Journal of Multimedia Information System: ISSN 2383-7632(Online)
VOL. 1, NO. 1, September 2014, pp. 55-67

Jointly Image Topic and Emotion Detection using Multi-Modal Hierarchical Latent Dirichlet Allocation

Author: Wanying Ding[1,*], Junhuan Zhu[2], Lifan Guo[3], Xiaohua Hu[1], Jiebo Luo[2], Haohong Wang[3]

Abstract

Image topic and emotion analysis is an important component of online image retrieval, which nowadays has become very popular in the widely growing social media community. However, due to the gaps between images and texts, there is very limited work in literature to detect one image's *Topics* and *Emotions* in a unified framework, although topics and emotions are two levels of semantics that often work together to comprehensively describe one image. In this work, a unified model, Joint Topic/Emotion Multi-Modal Hierarchical Latent Dirichlet Allocation (JTE-MMHLDA) model, which extends previous LDA, mmLDA, and JST model to capture topic and emotion information at the same time from heterogeneous data, is proposed. Specifically, a two level graphical structured model is built to realize sharing topics and emotions among the whole document collection. The experimental results on a Flickr dataset indicate that the proposed model efficiently discovers images' topics and emotions, and significantly outperform the text-only system by 4.4%, vision-only system by 18.1% in topic detection, and outperforms the text-only system by 7.1%, vision-only system by 39.7% in emotion detection.

Key Words: Hierarchical Latent Dirichlet Allocation, Multi-Modal Framework, Emotion Detection, Topic Detection

I. INTRODUCTION

With the emerging social media websites, like Flickr, YouTube, and Facebook, huge number of images are published and shared online. The image is becoming a rich and important information source for people's daily life. Images can not only express a specific topic, but also convey emotions to evoke viewers. Topics and emotions are two levels of semantics of one image, and they work together to deliver the whole meaning of the image. Thus, it will highly improve online image retrieval's performance if one can help to detect topics and emotions from one image simultaneously. However, with the online information explosion phenomenon, it is impossible to manually complete such a task, and automatically image topic/emotion detection has turned out to be an important and interesting research hotspot.

Generally speaking, a *Topic* can be defined as the object that an image describes, like a flower, a building or a person. An *Emotion* is the subjectivity an image conveys, like happy, sad or satisfied. Emotion analysis is closely related to sentiment analysis, which has been deeply explored in text mining area. The difference between the emotion and sentiment is that emotions are more specific. Classically, there are only three types of sentiments, namely positive, negative and neutral, and each type of sentiment contains various emotions, for example, positive sentiment contains emotions like happiness, interested, trust, and so on, and negative sentiment contains sadness, grief and rage. Although opinion mining and sentiment analysis has been intensively studied in text mining, limited similar works have been done in computer vision field. But, with the development of multimedia, more and more people choose to use image, which are more infectious and vivid, to express their opinions and emotions. Text analysis alone is becoming insufficient to cope with the huge influx of images online.

Topics and emotions are two semantic levels to interpret an image[1]. Topic detection is a hot research spot, and related methods can be grouped into two categories. The most commonly used one is the classification models[2-4]. After feature extraction and representation, researchers

Manuscript Received August 02, 2014; Revised September 03, 2014; Accepted September 20, 2014. (ID No. JMIS-2014-0005)

Corresponding Author(*): Wanying Ding, Drexel University, Tel:215-588-7328, E-mail:wd78@drexel.edu
[1]College of Computing and Informatics, Drexel University, Philadelphia, PA, USA, 19104
[2]Department of Computer Science, University of Rochester, Rochester, NY, USA, 14627
[3]TCL Research America, San Jose, CA, 95134

apply some classical classification models, such as Support Vector Machine (SVM), to group the images, and then find the topics for each group. Another method is the probabilistic latent variable models [5-8]. Latent Dirichlet Allocation (LDA)[9] and its extension models are widely used. Compared to the first approach, the advantage of LDA extended models is that they are unsupervised and do not need training dataset, so it is more flexible in use.

Despite its popularity, most emotion detection works focus on recognizing emotions from facial expressions[10], butthe perception of emotions from a non-face image is still an open area to explore. The other pertinent researches[11-15]are about to use low-level vision features alone to conduct the emotions analysis, and the results are always not that satisfying.

In summary, two problems exist in this area. First, many works just explore only topics or only emotions from an image. Topics and emotions are one image's two levels of semantics, and one of them alone could not make a comprehensive explanation about that image. Second, most researches only consider text or vision features, and ignore the other one, but both of them contribute to the final result. Common sense tells us the more useful information we use, the higher possibility to get an accurate result. All the available information should be well exploited but ignored.

