Online Multimodal Multiexpert Learning for Social Event Tracking

Shengsheng Qian, Tianzhu Zhang, *Member, IEEE*, and Changsheng Xu, *Fellow, IEEE*

***Abstract*—In this paper, we aim to automatically identify and track the interesting social event from vast amounts of social media data. However, there are two existing challenges: 1) how to model multimodal social event data over time and visualize the topic evolution and 2) how to alleviate the tracking drift problem to boost social event tracking accuracy. We propose a novel online multimodal multiexpert learning algorithm for social event tracking. Compared with existing methods, the proposed model has several advantages: First, it has a nonparametric online multimodal tracking module, which is able to not only automatically learn the number of topics from data over time, but also exploit the multimodal property of the social event. Second, it adopts a novel multiexpert minimization restoration scheme and allows the tracked model to evolve backwards to undo undesirable model updates, which helps alleviate the model drift problem of social event tracking. Third, it is able to not only effectively track the multimodal social event, but also automatically exploit the topic evolution of the social event for a deep understanding with multimodal topics. To evaluate the proposed model, we collect a real-world dataset for research on social event tracking with multimodality information. We have conducted extensive experiments, and both qualitative and quantitative evaluation results have demonstrated the effectiveness of the proposed model.**

***Index Terms*—Social event tracking, topic evolution, multimodality, topic model, social media.**

I. Introduction

With the rapid development of Internet, more and more social networking sites (e.g., Flickr, YouTube, Facebook, and Google News) appear and allow users to share ideas, pictures, posts, activities, events, and interests with people in their network. As a result, a popular event that is happening around us and around the world can spread very fast in different media sites, and has substantial amounts of media data with multi-modality (e.g., images, videos, and text). Here, an social event is something that occurs at specific place and time associated with some specific actions and consists of many stories over time [1]. However, most of these multimedia contents associated with social events uploaded by users are related to some specific topics, and it is very time-consuming for people to manually identify or cluster them to obtain the whole topic evolutions of the social events in real-world scenarios. Social event tracking can alleviate the above problem, and its goal is to automatically identify and track the interesting social event from vast amounts of social media data. For example, users may want to track the whole topic evolution of the event "2011 England Riot" from start to end. When they search for the recent related event on Google News to collect information given the well-defined query, they could get lots of related information as shown in the left panel of Fig. 1. However, all the results only tend to show what has happened recently and are too similar. They have to repeatedly switch back and forth from one to another in order to completely understand the content of the events. Moreover, it is time-consuming for people to browse such huge documents, and the users could not capture the whole topic evolution of the event. Thus, to support such an analysis process, it is important to automatically gather these semantic topics scattered in different documents and visualize the theme pattern over time. As a result, for the event "2011 England Riot", we can know the topic evolution of the event in different cities over time, as shown in the right panel of Fig. 1. Therefore, social event tracking is very important and helpful for social event understanding to know its topic evolution over time automatically.

Recently, how to mine and monitor social event in social media has attracted extensive research interests, such as social event mining [2]–[4], social event detection and tracking [5]–[9] and event evolution [10]–[12]. There are mainly two challenging factors to track the topic evolution of multiple events over time in social media data. Firstly, in the social media, it consists of a lot of unstructured metadata in multiple modalities, and is different from the traditional event tracking and evolution problem involving a single modality such as textual information. In different social media platforms, social media events have rich multi-modal information, such as text, images, and videos, which can complement each other and are helpful for the social event analysis [13]–[15]. For example, given the same social events, they may have different textual information due to different users, but their visual information may be similar. However, almost all existing work focuses on either textual features or images [16], [17] in isolation. In the social event analysis, we need to analyze these multi-modality data in a unified way for

Manuscript received July 15, 2016; revised March 16, 2017 and September 19, 2017; accepted February 16, 2018. Date of publication March 15, 2018; date of current version September 18, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61720106006, Grant 61432019, Grant 61572498, Grant 61532009, and Grant 61572296; in part by the Key Research Program of Frontier Sciences, CAS, under Grant QYZDJ-SSW-JSC039; and in part by the Beijing Natural Science Foundation under Grant 4172062. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. K. Selcuk Candan. *(Corresponding author: Changsheng Xu.)*

The authors are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences (e-mail: shengsheng.qian@nlpr.ia.ac.cn; tzzhang@nlpr.ia.ac.cn; csxu@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2018.2815785

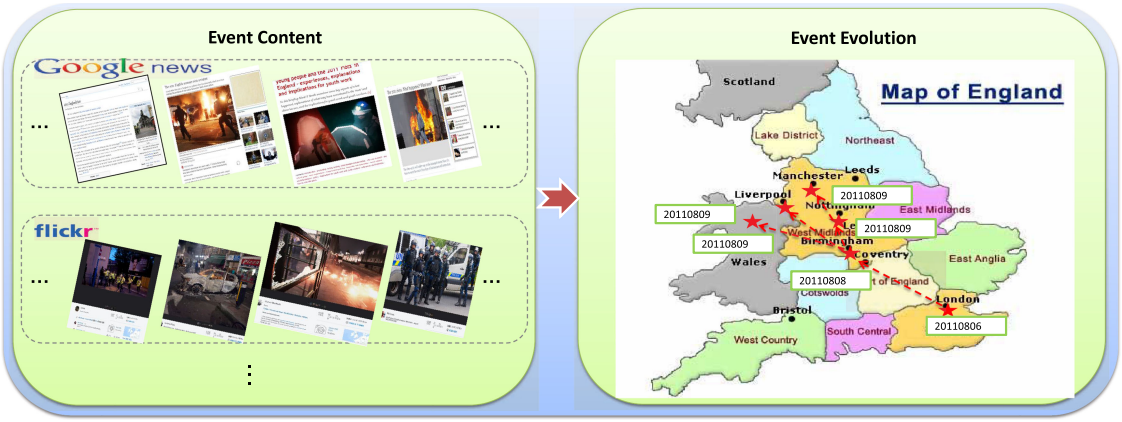


Fig. 1. The left panel includes the related event documents returned by querying "2011 England Riot" in different social media sites (Google News and Flickr). The right panel shows the whole topic evolution of "2011 England Riot" denoted with a red trajectory using event tracking method. Here, we show the location and its happening time on the map.

event modeling. Secondly, in the social event tracking process, events may be very similar in social media. For example, the event "Occupy Wall Street" and the event "United States presidential election" have many topics, and some topics are similar, such as some key words "United, States, government, president" in the topic "Government Departments". Moreover, since most information in these social media sites is generated by users, it may contain much noise. For example, the text and the images may be irrelevant to the event. The above issue makes social event tracking be prone to drift in an online model updating process. Therefore, it is necessary to design an effective model to explore a multi-modal fusion strategy and alleviate the model drift problem for social event tracking and evolution analysis.

To solve the above two challenges: (i) how to model multi-modal social event data over time and (ii) how to alleviate tracking drift problem, many methods have been proposed. In the latest study, many topic model methods [18]–[21] have been proposed to explore multi-modal topics for social event analysis. For example, Corr-LDA [18] and MoM-LDA [19] are proposed to capture the topic-level relations between images and annotations. In [21], a novel multi-modal event topic model (mmETM) is proposed to effectively model multi-modal social media documents including long text with related images and learn correlations between textual and visual modalities to separate the visual-representative topics and non-visual-representative topics. However, these approaches must either assume the number of topics or train multiple models with different settings and select the best one in the traditional topic models of the LDA variants. This means that the user must make an assumption about the structure of the collection, or carry some kind of model selection exercise. Experimentally, it has been shown [22] that the performance of the model is influenced by the number of topics. To deal with this issue, Teh *et al.* [22] and Yakhnenko *et al.* [23] propose the Hierarchical Dirichlet Process (HDP) and MoM-HDP model, which are the non-parametric approaches to automatically learn the number of topics from data for topic modeling. However, traditional HDP and MoM-HDP models become impractical when the data set size is large, and they become impossible when the data are streaming. Especially, for event tracking, we need consider the temporal order of event documents to online learn the topic evolution. To address this issue, we propose a novel online multi-modal tracking model (online-MMTM) based on the traditional MoM-HDP model by using a novel online variational inference algorithm, which can be suitable to model multi-modal social event documents for event tracking and evolution analysis. Moreover, the process of online model updating usually makes social event tracking be prone to drift due to the misalignment and noisy samples. To overcome the above problem, several approaches have been proposed to avoid bad model updating [24]–[26]. In [25], a novel paradigm is proposed for training a binary classifier from labeled and unlabeled examples called P-N learning, where one based on spatial constraints and the other based on temporal constraints, to alleviate drift. In [26], Zhang *et al.* propose a robust visual tracking algorithm via multiple experts using entropy minimization, and the proposed MEEM restoration scheme can significantly improve the robustness of the base tracker, especially in challenging scenarios with occlusions and repetitive appearance variations. Therefore, we utilize a novel multi-expert minimization restoration scheme to correct the effects of bad model updates by using an entropy-regularized optimization function as our expert selection criterion and online improve the social event model.

Inspired by the previous work [26], [27], we propose a novel online multi-modal multi-expert learning (OMMEL) algorithm to online obtain multiple topic models for social event tracking, as shown in Fig. 2. The basic idea is to effectively integrate an online multi-modal tracking model into the multi-expert ensemble framework. Our multi-expert ensemble framework adopts a novel multi-expert minimization restoration scheme and allows the tracked model to evolve backwards to undo undesirable model updates, where an expert ensemble consists of a learned tracker and its former snapshots. To implement the base tracker in our multi-expert tracking framework, we propose a novel online multi-modal tracking model method (online-MMTM) based on the traditional MoM-HDP model by using a novel online

