 2010. 3, 4

[31] M. Wang, X. Hua, R. Hong, J. Tang, G. Qi, and Y. Song, "Unified video annotation via multigraph learning," *IEEE TCSVT*, vol. 19, no. 5, pp. 733–746, 2009. 3, 4

[32] H. Tong, J. He, M. Li, C. Zhang, and W. Ma, "Graph based multi-modality learning," in *ACM Multimedia*, 2005, pp. 862–871. 3

[33] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas, "Query specific fusion for image retrieval," in *ECCV*, 2012, pp. 660–673. 3

[34] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003. 4, 5, 12

[35] X. Wang and E. Grimson, "Spatial latent dirichlet allocation," in *NIPS*, 2007. 4

[36] C. Wang, D. M. Blei, and F. Li, "Simultaneous image classification and annotation," in *CVPR*, 2009, pp. 1903–1910. 4

[37] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Learning multimodal latent attributes," *IEEE TPAMI*, vol. 36, no. 2, pp. 303–316, 2014. 4

[38] X. Wu, A. G. Hauptmann, and C. Ngo, "Practical elimination of near-duplicates from web video search," in *ACM Multimedia*, 2007, pp. 218–227. 4

[39] L. Xie, A. Natsev, J. R. Kender, M. L. Hill, and J. R. Smith, "Visual memes in social media: tracking real-world news in youtube videos," in *ACM Multimedia*, 2011, pp. 53–62. 4, 6, 8

[40] J. Zhu, S. C. H. Hoi, M. R. Lyu, and S. Yan, "Near-duplicate keyframe retrieval by nonrigid image matching," in *ACM Multimedia*, 2008, pp. 41–50. 4

[41] T. Chen, S. Jiang, L. Chu, and Q. Huang, "Detection and location of near-duplicate video sub-clips by finding dense subgraphs," in *ACM Multimedia*, 2011, pp. 1173–1176. 4

[42] C. Ngo, W. Zhao, and Y. Jiang, "Fast tracking of near-duplicate keyframes in broadcast domain with transitivity propagation," in *ACM Multimedia*, 2006, pp. 845–854. 4

[43] X. Wu, W. Zhao, and C. Ngo, "Near-duplicate keyframe retrieval with visual keywords and semantic context," in *CIVR*, 2007, pp. 162–169. 4

[44] W. H. Hsu and S. Chang, "Topic tracking across broadcast news videos with visual duplicates and semantic concepts," in *ICIP*, 2006, pp. 141–144. 4

[45] S. Speakman, Y. Zhang, and D. B. Neill, "Dynamic pattern detection with temporal consistency and connectivity constraints," in *ICDM*, 2013, pp. 697–706. 4

[46] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *VISAPP*, 2009, pp. 331–340. 6

[47] H. Liu and S. Yan, "Robust graph mode seeking by graph shift," in *ICML*, 2010, pp. 671–678. 7

[48] M. Pavan and M. Pelillo, "Dominant sets and pairwise clustering," *TPAMI*, vol. 29, no. 1, pp. 167–172, 2007. 7

[49] ——, "Dominant sets and hierarchical clustering," in *ICCV*, 2003, pp. 362–369. 7

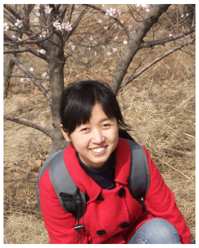
[50] S. Rota Bulò, M. Pelillo, and I. M. Bomze, "Graph-based quadratic optimization: A fast evolutionary approach," *CVIU*, vol. 115, no. 7, pp. 984–995, 2011. 7

[51] T. L. Griffiths and M. Steyvers, "Mapping knowledge domains: Finding scientific topics," 2004, pp. 5228–5235. 8, 9

[52] J. Cao, Y. Zhang, Y. Song, Z. Chen, X. Zhang, and J. Li, "Mcg-webv: A benchmark dataset for web video analysis," May. 2009. 8, 12

Lingyang Chu received the B.S. degree from the Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2009. He is currently a Ph.D student with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include duplicate image and video retrieval, Web event detection and large-scale Web multimedia data mining.

Yanyan Zhang received the Master's degree from the University of Chinese Academy of Sciences, Beijing, China, in 2014. She is currently an employee of the Bank of China. She was also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. Her research interests include image retrieval and Web multimedia data mining.



Guorong Li received the B.S. degree in technology of computer application from the Renmin University of China, Beijing, China, in 2006, and the Ph.D. degree in technology of computer application from the Graduate University of the Chinese Academy of Sciences (GUCAS), Beijing, in 2012. She is currently a Post-Doctoral Fellow with GUCAS. Her current research interests include cross-media analysis, object tracking, video analysis, pattern recognition, and computer vision.

Shuhui Wang (M'12) received the B.S. degree in Electronic Engineering from Tsinghua University, Beijing, China, in 2006, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2012. He is currently a researcher with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include semantic image analysis, image and video retrieval and large-scale Web multimedia data mining.

Weigang Zhang (M'13) received the B.S. and M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2003 and 2005. He is currently working toward the Ph.D. degree at the School of Computer Science and Technology, Harbin Institute of Technology. He is also a faculty member with School of Computer, Harbin Institute of Technology at Weihai, Weihai, China. His research interests include multimedia analysis and retrieval, image processing and pattern recognition.

Qingming Huang [SM'08] is a professor in the University of Chinese Academy of Sciences and an adjunct research professor in the Institute of Computing Technology, Chinese Academy of Sciences. He graduated with a Bachelor degree in Computer Science in 1988 and Ph.D. degree in Computer Engineering in 1994, both from Harbin Institute of Technology, China. His research areas include multimedia content analysis, image processing, computer vision, pattern recognition and machine learning. He has published more than 200 academic papers in prestigious international journals including IEEE Trans. on Image Processing, IEEE Trans. on Multimedia, IEEE Trans. on CSVT, etc, and top-level conferences such as ACM Multimedia, ICCV, CVPR and ECCV. He served as program chair, organization committee or TPC members in various well-known international conferences, including ACM Multimedia, ICCV, ICME, PSIVT, etc.