To solve the above problems, this paper proposes a Joint Topic/Emotion Multi-Modal Hierarchical LDA model (JTE-MMHLDA) to detect images' topics and emotions in a unified framework. Specifically, JTE-MMHLDA uses multi-level LDA to handle the joint topic/emotion detection problem. First, JTE-MMHLDA exploits both text and vision features, enriching the feature collection to get a better result. Second, to make a global view of the results, JTE-MMHLDA implements global and shared topic/emotion distributions, making the results more interpretable and comparable. Third, JTE-MMHLDA uses two kinds of hidden variables to function, one is for the topic distribution, and the other is for the emotion distribution, realizing topic/emotion detection in a uniform framework.

We test our model on an image collection with 36765 images extracted from Flickr, and the result shows that JTE-MMHLDA outperforms the text-only system by 4.4%, vision-only system by 18.1% in topic detection, and outperforms the text-only system by 7.1%, vision-only system by 39.7% in emotion detection.

The rest of the paper is organized as follows. Section 2 gives a brief introduction to the related models and works. Section 3 describes the JTE-MMHLDA in detail. Section 4 presents the experiment process and results. Finally we give a conclusion and future work description in Section 5.

II. RELATED WORK

1. Multi-Model Topic Model

Once Latent Dirichlet Allocation (LDA)[16]was proposed, this model has been widely applied to computer vision field. A lot of image classification and topic detection models are LDA extended [6-8, 17, 18]. The basic idea behind LDA is that it assumes each image has a distribution on the latent topics, and each topic has a distribution on the features. LDA utilizes the feature co-occurrence phenomenon to group images sharing similar feature vectors together. LDA discovers topics in an unsupervised fashion, so it is flexible to be widely applied.

Even though plain LDA has helped a lot in image topic detection, it can only handle one instance (one kind of feature). Some researchers argue that in the real world, one image might be associated with multiple instances[19]. One image not only has vision features, but also the text features, like descriptions, title and tags, around it. But plain LDA will be insufficient in handling with such heterogeneous features. Thus, Multi-Modal LDA (mmLDA)[20] is proposed to deal with such a problem. MmLDA can simultaneously exploit both vision and text features to generate topic distributions based on both of these two kinds of features. M3LDA[21] is a good example in extending mmLDA. M3LDA consists three parts of LDA, one is to process vision features, one is to process text features, and one is used to combine these two parts. Another application of mmLDA is the mm-SLDA, which combines two supervised LDA (sLDA)[22] together. The labels in mm-SLDA are shared by all the images and act as the bridge between the text features and vision features. Topic Regression Multi-Modal LDA (tr-mmLDA)[5] replaces the concrete and independent latent variables with a latent variable regression approach to correlate the latent variables of the two modalities.

Multi-Modal LDA has been broadly studied in image classification area. Most of the researches focus on how to design each modality or how to combine different modalities, but the number of hidden variables they use is just one. One hidden variable assumption does not always function in the real world. The topic and emotion are two hidden variables embedded in one image, and thus at least two hidden variables are in need. Taking images in Fig. 1 as an example, Figure 1(a) and Figure 1 (b) share the same topic "flower", but they have different emotions. Figure 1(a) and Figure 1(c) share the same emotion, like "happy", but have very different topics. Most current mmLDA extended models could detect that Fig. 1(a) and Fig. 1(b) describe the same topic, but could fail to distinguish the different emotions they have.

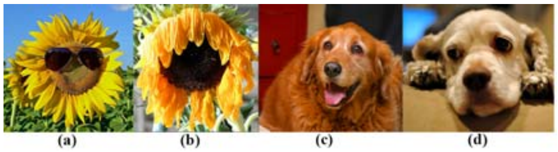


Fig. 1. Images share topics but have different emotions.

2. Sentiment/Emotion Analysis

Sentiment analysis is an important research topic in text mining area. In the beginning, a lot of related researches are based on lexicon or ontology construction methods[23, 24], but since the proposition of LDA, a lot of LDA extended models have been built to automatically recognize the sentiment distribution from a message. Joint Sentiment Topic Model(JST)[25] is first model to introduce the sentiment hidden variable into LDA. This paper has inspired a lot of text mining researchers to use the double hidden variable models in sentiment analysis. Similar models are MaxEnt-LDA[26], ASUM[27] and JAS[28], and all of them work very well on the text.