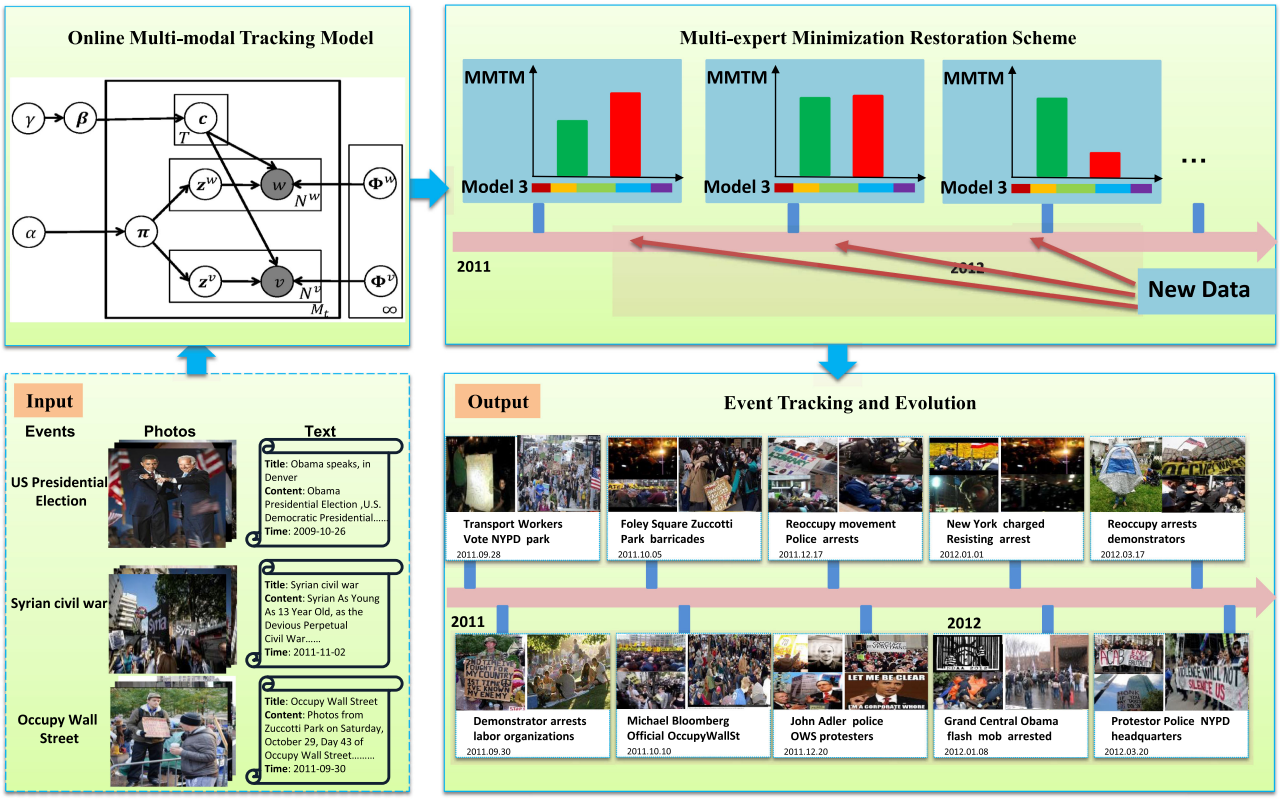


Fig. 2. The online multi-modal multi-expert learning framework. The input is the multi-modal data collected including images and the corresponding texts. Based on the input data, our multi-modal multi-expert tracking algorithm can learn multi-modality topics and track multiple events. Specifically, (i) the proposed online multi-modal tracking model method (online-MMTM) is able to not only automatically learn the number of topics from data over time, but also exploit the multi-modal property of social event. (ii) A novel multi-expert minimization restoration scheme is adopted to allow the tracked model to evolve backwards, which is to help alleviate the model drift problem of social event tracking. After tracking, for each event, it can be visualized with texts and images over time. Meanwhile, we can mine their semantic topics.

variational inference algorithm, which can be suitable to model multi-modal social event documents for event tracking and evolution analysis. In [27], the authors propose a novel online inference model for the Hierarchical Dirichlet Process (HDP) to exploit the topic evolution of time-series textual data, which does not address social event tracking problem. Different from the [27], we focus on social event tracking with multi-modal data including texts and images by using the online inference model. In [26], it focuses on single object tracking and utilizes a restoration scheme to boost tracking performance, which cannot learn the topic evolution of social event over time. In this paper, our goal is to track social event and learn its topic evolution over time. Because social event tracking is an online learning problem with multi-modal data and has model drift issue, we propose an online multi-modal multi-expert learning method, which can effectively model multi-modal data online and alleviate the model drift issue via a multi-expert learning strategy. Compared with existing methods, the contributions of this work are four-fold.

1) Our online multi-modal multi-expert learning algorithm adopts a novel multi-expert minimization restoration scheme and is able to evolve backwards to undo undesirable model updating, which can alleviate the drift problem of social event tracking.
2) The proposed online multi-modal tracking model is a non-parametric approach, which is able to not only automatically learn the number of topics from data over time, but also exploit the multi-modal property of social event.
3) Our online multi-modal multi-expert learning algorithm makes use of the supervised information of social event with an online multi-modal tracking model, which can obtain the whole topic evolution of social event over time and help deeply understand the event.
4) We collect a dataset for research on social event tracking with multi-modality information, and will release it for academic use. We evaluate the proposed model and demonstrate that it achieves much better performance than existing methods.

The rest of the paper is organized as follows. In Section II, the related work is reviewed. Our algorithm is presented in Section III. In Section IV, we report and analyze extensive experimental results. Finally, we conclude the paper with future work in Section V.

II. Related Work

In this section, we briefly review previous methods which are most related to our work including social event detection

and tracking, tracking model drift, and topic model learning methods.

Social Event Detection and Tracking: With the massive growth of social events in Internet, how to recognize and monitor social event becomes more and more challenging. Researchers have been working on social event analysis and propose many different methods [17], [28]–[30], which are based on single-modality (e.g., text, images) information or multi-modality information. In the single-modality analysis, many existing methods adopt visual information (e.g., images and videos) or textual information (e.g., names, time references, locations, title, tags, and description) in isolation [16], [29], [31] to model event data for event detection and tracking. In [16], a novel topic detection algorithm is developed. The idea is to first classify the incoming news into predefined categories and then use the topic-conditioned heuristics to identify the new events. In [29], the authors exploit the rich context associated social media data and use clustering algorithms to identify events. In the social event analysis, limited efforts have been devoted to analyzing multi-modality data in a unified way. In different social media platforms, social media events have rich multi-modal information, such as text, images, and videos, which complement each other and are helpful for the social event analysis [13], [14], [32]. Recently, social event analysis with multi-modality has received the considerable attention. Kender and Naphade [33] study the correlation between manually annotated visual concepts (e.g., sites, people, and objects) and topic annotations, then use graph-cut techniques in story clustering. Zhai and Shah [34] propose a concept tracking method to link news stories from different TV channels by textual correlation and keyframe matching. Zhang and Xu [8] propose a CO-PMHT algorithm for cross domain multi-event tracking, which can track events by use of cross-domain knowledge and obtain their summary information over time. Different from the above methods which focus on feature designing strategy, we make use of the rich multi-modal contents associated with social events, including visual information and the corresponding text, and propose a novel online multi-modal tracking model method (online-MMTM) based on the traditional MoM-HDP model by using a novel online variational inference algorithm, which can efficiently model multi-modal social event documents.

Tracking Model Drift: In the social event tracking, online model updating usually brings the model drift problem due to the misalignment and noisy of training samples, and some approaches are proposed to avoid bad model updates [24]–[26], [35]. Matthews *et al.* [24] propose a template update algorithm that avoids the drifting inherent in the naive algorithm. Kalal *et al.* [25] propose a novel paradigm for training a binary classifier from labeled and unlabeled examples called P-N learning to alleviate drift. Zhang *et al.* [26] propose a novel MEEM restoration scheme by using entropy minimization to effectively improve the robustness of the base tracker in the visual tracing. In this paper, we mainly utilize a novel multi-expert minimization restoration scheme to correct the effects of bad model updates and online improve the social event tracking model. Moreover, the proposed online multi-modal multi-expert learning framework can obtain the whole topic evolution of social events over time and help deeply understand the events by combining the supervised information of social event with an online multi-modal tracking model.

Topic Model Learning: Topic models, such as Latent Dirichlet Allocation (LDA) [36] and Probabilistic Latent Semantic Analysis (PLSA)[37], have been widely applied to various applications and have many extensions, such as supervised Latent Dirichlet Allocation (SLDA) [38]–[40], multi-modal Latent Dirichlet Allocation (MoM-LDA) [41], [42], multi-modal event topic model (mmETM) [21]. Wang and McCallum [43] propose the topic over time (TOT) model to capture not only the low-dimensional structure of data, but also how the structure changes over time. In the multi-modal data analysis, many topic model methods [18], [20], [42] have been proposed to explore multi-modal topics. The Corr-LDA [18] and MoM-LDA [41] are proposed to capture the topic-level relations between images and annotations. The mmETM [21] mainly focuses on the multi-modal data feature representation and model multi-modal document including long text and relative image to learn the correlations between the textual and visual modalities. However, the above approaches must assume the the number of topics in the traditional topic models of the LDA variants, and the user must make an assumption about the structure of the collection, or carry some kind of model selection exercise. It would be better if the model could learn the number of topics from the data itself. To deal with this issue, Teh *et al.* [22] propose the Hierarchical Dirichlet Process (HDP), which is the non-parametric approach to topic modeling to automatically learn the number of topics from data. Similarly, Yakhnenko *et al.* [23] extend the MoM-LDA model to the nonparametric Bayesian model and propose a MoM-HDP model to address the number of topics based on the training data. However, the limitation of traditional HDP and MoM-HDP methods in model inference process is that existing posterior inference algorithms require multiple passes through all the data, and these algorithms are intractable for very large scale applications. Therefore, traditional HDP and MoM-HDP models become impractical when the data set size is large, which can not online update model in the streaming data. For event tracking step, the event topic evolution should consider the temporal ordering about event documents and online update social event model. To address this issue, we propose a novel online multi-modal tracking model method (online-MMTM) based on the traditional MoM-HDP model by using a novel online variational inference algorithm [27], which can be suitable to model multi-modal social event documents for event tracking and evolution analysis. Compared with our previous work [21], the difference is as follows: (1) We deal with different problems. In [21], we mainly focus on the multi-modal data feature representation and model multi-modal document including long text and related image to learn the correlations between the textual and visual modalities. In this work, we focus on the social event tracking and propose a novel online multi-modal multi-expert tracking algorithm to improve the event tracking performance. (2) The models are different. In [21], our work is the multi-modal extensions of the Corr-LDA [18] and the mm-LDA [41], and is applied to social event evolution and visualization analysis by adopting an incremental learning strategy. Especially

in the event tracking process, we only use the traditional similarity computing identification method. In this work, we adopt a novel multi-expert minimization restoration scheme to deal with the model updating issue for social event tracking. (3) We have different models for social event representation. In [21], the proposed model is to model multi-modal data, and it must either assume the number of topics or train multiple models with different settings. However, the proposed method in this paper adopts a nonparametric approach to model multi-modal data, which is able to not only automatically learn the number of topics from data over time, but also exploit the multi-modal property of social event.

III. The Proposed Algorithm

In this section, we first overview our proposed algorithm for multi-modal social event tracking and evolution analysis. We then introduce our proposed online multi-modal tracking model, classifier learner designing and multi-expert minimization restoration scheme in details. Finally, we show how to use the proposed model to build effective tracks for social event tracking analysis.

A. Preliminaries

In this paper, the social event tracking and evolution aims at linking together evolving and historical stories. To describe multi-modal social event data, we adopt the traditional bag-of-word method for both image and text as other existing topic model methods [36], [38], [39]. In this way, word ordering is ignored and a document is simply reduced to a vector of word count. Here, a multimedia document consisting of image and the corresponding text is thus summarized as a pair of vectors of word counts corresponding to visual and textual information, respectively. Specifically, an image word is denoted as a unit-basis vector v of size D_v with exactly one non-zero entry representing the membership to only one word in a dictionary of D_v words. A text word w_n is similarly defined for a dictionary of size D_w. An image is a collection of N^v word occurrences denoted by $\mathbf{v} = \{v_1, v_2, \ldots, v_{N^v}\}$, and the text is a collection of N^w word occurrences denoted by $\mathbf{w} = \{w_1, w_2, \ldots, w_{N^w}\}$.

Given a set of social media documents over time, the problem that we address in this paper is to track multiple events (e.g., Syrian civil war, US presidential election) that are reflected in the documents, as well as the documents that correspond to each event. Suppose we are given a collection of social media documents S_t related to the same social event with a sliding windows time t, where each document d consists of two components: textual component w_d and visual component v_d, our aim is to learn an online tracking model to predict the class label Y_t of event documents S_t at epoch t, $t \in \{1, \ldots, T\}$. Here, the epoch t is a discrete variable, we set time period for an epoch according to the evolution time of social events. To achieve this goal, we propose an online multi-modal multi-expert learning (OMMEL) algorithm by introducing a novel online multi-modal tracking model and utilizing a novel multi-expert minimization restoration scheme, which is to obtain their informative summary details and the evolutionary trends of social events over time. At the tracking step of each epoch t, our algorithm has three major components, including online multi-modal tracking model, classifier learner designing, and multi-expert minimization restoration scheme, described as follows.

B. Online Multi-Modal Tracking Model

To consider the temporal ordering of multi-modal social event documents and track multiple events over time, we propose a novel online multi-modal tracking model (online-MMTM) based on the traditional MoM-HDP model by using a novel online variational inference algorithm. The proposed method can sequentially update the model at each epoch and model multi-modal social event documents for event tracking and evolution analysis. In the following, we first introduce the MoM-HDP model, and then show the details of the proposed online variational inference algorithm.

1) MoM-HDP: Before introducing the MoM-HDP model [23], we overview the multi-modal LDA (MoM-LDA), which is the basis of our model. In the MoM-LDA [41], one social media document contains an image and its corresponding text. Each document has topic distribution π_d, for each of the textual words N_d^w in the document, topic $z_{d,n}^w$ is chosen from the topic distribution, and then textual word $w_{d,n}$ is generated from a topic-specific multinomial distribution over words ϕ^w. And the generative process of the visual word v is similar. Fig. 3(a) shows the graphical model of the MoM-LDA, where shaded and unshaded nodes indicate observed and latent variables, respectively. Since the MoM-LDA must either assume the number of topics or train multiple models with different settings, the choice of number of the mixture components can have a major effect on how well the model fits the data. To deal with this issue, Yakhnenko *et al.* [23] extend the MoM-LDA model to the nonparametric Bayesian model and propose a MoM-HDP model to address the number of topics based on the training data, as shown in Fig. 3(b).

The MoM-HDP model mainly applies the traditional Teh's stick-breaking construction [22] about the priors for the hierarchical Dirichlet process to the multi-modal generative model, where the Dirichlet Process (DP) is a generalization of a finite mixture model, and it assumes countably infinite number mixture components. Like in the case of MoM-LDA, the MoM-HDP model assumes that each observable modality is clustered by the mixture components, so that textual word $w_{d,n}$ is generated from a topic-specific multinomial distribution over words ϕ^w, and visual word $v_{d,n}$ is generated from a topic-specific multinomial distribution over words ϕ^v. The topic distribution π_d is drawn from $DP(\alpha^\pi, \beta)$, where β is constructed using a stick-breaking distribution. Furthermore, the parameters for observations given their topic distribution ϕ^w and ϕ^v are generated from some base distribution G_0 (such as a Dirichlet distribution). In Fig. 3, we also note that if the prior β is assumed to be drawn from finite Dirichlet instead of a stick-breaking distribution, this model becomes a Dirichlet-smoothed version of the MoM-LDA.

In the derivation process of MoM-HDP model, they use traditional Teh's Stick-breaking Construction method [22]. Specifically, a two-level hierarchical Dirichlet process (HDP) is a

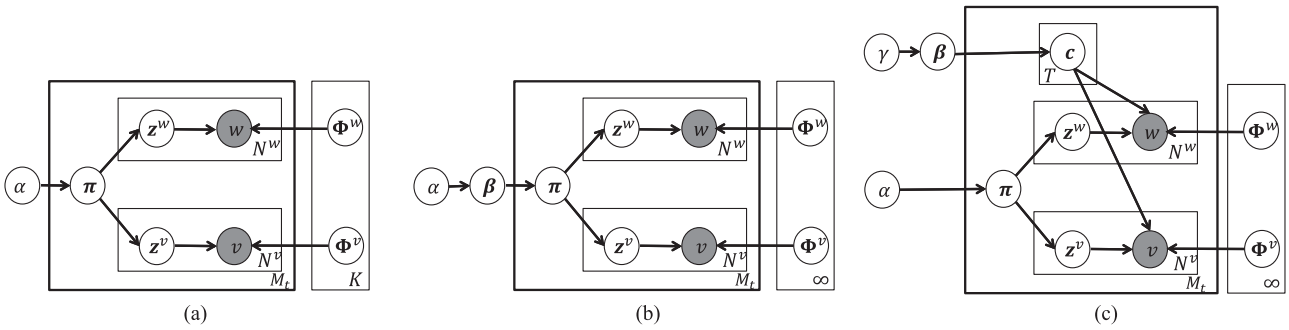


Fig. 3. The graphical models of (a) Multi-modal Latent Dirichlet Allocation (MoM-LDA) [41], (b) Multi-Modal Hierarchical Dirichlet Process (MoM-HDP) [23], and (c) Online multi-modal tracking model (online-MMTM).

collection of Dirichlet processes (DP) [44] that shares a base distribution G_0, which is also drawn from a DP. Mathematically,

$$G_0 \sim \mathrm{DP}(\gamma H), \tag{1}$$

$$G_d \sim \mathrm{DP}(\alpha G_0),\ for\ each\ document\ d, \tag{2}$$

where the global Dirichlet process G_0 has a concentration parameter γ and an underlying base distribution H, the document level Dirichlet processes $(G_d)_{d=1}^{D}$ are independent given G_0, and a notable feature of the HDP is that all DPs G_d share the same set of atoms and only the atom weights differ. In the Teh's Stick-breaking Construction process, they extend the (1) and (2), and propose a more constructive representation of the HDP using two stick-breaking representations of a Dirichlet distribution. For the corpus-level DP draw, this representation is

$$\begin{aligned}
&\beta'_k \sim \mathrm{Beta}(1,\gamma)\\
&\beta_k = \beta'_k \prod_{l=1}^{k-1}(1-\beta'_l)\\
&\Phi_k \sim H\\
&G_0 = \sum_{k=1}^{\infty}\beta_k \delta_{\Phi_k},
\end{aligned} \tag{3}$$

and each document-level G_d is defined as:

$$\begin{aligned}
&\pi'_{dk} \sim \mathrm{Beta}\left(\alpha\beta_k, \alpha\left(1-\sum_{l=1}^{k}\beta_l\right)\right)\\
&\pi_{dk} = \pi'_{dk}\prod_{l=1}^{k-1}(1-\pi'_{dl})\\
&G_d = \sum_{k=1}^{\infty}\pi_{dk}\delta_{\Phi_k},
\end{aligned} \tag{4}$$

In the above construction, the stick-breaking weights are tightly coupled between the bottom and top-level DPs. This construction method can obtain the parameter estimation of the MoM-HDP model. However, this construction cannot be used in an online variational inference algorithm.

2) Online Multi-Modal Tracking Model: Since the limitation of the traditional HDP and MoM-HDP methods in model inference process is that existing posterior inference algorithms require multiple passes through all the data, and these algorithms are intractable for very large scale applications. Therefore, the traditional HDP and MoM-HDP models become impractical when the data set size is large and the data are streaming. Especially, for event tracking, we need consider the temporal ordering about event documents to online learn the topic evolution. To address this issue, we propose a novel online multi-modal tracking model method (online-MMTM) based on the traditional MoM-HDP model by using a novel online variational inference algorithm [27], which can be suitable to model multi-modal social event documents for event tracking and evolution analysis.

Specifically, we use an alternative construction for the MoM-HDP model that allows for closed-form coordinate-ascent variational inference. This alternative construction for the document level Dirichlet is defined as:

$$\begin{aligned}
&\psi_{dt} \sim G_0\\
&\pi'_{dt} \sim \mathrm{Beta}(1,\alpha)\\
&\pi_{dt} = \pi'_{dt}\prod_{l=1}^{t-1}(1-\pi'_{dl})\\
&G_d = \sum_{t=1}^{\infty}\pi_{dt}\delta_{\psi_{dt}},
\end{aligned} \tag{5}$$

where π_{dt} is the weight distribution for document topic ψ_{dt} of document d. The important difference is how the topics themselves are drawn. On the corpus level, topics are drawn from the prior distribution H, but on the document level, the topics ψ_{dt} are drawn from G_0 in the following way:

$$\begin{aligned}
&c_{dt} \sim Mult(\beta),\\
&\psi_{dt} \sim \Phi_{c_{dt}},
\end{aligned} \tag{6}$$

where c_{dt} is the indicator variable which indexes the corpus-level topic corresponding to ψ_{dt}. Fig. 3(c) illustrates the graphical representation of online-MMTM. From the above construction process, we do not need to explicitly represent the document topics ψ_{dt}, which can further simplify online inference.

TABLE I
KEY NOTATIONS OF OUR PROPOSED ONLINE-MMTM MODEL

Notations	Description
$d \in \{1, 2, ..., D\}$	The index of social event document
β'	the corpus level stick breaking proportion
N^v, N^w	The collection of images and associated text for a social event document d
Φ_k^w, Φ_k^v	The multinomial distributions of textual word and visual descriptor specific to the topic k
π'_d	per-document stick breaking proportion
$\mathbf{c}$	per-document topic indices which map the document topics to corpus level topics Φ^w, Φ^v
w, v	The textual word and visual descriptor vectors
z^w, z^v	The topic assignments for textual word and visual descriptor

Accordingly, the generative processes modeled by the MoM-HDP based on the alternative construction method for a document d with N_d^v visual words and N_d^w textual words can be described as follows:

1) For each topic $k \in \{1, 2, \ldots, K\}$,
   a) Draw a global topic proportion $\beta'_k \sim \text{Beta}(1, \gamma)$
   b) For textual topic z^w, Draw $\Phi_k^w \sim Dirichlet(\eta)$
   c) For visual topic z^v, Draw $\Phi_k^v \sim Dirichlet(\eta)$
2) For each topic to truncation level $t = 1, ..., T$
   a) Draw a set of topic indicators $c_{dt} \sim Mult(\beta)$
   b) Draw document topic proportions $\pi'_{dt} \sim \text{Beta}(1, \alpha)$
3) For each visual word v_n, $n \in \{1, 2, ..., N^v\}$
   a) Draw a topic assignment $z_n^v \sim Mult(\pi_d)$
   b) Draw a visual patch $v_n \sim Mult(\Phi_k^v), k = c_{dt}, t = z_n^v$
4) For each textual word w_n, $n \in \{1, 2, ..., N^w\}$
   a) Draw a topic assignment $z_n^w \sim Mult(\pi_d)$
   b) Draw a textual patch $w_n \sim Mult(\Phi_k^w), k = c_{dt}, t = z_n^w$

Table I lists the key notations. During the model learning process, we assume that the priors distributions follow symmetric Dirichlet, which are conjugate priors for multinomial. Every event document has two different kinds of features: w_n, v_n. The w_n and v_n are textual information and visual information, respectively, and they are semantically related to each other in each social event document.

Given the model, we need to estimate all the latent variables, the corpus level stick breaking proportion β', per-document stick breaking proportion π'_d, per-document topic indices $\mathbf{c}$ (which map the document topics to corpus level topics Φ), textual word topic assignment $\mathbf{z}^w$ (which is the document level topic index for each textual word in the document), visual word topic assignment $\mathbf{z}^v$ (which is the document level topic index for each visual word in the document), the textual topics Φ^w and the visual topics Φ^v. Exact inference is often intractable in many topic models and appropriate methods must be used, such as variational inference [36] and Gibbs sampling [45]. We use a fully factorized variational distribution and perform mean-field variational inference [36]. The MoM-HDP model can be described as:

$$q(\beta', \pi', c, z^w, z^v, \Phi^w, \Phi^v) \tag{7}$$

This further factorizes into:

$$q(\beta') = \sum_k^{K-1} q(\beta'_k | u_k, v_k) \tag{8}$$

$$q(\pi') = \prod_d \prod_t^{T-1} q(\pi'_{dt} | a_{dt}, b_{dt}) \tag{9}$$

$$q(\mathbf{c}) = \prod_d \prod_t q(c_{dt} | \varphi_{dt}) \tag{10}$$

$$q(z^w) = \prod_d \prod_n q(z_{dn}^w | \varsigma_{dn}^w) \tag{11}$$

$$q(z^v) = \prod_d \prod_n q(z_{dn}^v | \varsigma_{dn}^v) \tag{12}$$

$$q(\Phi^w) = \prod\nolimits_k q(\Phi_k^w | \lambda_k^w) \tag{13}$$

$$q(\Phi^v) = \prod_k q(\Phi_k^v | \lambda_k^v) \tag{14}$$

where (u_k, v_k), (a_{dt}, b_{dt}) are parameters of Beta distributions. The stick breaking proportions are truncated to finite values by setting $q(\beta'_K = 1) = 1$ and $q(\pi'_{dT} = 1) = 1$. $q(c_{dt})$, $q(z_{dn}^w)$ and $q(z_{dn}^v)$ are multinomial distributions with parameters φ_{dt}, ς_{dn}^w and ς_{dn}^v, respectively. $q(\Phi^w)$ and $q(\Phi^v)$ are Dirichlet distributions with parameters λ_k^w and λ_k^v. Note that, a particular advantage of this two-level stick-breaking distribution is that the document truncation T can be much smaller than K. Though there may be hundreds of topics in a large corpus, we expect each document will only exhibit a small subset of them. In a standard variational inference [36], using Jensens inequality, we obtain the lower bound of the marginal log likelihood of the observed data $D = (w_d, v_d)_{d=1}^D$,

$$\begin{aligned}
&\log p(D | \gamma, \alpha, \eta^w, \eta^v) \\
&\geq \mathbf{E}_q[\log p(D, \beta', \pi', c, z^w, z^v, \Phi^w, \Phi^v)] + H(q) \\
&= \sum_d \{\mathbf{E}_q[\log(p(w_d | c_d, z_d^w, \Phi^w) p(v_d | c_d, z_d^v, \Phi^v) p(c_d | \beta') \\
&\quad p(z_d^w | \pi'_d) p(z_d^v | \pi'_d) p(\pi'_d | \alpha)) \\
&\quad + H(q(c_d)) + H(q(z_d^w)) + H(q(z_d^v)) + H(q(\pi'_d))\} \\
&\quad + \mathbf{E}_q[\log p(\beta') p(\Phi^w) p(\Phi^v)] + H(q(\beta')) + H(q(\Phi^w)) \\
&\quad + H(q(\Phi^v)) = \ell(q).
\end{aligned} \tag{15}$$

where $H(\cdot)$ is the entropy term for the variational distribution. Then, we take derivatives of this lower bound with respect to each variational parameter, and use them to derive the corresponding update equations. The derivation of update rules is detailed in the Appendix.

3) Online Variational Inference: We now extend the variational inference to an online setting which can fit well with sequential data. The online inference algorithm for multi-modal tracking model is based on the stochastic learning algorithm developed for LDA and HDP in [27], [46]. The basic idea is that we use the document level parameters to calculate the natural

gradient of the topic level parameters, and then use the obtained gradients to update the topic level parameters. As in [27], we can redefine the variational lower bound for the document d,

$$\begin{aligned}\ell_d = \mathbf{E}_q[&\log(p(w_d|c_d, z_d^w, \Phi^w)p(v_d|c_d, z_d^v, \Phi^v)p(c_d|\beta') \\ &p(z_d^w|\pi_d')p(z_d^v|\pi_d')p(\pi_d'|\alpha))] \\ &+ H(q(c_d)) + H(q(z_d^w)) + H(q(z_d^v)) + H(q(\pi_d')) \\ &+ \frac{1}{D}\mathbf{E}_q[\log p(\beta')p(\Phi^w)p(\Phi^v)] + H(q(\beta')) \\ &+ H(q(\Phi^w)) + H(q(\Phi^v)). \end{aligned} \tag{16}$$

where D is the total number of documents in the event dataset. The original lower bound ℓ can be recovered by summing over the documents, and the lower bound ℓ in (15) can be written as:

$$\ell_d = \sum_d \ell_d = \mathbf{E}_d[D\ell_d] \tag{17}$$

Our approach is similar to [27], and we simply take the corpus topics updates in (42)–(45), and replace the summation over documents with multiplication by D:

$$\hat{u}_k = 1 + D\sum_{t=1}^{T}\varphi_{dtk} \tag{18}$$

$$\hat{v}_k = \gamma + D\sum_{t=1}^{T}\sum_{l=k+1}^{K}\varphi_{dtl} \tag{19}$$

$$\hat{\lambda}_k^w = \eta^w + D\sum_{t=1}^{T}\varphi_{dtk}\left(\sum_n \varsigma_{dnt}^w I[w_{dn} = w]\right) \tag{20}$$

$$\hat{\lambda}_k^v = \eta^v + D\sum_{t=1}^{T}\varphi_{dtk}\left(\sum_n \varsigma_{dnt}^v I[v_{dn} = v]\right) \tag{21}$$

In the online inference for multi-modal tracking model, we need set an appropriate learning ρ_t to ensure convergence of the corpus topics updates. Final, we can obtain the corpus topics updates with the learning rate:

$$u_k \leftarrow (1-\rho_{t_0})u_k + \rho_{t_0}\hat{u}_k \tag{22}$$

$$v_k \leftarrow (1-\rho_{t_0})v_k + \rho_{t_0}\hat{v}_k \tag{23}$$

$$\lambda_k^w \leftarrow (1-\rho_{t_0})\lambda_k^w + \rho_{t_0}\hat{\lambda}_k^w \tag{24}$$

$$\lambda_k^v \leftarrow (1-\rho_{t_0})\lambda_k^v + \rho_{t_0}\hat{\lambda}_k^v \tag{25}$$

We use the learning rate in our experiment as in [27], [46], $\rho_t = (t+\tau)^{-\kappa}$, and set $\kappa \in (0.5, 1.0]$ and $\tau > 0$. An overview of the proposed online-MMTM algorithm is shown in Algorithm 1.

Therefore, in each epoch t, event document set $\mathbf{S}_t$ consists of two components: textual component $\mathbf{w}_t$, and visual component $\mathbf{v}_t$, and each document is considered as the track's object. Then, on this subset, we apply our proposed online-MMTM model to learn the corresponding topic model $\mathbf{F}_t(\mathbf{S}_t) = \{\mathbf{c}, \Phi^w, \Phi^v\}$. Once obtaining the topic model $\mathbf{F}_t(\mathbf{S}_t)$, we can obtain the document-topic representation of multi-modal event document, and predict the label of a new event document.

Algorithm 1: Online-MMTM for parameter inference

input : Initialize $\mathbf{u} = (u_k)_{k=1}^{K-1}, \mathbf{v} = (v_k)_{k=1}^{K-1}, \lambda^w = (\lambda_k^w)_{k=1}^{K}, \lambda^v = (\lambda_k^v)_{k=1}^{K}$, randomly. Set t_0 = 1.

for $t = 1$ *to* T **do**

- For each document, update document-level parameters, $\mathbf{a_d}, \mathbf{b_d}, \varphi_\mathbf{d}, \varsigma_\mathbf{d}^\mathbf{w}, \varsigma_\mathbf{d}^\mathbf{v}$ according to Eq.(37), Eq.(38), Eq.(39), Eq.(40).
- Compute the natural gradients, $\hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k, \hat{\lambda}_k^w, \hat{\lambda}_k^v$ according to Eq.(18), Eq.(19), Eq.(20), Eq.(21).
- Set $\rho_{t_0} = (t_0 + \tau_0)^{-\kappa}, t_0 \leftarrow t_0 + 1$
- Update $\mathbf{u}, \mathbf{v}, \lambda^w, \lambda^v$ according to Eq.(22), Eq.(23), Eq.(24), Eq.(25).

end

C. Classifier Learner Designing

Once we obtain the topic model $\mathbf{F}_t(\mathbf{S}_t)$, we need to design an effective classifier learner $\mathbf{h}_t(\mathbf{E}_t)$. Specifically, we adopt an effective softmax regression function. Here, the softmax function provides the following distribution

$$p(y_d|\bar{\theta}_d, \mu_t) = \exp\left(\mu_{t,y_d}^T\bar{\theta}_d\right) \Big/ \sum_{l=1}^{C}\exp\left(\mu_{t,l}^T\bar{\theta}_d\right), \tag{26}$$

where $\mu_{t,1:C}$ represent a set of class coefficients that will be inferred from data, C represents the number of event class, y_d represents the predict label of the d-th event document, and $\bar{\theta}_d$ represents the document-topic representation of document d which can be obtained in Section III-B, specifically, $\bar{\theta}_d = \sum_{n=1}^{N^w}\sum_{t=1}^{T}\varphi_{dtk}\cdot\varsigma_{dnt}^w + \sum_{n=1}^{N^v}\sum_{t=1}^{T}\varphi_{dtk}\cdot\varsigma_{dnt}^v$.

The softmax classifier is defined as in (27).

$$\begin{aligned}\mathbf{h}_t(\mathbf{e}_{new}) &= \underset{y_{\mathbf{e}_{new}}\in\{1,2\cdots,C\}}{\arg\max}\left(p(y_{\mathbf{e}_{new}} = c|\overline{\mathbf{e}}_{new}, \mu_\mathbf{t})\right) \\ &= \underset{y_{e_{new}}\in\{1,2\cdots,C\}}{\arg\max}(\mu_{t,c}^T\overline{\mathbf{e}}_{new}),\end{aligned} \tag{27}$$

where $p(y_{\mathbf{e}_{new}} = c|\overline{\mathbf{e}}_{new}, \mu_\mathbf{t}) = \exp(\mu_{t,c}^T\overline{\mathbf{e}}_{new})/\sum_{l=1}^{C}\exp(\mu_{t,l}^T\overline{\mathbf{e}}_{new})$.

In the epoch $t+1$ for event tracking, given a new social event document $\mathbf{e}_{new}$, which is composed of many textual words $\mathbf{w}_{new}$ and associated visual words $\mathbf{v}_{new}$, we first obtain the document-topic representation of document $\mathbf{e}_{new}$, then adopt the learned class coefficients μ_t for prediction. The softmax classifier $\mathbf{h}_t(\mathbf{E}_t)$ in the epoch t contains a set of class coefficients $\mu_{t,1:C}$ to summarize the previous tracking data. In the epoch $t+1$, we retrain the softmax classifier $\mathbf{h}_{t+1}$ using the newly obtained event documents $\mathbf{e}_{new}$ and the parameters of the previous epoch μ_t .

D. Multi-Expert Minimization Restoration Scheme

Based on the obtained multiple topic models in different epoches, we utilize a multi-expert minimization restoration scheme [26] to correct the effects of bad model updates in the social tracking process. This strategy considers an entropy-regularized optimization function as our expert selection criterion, which can effectively avoid the topic model drift problem

due to the misalignment and noisy of training samples. We use a learned tracker h_t of the epoch t and its former snapshots to constitute an expert ensemble. Then $\mathbf{E} = \{h_t, h_{t0}, h_{t1}, h_{t2} \cdot \cdot\}$ is an expert ensemble, where $\mathbf{E}$ denotes an expert in the ensemble. As in [26], we determine the best expert by its cumulative loss within a recent temporal window:

$$E^* = \arg\max_{E \in \mathbf{E}} \sum_{k \in [t-\Delta, t]} \wp_E^k, \tag{28}$$

where $\wp_E^t$ denotes the loss value of the expert E, and Δ denotes the size of the temporal windows.

In this paper, we need set a proper loss function in our multi-expert tracking model, and consider the event tracking problem as an extended semi-supervised partial-label learning (PLL) problem [47]. In the traditional PLL problem, given the partially labeled training samples $\Gamma = \{(x_d, z_d)\}$, the PLL can be solved with an MAP framework that maximizes the log posterior probability of the model parameterized by θ:

$$\ell(\theta, \lambda|\Gamma) = L(\theta|\Gamma) - \lambda H_{emp}(Y|X, Z; \Gamma, \theta), \tag{29}$$

where $L(\theta|\Gamma)$ denotes the log likelihood of the model parameters θ, and $H_{emp}(Y|X, Z; \Gamma, \theta)$ denotes the empirical conditional entropy of class labels conditioned on the training data and the possible label sets, we can consider $H_{emp}(Y|X, Z; \Gamma, \theta)$ as an empirical approximation of the logarithm of the prior probability of θ. Here, the MAP framework provides an effective approach to favor models with low ambiguity with respect to the partial label sets by using the entropy regularization term.

In our multi-expert event tracking process, we use the minimum entropy criterion like [26]. Through the online multi-modal tracking model learning, we can obtain the document-topic distributions for each social event over time. At each epoch, given new documents $\{\mathbf{x}, \mathbf{z}\}$, we need predict the category of the document. Here, $\mathbf{x}$ denotes the representative of new documents, and $\mathbf{z}$ denotes a possible label set that encodes the specific constraints of the event tracking problem. According to (29), we set the following loss function for our expert tracking problem (28),

$$\wp_E(\mathbf{x}, \mathbf{z}) = -\ell(\theta, \lambda|\Gamma) = -L(\theta_E|\mathbf{x}, \mathbf{z}) + \lambda H(y|\mathbf{x}, \mathbf{z}; \theta_E), \tag{30}$$

where the log likelihood is defined as:

$$L(\theta_E|\mathbf{x}, \mathbf{z}) = \arg\max_{y \in \mathbf{z}} \log p(y|\mathbf{x}; \theta_E), \tag{31}$$

and the entropy term is defined as:

$$H(y|\mathbf{x}, \mathbf{z}; \theta_E) = \sum_{y \in Y} p(y|\mathbf{x}, \mathbf{z}; \theta_E) \log p(y|\mathbf{x}, \mathbf{z}; \theta_E), \tag{32}$$

where $p(y|\mathbf{x}; \theta_E)$ is the classifier score of the new documents $\mathbf{x}$, and $p(y|\mathbf{x}, \mathbf{z}; \theta_E)$ is defined as:

$$p(y|\mathbf{x}, \mathbf{z}; \theta_E) = \frac{\delta_z(y) p(y|\mathbf{x}; \theta_E)}{\sum_{y' \in Y} \delta_z(y') p(y'|\mathbf{x}; \theta_E)}, \tag{33}$$

where the value $\delta_z(y)$ takes 1 if $y \in z$ and 0 otherwise.

Algorithm 2: Our Proposed Algorithm for Multi-expert Event Tracking

input : Event document streams at epoch t for the event c: $S_{t,c}, t \in \{1, \ldots, T\}, c \in \{1, \ldots, C\}$

output: The class label set of event documents at epoch t: $Y_t, t \in \{1, \ldots, T\}$

for $t = 1$ *to* T **do**
- **If** $t = 1$ **then**
  - Initialize the class property variables Y_1 for all events.
  - Initialize the online multi-modal tracking model for each social event as in Section III-B, and initialize the snapshot of the tracker classifier for each social event as in Section III-C.
- **Else**
  - if mod(t, ξ)=0 then
    - $\mathbf{E} \leftarrow \mathbf{E} \cup \{h_t\}$, and the oldest snapshots will be discarded if the number of experts exceeds N.
  - end if.
  - foreach $E \in \mathbf{E}$
    - compute $\wp_E^t$ in Section III-D.
  - then
    - the best expert will be assigned to the current tracker described in Section III-D.
  - the obtained tracker will output the prediction labels Y_t of new documents S_t.
  - Update the tracker using S_t, and Y_t.
- **end if**

end

E. Social Event Tracking

The whole process of the proposed online multi-modal multi-expert event tracking algorithm is shown in Algorithm 2. In our event tracking algorithm, firstly, we need to initialize the online multi-modal tracking model for each social event described in Section III-B, and initialize the snapshot of the tracker classifier for each social event described in Section III-C. The related experiment settings are shown in Section IV. Secondly, many coming event documents will be determined to which class in the next moment described in Section III-C. Thirdly, we can obtain the classifiers and save a snapshot of the tracker every ξ times, and update the expert ensemble. Note that, the oldest snapshots will be discarded if the number of experts exceeds N. Fourthly, the best expert will be assigned to the current tracker described in Section III-D. We will judge whether the current tracker is the best one, and determine whether restore a tracker or not. Finally, the obtained tracker will output the prediction of new document in next moment. In this way, we can track multi-modal social event documents from multiple events over time.

IV. Experimental Results

In this section, we show extensive experimental results on our collected dataset to demonstrate the effectiveness of our model.

TABLE II
ILLUSTRATION OF THE EVENT NAME, DURATION TIME AND NUMBER OF DOCUMENTS FOR EACH EVENT IN OUR COLLECTED SOCIAL EVENT DATASET

Event ID	Event Name	Start Time	End Time	Dataset	
				#Images	#Text
1	North Korea nuclear program	2004.01	2012.04	9745	3640
2	Greek protests	2011.05	2012.04	9260	1988
3	Mars Reconnaissance Orbiter	2005.04	2012.08	8390	2055
4	Syrian civil war	2011.01	2013.01	5840	1945
5	Senkaku Islands dispute	2008.06	2012.12	5862	1866
6	Occupy Wall Street	2011.09	2012.09	6842	2142
7	United States presidential election	2009.10	2013.01	5890	2001
8	War in Afghanistan	2001.10	2012.08	10425	4489

We first introduce dataset construction and then show feature extraction. Finally, we give results and analysis.

A. Dataset Collection

Nowadays, there are already some public event datasets, such as the MediaEval social event detection(SED) [48]. However, the existing MediaEval SED dataset is organized and attended by people and illustrated by social media content and does not include current hot social events. Moreover, to the best of our knowledge, there are no multi-modality social event datasets available for social event tracking analysis. Therefore, to comprehensively analyze event data with the proposed algorithm, we mainly focus on 8 complex and public social events happened in the past few years, and collect the dataset by ourselves from Google News and Flcikr websites. For these events, we manually create the introduction page of each event or download it from the Wikipedia page,[1] which contains the whole stories of each event. Specifically, the timeline of the event "Senkaku Islands dispute" can be manually obtained based on the content on the page.[2] Then, we manually create the queried keywords of different time for each event based on the timeline of the event. Therefore, the queried keywords are fixed and different in different time for each event. We then search and download related text and its corresponding images from Google News and Flickr websites based on the keywords in the whole timeline of each social event. Here, each social media document is a multimedia document consisting of image and the corresponding text. We adopt some simple rules to delete the unnecessary documents without including the queried keywords of the event, and ensure the reliability of the most of documents. Table II shows the detail of our collected dataset. The collected 8 social events cover a wide range of topics including politics, economics, military, society, and so on. There are about 2000 to 5000 documents with about 30 to 50 epochs for each social event.

[1]http://www.wikipedia.org
[2]http://en.wikipedia.org/wiki/Senkaku_Islands_dispute

B. Feature Extraction

For textual description, we use stemming method and stop words elimination and remove words with a corpus frequency less than 15 in the whole stories of the event, and take the commonly used vector space model to represent the text feature. For visual description, we extract the visual words from images at the region level instead of using low-level feature-based visual descriptors. Previous work [49], [50] has validated that the visual regions can model visual structures and capture interpretable semantics, which is important for illustration of the mined multimodal topics of social event. Therefore, we use regions to represent visual words. In our implementation, each image in the dataset is segmented into 21 regions using rectangular grids with three-level spatial pyramid ($1 \times 1, 2 \times 2$, and 4×4 as in [51]). Each image patch is extracted and represented as 809 dimensional feature vector consisting of color moment, edge histogram, wavelet texture, LBP, and GIST features. Then, we conduct clustering on all regions and apply the affinity propagation algorithm [52] to construct the visual codebook, and each photo is represented using the bag-of-words model with the learned vocabulary.

C. Results and Analysis

In this section, we show experimental results and analysis. In the Section IV-C1, we show quantitative results compared with existing methods. In the Section IV-C2, we give qualitative evaluation of the mined visual and textual topics.

1) Quantitative Evaluation: In this section, we evaluate the effectiveness of online-MMTM model and online multi-modal multi-expert learning (OMMEL) model for social event analysis, respectively.

a) Evaluations of the online-MMTM model: Experimental settings: We evaluate the experimental results of the online MoM-LDA and the proposed online-MMTM, as in [27]. In the online MoM-LDA, since the number of topics is fixed in advance, we vary the number of topics K to be 25, 50, 100, 150, 200. We set the Dirichlet hyperparameters a = 1/K. In the online-MMTM, the hyperparameters are set as $\alpha = 1$, $\gamma = 1$, $\eta = 0.5$, and the corpus-level topic truncation is set as $K = 150$, the document-level truncation is set as $T = 20$. We set $\tau = 64$ and $\kappa = 0.8$ according to the [27]. We choose the best topic number K in the following experiments. We apply the traditional online-LDA [46] strategy to realize the online extension of the MoM-LDA. We measure the performance of online topic modeling by use of the negative predictive likelihood and the soft clustering quality on a held-out test set.

The comparison of the negative predictive likelihood: We utilize the negative predictive likelihood method (similar to the perplexity) to evaluate the proposed model, as in [27]. For each social event, we randomly select 20% of the documents as a held-out test set D_{test} with the remaining documents as training set D_{train}. In the test step, the test document d is split into two parts d_1 and d_2. The predictive distribution of words is estimated by the first part of the test data and the training data. Then the negative predictive likelihood of each word in the second part of test document can be estimated via the obtained predictive

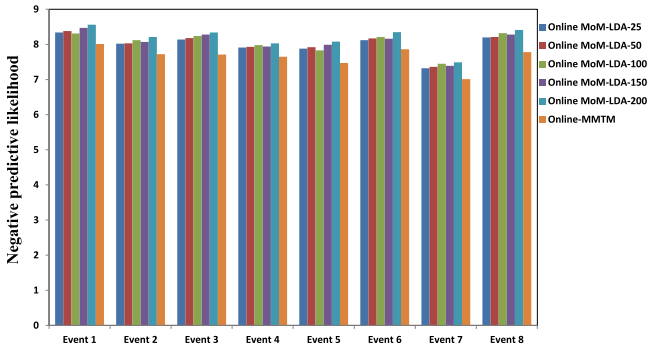


Fig. 4. The negative predictive likelihood values for different topic models on our collected dataset.

distribution of words. The lower negative predictive likelihood value shows the better performance, and it is defined as

$$\text{likelihood}_{\text{pre}} = -\frac{\sum_{d\in D_{\text{test}}} \log p(d_2|d_1, D_{\text{train}})}{\sum_{d\in D_{\text{test}}} |d_2|} \tag{34}$$

where

$$p(d_2|d_1, D_{\text{train}}) = \prod_{w\in d_2} \sum_{k} \bar{\pi}_{d_2 k} \bar{\Phi}_{kw} \tag{35}$$

In (34) and (35), $|d_2|$ represents the number of words in d_2, $\bar{\Phi}$ represents the variational expectation of the topics Φ in the training data D_{train}, and $\bar{\pi}$ represents the variational expectation of document weights in the test data d_2. Fig. 4 shows the negative predictive likelihood values for both models.

The comparison of the soft clustering quality: The soft clustering quality method is adopted to measure the quality of identifying the topic clusters on the test set [53]. Since the proposed model is actually a soft clustering technique, for each document including images and corresponding text, the document's topic with the highest probability is selected as its topic cluster. The intuition is that the topic model learns better if the predicted document's topic clusters are more like the human labeled results. In the experiment, we manually build a set of true topic clusters for each event as the ground truth for evaluation, where a cluster is made up of many documents that describe common topics. We invited four subjects who were graduates and familiar to the studied entities. Each annotator was assigned two social event categories. Note that, since it is time-consuming to label a large amount of data, we only randomly sample 600 documents for each event from the collected dataset. For each social event, we hold out 80% of the documents for training. For each event, many documents regarding different topics were selected from the collected dataset. Documents that describe common topics are grouped into one cluster, which describes one certain property of the social event. Table III shows the statistics of the built topic clusters for each event. We utilize the purity [53] as the evaluation metric. The purity is a standard measure of clustering quality in tradition clustering problem. For each event, let $S = \{s_1, \ldots, s_J\}$ denote the mined topics, and $\widehat{S} = \{\widehat{s}_1, \ldots, \widehat{s}_K\}$ denote its manually labeled topic groups. More formally, the purity metric is defined as

$$Purity(S, \widehat{S}) = \frac{1}{N} \sum_{j} \max_{K} |\{s_j \cap \widehat{s}_k|, \tag{36}$$

where N is the total number of topics in the ground truth dataset, and a higher purity score means a better performance. Fig. 5 shows the purity scores for both models.

Based on the comparison results of the soft clustering quality and the negative predictive likelihood as shown in Figs. 4 and 5, we can make the following conclusions.

1) The online MoM-LDA-25 shows inferior performance in the soft clustering quality. This is because when the number of topics is too small, the result suffers from underfitting.
2) The online MoM-LDA method will become better with the topic number increasing. However, blindly increasing the topic numbers could make the computation cost extremely high without further improving the performance. Typically, researchers could find the "best" number of topics with cross-validation [36]. This approach will become not practical on a very large dataset.
3) The proposed online-MMTM consistently and significantly outperforms the online MoM-LDA. The major reason is that the proposed online-MMTM is a nonparametric approach, which can automatically learn the number of topics from data over time. We observe that the online-MMTM uses about 120 and 110 topics out of its potential 150 in two experiment evaluations: the soft clustering quality and the negative predictive likelihood, respectively. Especially, the non-parametric Dirichlet process variables make the online-MMTM have the ability that certain topics have a higher priority to appear than others, whereas the exchangeable Dirichlet prior used in online MoM-LDA assumes that all topics are equally common.

The comparison of the runtime efficiency: We propose a novel online multi-modal tracking algorithm to model time-series multi-modal social event data. Different from the traditional MoM-HDP model, the proposed online-MMTM can work in an online mode. Fig. 6 shows the comparison of the runtime efficiency for the MoM-HDP and online-MMTM to online train the model at each epoch of social event. It is clear that the online-MMTM stores only the content of each document in current time and requires approximately a constant time in different epoches, while the MoM-HDP requires the whole data to be stored to train the model and costs more time. The online-MMTM is more superior than the MoM-HDP in terms of time and memory efficiency. As a result, the online-MMTM allows for closed-form coordinate ascent variational inference, which is a key factor in developing the online algorithm. In this paper, we use this online mode to update the model parameters for each document of each epoch in sequential social event data.

b) Evaluations on event tracking: To evaluate different event tracking methods, we adopt the popular mean average precision (MAP) as a metric. Because this is a tracking problem, the supervised labels are only obtained to initialize the model parameters

TABLE III
STATISTICS OF THE BUILT TOPIC CLUSTERS

	Event 1	Event 2	Event 3	Event 4	Event 5	Event 6	Event 7	Event 8
#clusters	9	8	8	6	9	6	8	9
#document per cluster	30∼100	20∼100	30∼100	30∼100	30∼100	20∼100	30∼100	20∼100
#total documents	600	600	600	600	600	600	600	600

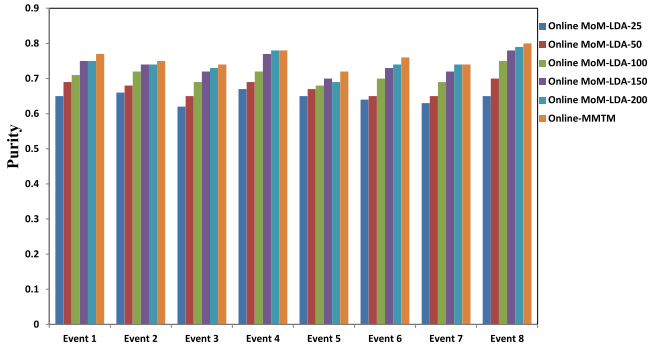


Fig. 5. The purity scores of topic identification for different topic models on our collected dataset.

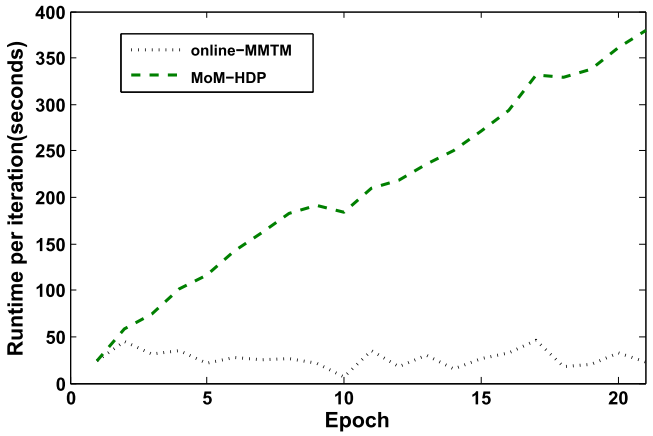


Fig. 6. The comparisons of the runtime efficiency required to train the MoM-HDP and the online-MMTM model for the evolution of event "North Korea nuclear program".

at the first epoch of each social event, and we do not know the label information in the rest of time. Therefore, we just need to know the labeled data of the first epoch for each event. In the experiment, we manually label ground-truth tracks for all the events, which can be obtained when we download the data by the corresponding queried keywords. After event tracking, we know the associated label for each document of the story. Based on the tracking results, we can calculate the average precision for each event. Finally, we can calculate the mean of the average precisions of multiple events and obtain the MAP.

To demonstrate the effectiveness of our online multi-modal multi-expert learning (OMMEL) algorithm for event tracking analysis, we compare it with the most related baseline methods including BOW, online-LDA, online-HDP, online-MoM-LDA, mmETM, online-MMTM.

- BOW: This is a traditional Bag-of-Words method by concatenating the textual and visual features.
- online-LDA [46]: This is an online-LDA extension by concatenating the textual and visual features.
- online-HDP [36]: This is an online-HDP extension by concatenating the textual and visual features.
- online-MoM-LDA [19]: This model extends LDA for multi-modal data to represent the joint distribution of images and text by using textual and visual data jointly.
- mmETM [21]: This model extends MoM-LDA to learn correlations between textual and visual modalities to separate the visual-representative topics and non-visual-representative topics.
- online-MMTM: This model is a novel online multi-modal tracking model (online-MMTM) based on the traditional MoM-HDP model by using a novel online variational inference algorithm.

Note that, we apply our tracking strategy to implement all baseline methods, but they do not adopt the multi-expert minimization restoration scheme. In the experimental setting of social event tracking, for the online-LDA, online-MoM-LDA, and mmETM model, we vary the different topic numbers and choose the best tracking results. The online-HDP and online-MMTM can automatically infer the number of topics needed, and we observe that the online-HDP and online-MMTM use about 100 and 120 topics in the experiments. The same experimental settings are adopted as in Section IV-C1a. Table IV and Fig. 7 show the comparisons of different models for multi-event tracking. Based on these results, we have the following conclusions.

1) The BOW, online-LDA and online-HDP show much worse tracking performance than other multi-modal topic models. This is due to their incapability of learning the latent association relationships between textual and visual topics. The BOW shows much better performance than the online-LDA and online-HDP, and it might be because the high dimension feature of the BOW can be effective for social event tracking.
2) The online-MoM-LDA, mmETM and online-MMTM achieve better tracking results. In the multi-modal event tracking, the online-MoM-LDA, mmETM and online-MMTM are effective to capture the intrinsic relationships between text words and visual words. However, the process of online model updating usually has the model drift problem due to the misalignment and noisy of event documents, which makes these models have limited performance.
3) The proposed OMMEL consistently outperforms other existing state-of-the-art methods, which is consistent with

TABLE IV
MEAN AVERAGE PRECISION ON EVENT DATASET FOR DIFFERENT MODELS

Model	BOW	online-LDA	online-HDP	online-MoM-LDA	mmETM	online-MMTM	**OMMEL**
MAP	0.62	0.57	0.61	0.66	0.69	0.70	**0.73**

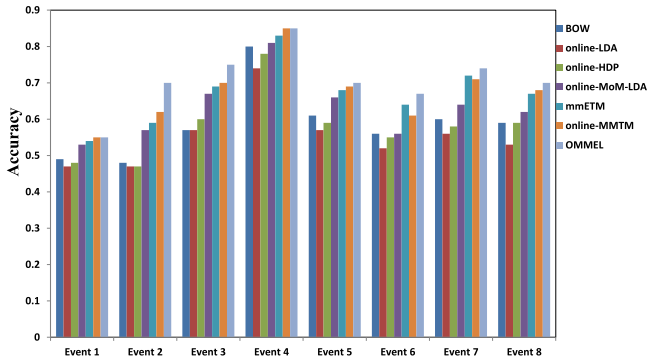


Fig. 7. The accuracy of different event tracking methods on our collected dataset. Compared with the existing methods, our OMMEL achieves the best.

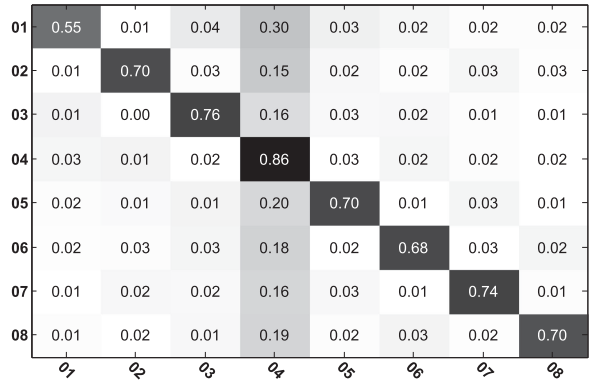


Fig. 9. The confusion matrix of our tracking methods on our dataset.

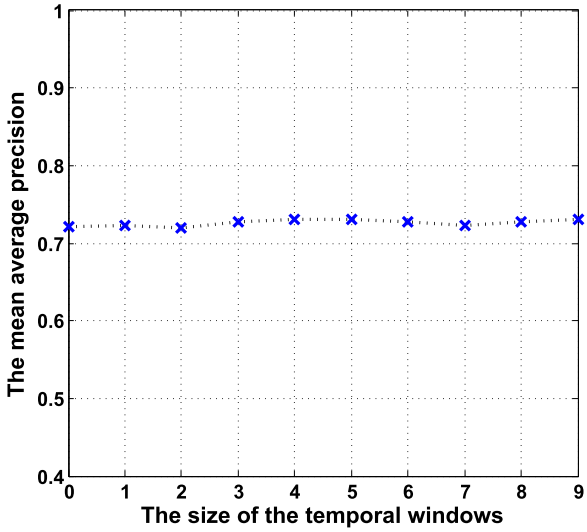


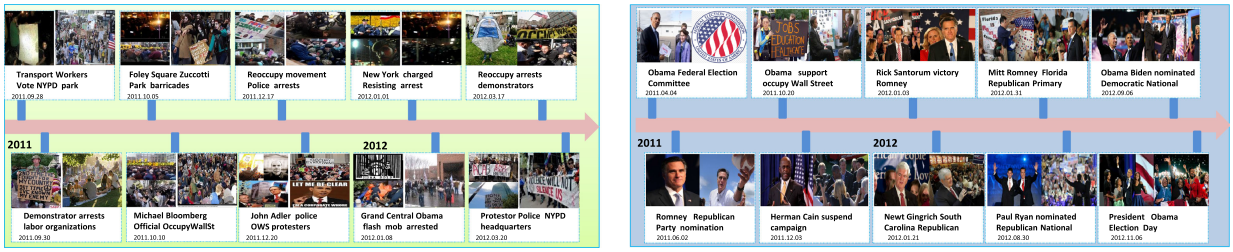
Fig. 8. The mean average precision with different sizes of the temporal windows.

the experimental results of online-MMTM in Section IV-C1a. This is because our model uses a novel online multi-modal tracking model (online-MMTM) to model multi-modal social event documents for event tracking and evolution analysis, and utilizes a novel multi-expert minimization restoration scheme to correct the effects of bad model updates by considering an entropy-regularized optimization function as our expert selection criterion and online improve the social event model.

In Fig. 8, we show the mean average precision with different sizes of the temporal window, and can observe that the performance of social event tracking tends to be insensitive to the size of the temporal window. We set $\Delta = 3$ in the experiment. In Fig. 9, we show the confusion matrix to visualize the tracking performance of each event on the dataset. Each row of the matrix represents the actual class of instances while each column of the matrix represents the predicted class of instances. The (i, j) value of the confusion matrix shows that the percentage value of the i-th category is classified to the j-th category. Based on the confusion matrix, we can clearly observe the effects of different event classes. From the results, we can see that some events have bad tracking results because they are very similar to other events. For example, the event "Greek protests" and "Occupy Wall Street" are confused, because the two events have some similar topics, such as, economy crisis, demonstration and government officials.

2) *Qualitative Evaluation*

In this section, we will qualitatively demonstrate the effectiveness of the proposed online-MMTM for social event analysis. For the multi-modal event evolution and visualization, we can mine and update representative topics over time by the proposed online-MMTM. Given a time-order event in a time span, we can mine the event theme patterns and visualize the events' multi-modal topic information over time in order to get their evolutionary trends. Fig. 10 shows the mined multi-modal topics over time for the event (a) "Occupy Wall Street" and the event (b) "United States presidential election". With the discovered multi-modal topics, represented by textual words and photos, we organize the topics and present the summary in the way of timeline for each social event. From such summary of the timeline, users can easily know what happened and the topic evolution of social events over time. From the Fig. 10, we can see that the event "United States presidential election" has some keywords such as " Obama, support, occupy, Wall, Street" in October 20, 2011, and they are very related to what happened in October and November for the event "Occupy Wall Street".



(a) (b)

Fig. 10. The timeline for the event (a) "Occupy Wall Street" and the event (b) "United States presidential election". Based on the results, it is clear and easy for users to know what happened and the topic evolution of social events over time.

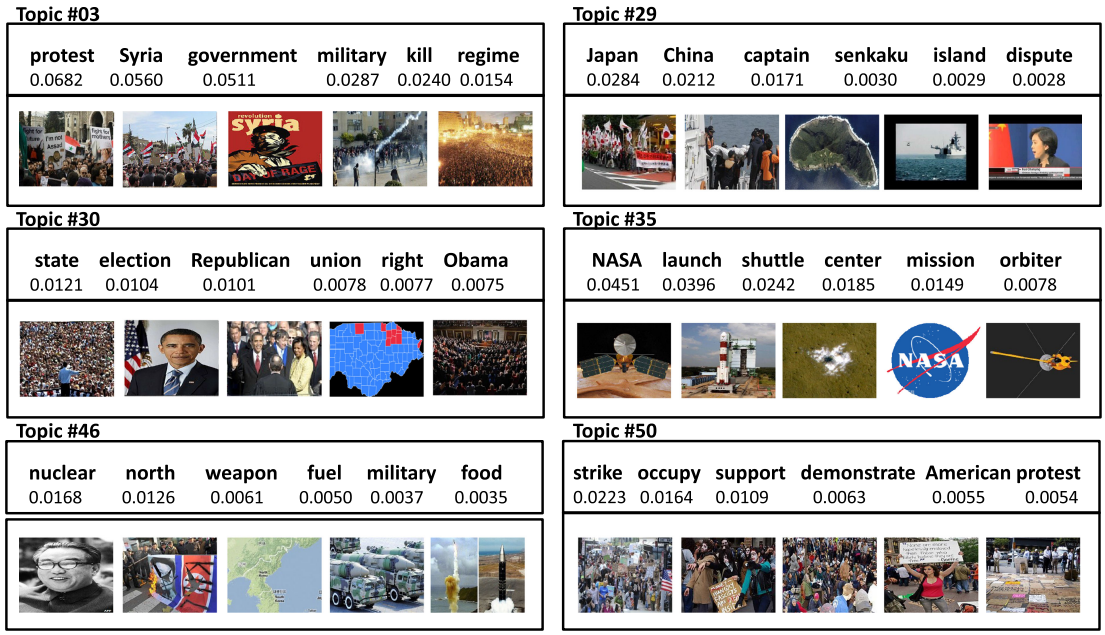


Fig. 11. Illustration of samples of discovered textual and visual topics. Here, we show six topics with their top six textual words and the five most related images.

For simplicity, we then visualize the learned textual and visual topics in the first epoch of social event in Fig. 11, which can validate the effectiveness of the proposed model for topic learning. We use the proposed model in all social events to learn the multi-modal topics. In Fig. 11, we show an example of the discovered six topics with their top six textual words and the five most related images. By providing the multi-modal information of the representative textual and visual words, it is very intuitive to interpret the social events with each associated topic. For the visualization of learned topics, we present the top-ranked textual words and visual patches for each topic. The textual words and visual patches are sorted by the probability $p(w|z)$, $p(v|z)$, respectively. As shown in Fig. 11, the results are impressive and satisfy our expectation, where most of the extracted event topics are meaningful and textual words are well aligned with the corresponding visual image content. For example, "topic #03" shows the top six textual words and the five most related image patches. Specifically, textual words, such as protest, Syria, government, military, kill, regime, are shown according to the rank of their probability values in "topic #03". Similarly, image patches are shown. Based on the results, we can confirm that our model can effectively mine the topics of social events.

V. Conclusion

In this paper, we propose a novel online multi-modal multi-expert learning algorithm to obtain informative summary details and the topic evolution of social events over time. The proposed tracking algorithm utilizes a novel multi-expert minimization restoration scheme to correct the effects of bad model updates by using an entropy-regularized optimization function as our expert selection criterion and online improve the social event model. Moreover, the proposed online multi-modal tracking model is a non-parametric approach, which is able to not only automatically learn the number of topics from data over time, but also can exploit the multi-modal property of social event. We have conducted experiments on our collected dataset and extensive results have demonstrated that our model outperforms all other existing models. In the future, we will

investigate event summarization and event detection in social media and combine them to build a real system.

APPENDIX
DERIVATION OF ONLINE MULTI-MODEL TRACKING MODEL

For the document-level parameters, the updates are:

$$a_{dt} = 1 + \left(\sum_n \varsigma_{dnt}^w + \sum_n \varsigma_{dnt}^v \right) \tag{37}$$

$$b_{dt} = \alpha + \left(\sum_n \sum_{s=t+1}^{T} \varsigma_{dns}^w + \sum_n \sum_{s=t+1}^{T} \varsigma_{dns}^v \right) \tag{38}$$

$$\varphi_{dtk} \propto \exp \left(\mathbf{E}_q[\log \beta_k] + \left(\sum_n \varsigma_{dnt}^w \cdot \mathbf{E}_q[\log(p(w_{dn}|\Phi_k^w))] + \sum_n \varsigma_{dnt}^v \cdot \mathbf{E}_q[\log(p(v_{dn}|\Phi_k^v))] \right) \right) \tag{39}$$

$$\varsigma_{dnt}^w \propto \exp \left(\mathbf{E}_q[\log \pi_{dt}] + \sum_{k=1}^{K} \varphi_{dtk} \cdot \mathbf{E}_q[\log(p(w_{dn}|\Phi_k^w))] \right) \tag{40}$$

$$\varsigma_{dnt}^v \propto \exp \left(\mathbf{E}_q[\log \pi_{dt}] + \sum_{k=1}^{K} \varphi_{dtk} \cdot \mathbf{E}_q[\log(p(v_{dn}|\Phi_k^v))] \right) \tag{41}$$

For the corpus-level parameters about the topic-level stick-breaking proportions and the topics, the updates are:

$$u_k = 1 + \sum_d \sum_{t=1}^{T} \varphi_{dtk} \tag{42}$$

$$v_k = \gamma + \sum_d \sum_{t=1}^{T} \sum_{l=k+1}^{K} \varphi_{dtl} \tag{43}$$

$$\lambda_k^w = \eta^w + \sum_d \sum_{t=1}^{T} \varphi_{dtk} \left(\sum_n \varsigma_{dnt}^w I[w_{dn} = w] \right) \tag{44}$$

$$\lambda_k^v = \eta^v + \sum_d \sum_{t=1}^{T} \varphi_{dtk} \left(\sum_n \varsigma_{dnt}^v I[v_{dn} = v] \right) \tag{45}$$

Here, the expectations involved above are taken under the variational distribution q,

$$\mathbf{E}_q[\log \beta_k] = \mathbf{E}_q[\log \beta_k'] + \sum_{l=1}^{k-1} \mathbf{E}_q[\log(1 - \beta_l')] \tag{46}$$

$$\mathbf{E}_q[\log \beta_k'] = \Psi(u_k) - \Psi(u_k + v_k) \tag{47}$$

$$\mathbf{E}_q[\log(1 - \beta_k')] = \Psi(v_k) - \Psi(u_k + v_k) \tag{48}$$

$$\mathbf{E}_q[\log \pi_{dt}] = \mathbf{E}_q[\log \pi_{dt}'] + \sum_{s=1}^{t-1} \mathbf{E}_q[\log(1 - \pi_{dt}')] \tag{49}$$

$$\mathbf{E}_q[\log \pi_{dt}'] = \Psi(a_{dt}) - \Psi(a_{dt} + b_{dt}) \tag{50}$$

$$\mathbf{E}_q[\log(1 - \pi_{dt}')] = \Psi(b_{dt}) - \Psi(a_{dt} + b_{dt}) \tag{51}$$

$$\mathbf{E}_q[\log(1 - \pi_{dt}')] = \Psi(b_{dt}) - \Psi(a_{dt} + b_{dt}) \tag{52}$$

$$\mathbf{E}_q[\log(p(w_{dn} = w'|\Phi_k^w))] = \Psi(\lambda_{kw'}^w) - \Psi\left(\sum_{w'} \lambda_{kw'}^w \right) \tag{53}$$

$$\mathbf{E}_q[\log(p(v_{dn} = v'|\Phi_k^v))] = \Psi(\lambda_{kv'}^v) - \Psi\left(\sum_{v'} \lambda_{kv'}^v \right) \tag{54}$$

where $\Psi(\cdot)$ is the digamma function.

REFERENCES

[1] Y. Yang *et al.*, "Learning approaches for detecting and tracking news events," *IEEE Intell. Syst.*, vol. 14, no. 4, pp. 32–43, Jul./Aug. 1999.

[2] K. Radinsky and E. Horvitz, "Mining the web to predict future events," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, 2013, pp. 255–264.

[3] D. Patel, W. Hsu, and M. L. Lee, "Mining relationships among interval-based events for classification," in *Proc. ACM Int. Conf. Manag. Data,* 2008, pp. 393–404.

[4] S. Qian, T. Zhang, and C. Xu, "Multi-modal multi-view topic-opinion mining for social event analysis," in *Proc. ACM Conf. Multimedia*, Amsterdam, The Netherlands, Oct. 15–19, 2016, pp. 2–11.

[5] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1998, pp. 37–45.

[6] G. Kumaran and J. Allan, "Text classification and named entities for new event detection," in *Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2004, pp. 297–304.

[7] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic model vectors for complex video event recognition," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 88–101, Feb. 2012.

[8] T. Zhang and C. Xu, "Cross-domain multi-event tracking via CO-PMHT," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 10, no. 4, 2014, Art no. 31.

[9] S. Qian, T. Zhang, C. Xu, and M. S. Hossain, "Social event classification via boosted multimodal supervised latent Dirichlet allocation," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 11, no. 2, 2015, Art no. 27.

[10] L. Xie *et al.*, "Discovering meaningful multimedia patterns with audio-visual concepts and associated text," in *Proc. Int. Conf. Image Process.*, 2004, pp. 2383–2386.

[11] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 113–120.

[12] X. Wang, N. Mohanty, and A. McCallum, "Group and topic discovery from relations and text," in *Proc. 3rd Int. Workshop Link Discovery*, 2005, pp. 28–35.

[13] X. Wu, C.-W. Ngo, and A. G. Hauptmann, "Multimodal news story clustering with pairwise visual near-duplicate constraint," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 188–199, Feb. 2008.

[14] I. Kalamaras, A. Drosou, and D. Tzovaras, "Multi-objective optimization for multimodal visualization," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1460–1472, Aug. 2014.

[15] X. Yang, T. Zhang, and C. Xu, "Cross-domain feature learning in multimedia," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 64–78, Jan. 2015.

[16] Y. Yang, J. Zhang, J. Carbonell, and C. Jin, "Topic-conditioned novelty detection," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 688–693.

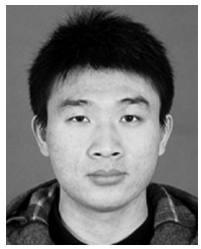
[17] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi, "Simple semantics in topic detection and tracking," *Inf. Retrieval*, vol. 7, no. 3/4, pp. 347–368, 2004.

[18] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2003, pp. 127–134.

[19] K. Barnard *et al.*, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, 2003.

[20] J. Bian, Y. Yang, H. Zhang, and T. Chua, "Multimedia summarization for social events in microblog stream," *IEEE Trans. Multimedia*, vol. 17, no. 2, pp. 216–228, Feb. 2015.

[21] S. Qian, T. Zhang, C. Xu, and J. Shao, "Multi-modal event topic model for social event analysis," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 233–246, Feb. 2016.
[22] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Stat. Assoc.*, vol. 101, no. 476, pp. 1566–1581, 2006.
[23] O. Yakhnenko and V. Honavar, "Annotating images and image objects using a hierarchical Dirichlet process model," in *Proc. 9th Int. Workshop Multimedia Data Mining: Held Conjunction ACM SIGKDD*, 2008, pp. 1–7.
[24] I. A. Matthews, T. Ishikawa, and S. Baker, "The template update problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 6, pp. 810–815, Jun. 2004.
[25] Z. Kalal, J. Matas, and K. Mikolajczyk, "P-N learning: Bootstrapping binary classifiers by structural constraints," in *Proc. 23rd IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 13–18, 2010, pp. 49–56.
[26] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: Robust tracking via multiple experts using entropy minimization," in *Proc. 13th Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Sep. 6–12, 2014, pp. 188–203.
[27] C. Wang, J. W. Paisley, and D. M. Blei, "Online variational inference for the hierarchical Dirichlet process," in *Proc. 14th Int. Conf. Artif. Intell. Stat.*, Fort Lauderdale, FL, USA, Apr. 11–13, 2011, pp. 752–760.
[28] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, "Diamonds in the rough: Social media visual analytics for journalistic inquiry," in *Proc. IEEE Symp. Vis. Analytics Sci. Technol.*, 2010, pp. 115–122.
[29] H. Becker, M. Naaman, and L. Gravano, "Event identification in social media," in *Proc. 3rd ACM Int. Conf. Web Search Data*, 2009, pp. 291–300.
[30] C. S. Firan, M. Georgescu, W. Nejdl, and R. Paiu, "Bringing order to your photos: Event-driven classification of Flickr images based on social knowledge," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage.*, 2010, pp. 189–198.
[31] J. Allan, C. Wade, and A. Bolivar, "Retrieval and novelty detection at the sentence level," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2003, pp. 314–321.
[32] X. Yang, T. Zhang, C. Xu, and M. S. Hossain, "Automatic visual concept learning for social event understanding," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 346–358, Mar. 2015.
[33] J. R. Kender and M. R. Naphade, "Visual concepts for news story tracking: Analyzing and exploiting the NIST TRECVID video annotation experiment," in *Proc. Comput. Vis. Pattern Recognit.*, 2005, pp. 1174–1181.
[34] Y. Zhai and M. Shah, "Tracking news stories across different sources," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, 2005, pp. 2–10.
[35] T. Zhang, S. Liu, N. Ahuja, and M.-H. Yang, "Robust visual tracking via consistent low-rank sparse learning," *Int. J. Comput. Vis.*, vol. 111, no. 2, pp. 171–190, 2015.
[36] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[37] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. 15th Conf. Uncertainty Artif. Intell*, 1999, pp. 289–296.
[38] D. Blei and J. McAuliffe, "Supervised topic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 121–128.
[39] S. Lacoste-Julien, F. Sha, and M. I. Jordan, "DiscLDA: Discriminative learning for dimensionality reduction and classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 897–904.
[40] Z. Niu, G. Hua, X. Gao, and Q. Tian, "Spatial-DiscLDA for visual recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2011, pp. 1769–1776.
[41] D. Ramage, P. Heymann, C. D. Manning, and H. Garcia-Molina, "Clustering the tagged web," in *Proc. 2nd ACM Int. Conf. Web Search Data Mining*, 2009, pp. 54–63.
[42] J. Sang and C. Xu, "Right buddy makes the difference: An early exploration of social relation analysis in multimedia applications," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 19–28.
[43] X. Wang and A. McCallum, "Topics over time: A non-Markov continuous-time model of topical trends," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 424–433.
[44] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Ann. Stat.*, vol. 1, no. 2, pp. 209–230, 1973.
[45] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, pp. 5228–5235, 2004.
[46] M. D. Hoffman, D. M. Blei, and F. R. Bach, "Online learning for latent Dirichlet allocation," in *Proc. 23rd Int. Conf. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2010, pp. 856–864.
[47] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. 17th Int. Conf. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 13–18, 2004, pp. 529–536.
[48] T. Reuter *et al.*, "Social event detection at mediaeval 2013: Challenges, datasets, and evaluation," in *Proc. MediaEval Multimedia Benchmark Workshop*, Barcelona, Spain, Oct. 18–19, 2013, pp. 1–2.
[49] Q. Fang, J. Sang, and C. Xu, "Giant: Geo-informative attributes for location recognition and exploration," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 13–22.
[50] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 73–86.
[51] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 17–22, 2006, pp. 2169–2178.
[52] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.
[53] Y. Zhao and G. Karypis, "Empirical and theoretical comparisons of selected criterion functions for document clustering," *Mach. Learn.*, vol. 55, pp. 311–331, 2004.



Shengsheng Qian received the B.E. degree from the Jilin University, Changchun, China, in 2012, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2017. He is currently an Assistant Professor with the Institute of Automation, Chinese Academy of Sciences. His current research interests include social media data mining and social event content analysis.

Tianzhu Zhang (S'09–M'11) received the B.S. degree in communications and information technology from the Beijing Institute of Technology, Beijing, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011. He is currently an Associate Professor with the Institute of Automation, Chinese Academy of Sciences. His current research interests include computer vision and multimedia, especially action recognition, object classification, and object tracking.

Changsheng Xu (M'97–SM'99–F'14) is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China and the Executive Director of the China-Singapore Institute of Digital Media, Singapore. He holds 30 granted/pending patents and published more than 200 refereed research papers in these areas. His research interests include multimedia content analysis/indexing/retrieval, pattern recognition and computer vision. Dr. Xu is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, the *ACM Transactions on Multimedia Computing, Communications, and Applications*, and the ACM/Springer *Multimedia Systems* Journal. He was the recipient of the Best Associate Editor Award for the *ACM Transactions on Multimedia Computing, Communications, and Applications* in 2012 and the Best Editorial Member Award for the ACM/Springer *Multimedia Systems* Journal in 2008. He was a Program Chair of ACM Multimedia 2009. He has been an Associate Editor, Guest Editor, General Chair, Program Chair, Area/Track Chair, Special Session Organizer, Session Chair, and TPC member for more than 20 IEEE and ACM prestigious multimedia journals, conferences, and workshops. He is the IAPR Fellow and ACM Distinguished Scientist.