Although comparing the achievement of sentiment analysis in text analysis, the one for image analysis falls far behind, some efforts still have been made and inspire a lot. Jana Machajdik and Allan Hanbury[14] identified a series of features, including color and texture, which are helpful for image sentiment detection, and their work laid a solid foundation for feature definition in image emotion analysis. Damian Borth et al.[29] introduced the Adjective Noun Pairs (ANP), to recognize the concepts and emotions for an image, but the information they use is just text data. This research has proved that one image's text is indicative for its sentiment, and needs to be incorporate with vision features together for emotion detection. Wang et al.[30] and Guo[31] et.al tried to train a mapping function between low-level features and high-level emotions and to realize image emotion recognition. However, the problems of such function-based models are that they heavily rely on the training dataset, while in the social media environment, a well-defined training dataset, which conveys all the possibilities in the real world, is hard to construct. Shuoyan Liu et al[32] combined the unsupervised generative model pLSA with domain knowledge together to make the image emotion categorization. However, pLSA has been proved to be inferior to LDA, especially in modeling the distribution on the document level.

Besides, most image emotion researches care about just the emotion in images, but ignore the topics which convey emotions. In many cases the topic-emotion pairs are more informative than emotions alone, like "happy babies" is more meaningful than "happy". As far as we are concerned, very limited works have been done in the computer vision field to identify both topic and emotion at same time.

This paper proposes a framework, JTE-MMHLDA, to identify the topic and emotion contained in an image simultaneously. JTE-MMHLDA extends the mmLDA model, but uses two kind of hidden variables to infer topic distribution and emotion distribution respectively. In addition, JTE-MMHLDA assumes each hidden variable has distributions on both of the text and vision features. Furthermore, JTE-MMHLDA introduces a two-level model structure to realize the topic/emotion sharing among the whole document collections.

III. JOINTLY TOPIC/EMOTION MULTI-MODEL HIERARCHICAL LDA

1. Foundation Models for JTE-MMHLDA

The theory foundation of JTE-MMHLDA is the LDA model. LDA assumes two levels of multinomial distributions, which are controlled by two levels of Dirichlet distributions. LDA is the generative model shown as Fig. 2 (a), and its generative process is as follows:

(1) For each topic$k\epsilon\{1,2,3 \ldots K\}$, sample a distribution □(k) according to a Dirichlet distribution parameterized by β, $\phi_k \sim Dir(\beta)$

(2) For each document$d\epsilon\{1,2,3 \ldots D\}$, sample a distribution θ(d) according to a Dirichlet distribution parameterized by α. $\theta_d \sim Dir(\alpha)$

(2.1) For each word position w in document d, sample a topic according to θ_d, $z_w \sim Multinomial(\theta_d)$

(2.2) For each word position w in document d, sample a word according to z_w, ϕ_{z_w}, $w \sim Multinomial(z_w, \phi_{z_w})$

The problem of LDA is that it can only process homogenous features, but in image process, an image might have multiple types of features, such as text features, vision features, or even cognitive features. In order to extend LDA model to cope with multiple-type-feature problem, mmLDA has been created. mmLDA combines different modalities together, and each modality is also a

LDA model dealing with one kind of features. Figure2(b) shows the graphic model of mmLDA. In order to simplify the presentation, we assume two modalities in mmLDA, and one is for text features, and the other one is for vision features. The generative process of mmLDA is shown as follows:

(1) For each text topic $k\epsilon\{1,2,3 \dots K\}$ sample a distribution according to a Dirichlet distribution parameterized by β.$\phi_k \sim Dir(\beta)$

(2) For each vision topic $t\epsilon\{1,2,3 \dots T\}$ sample a distribution according to a Dirichlet distribution parameterized by γ. $\varphi_t \sim Dir(\gamma)$

(3) For each document d $d\epsilon\{1,2,3 \dots D\}$, sample a distribution according to a Dirichlet distribution parameterized by α. $\theta_d \sim Dir(\alpha)$

(3.1) For each text feature position in document d, sample a text topic z_w according to θ_d. $z_w \sim Multinomial\ (\theta_d)$

(3.2) For each vision feature position in document d, sample a vision topic r_v according to θ_d, $s_v \sim Multinomial(\theta_d)$

(3.3) For each text feature position in document d, sample a text feature w according to z_w, ϕ_k, $w \sim Multinomial(z_w,\phi_k)$

(3.4) For each vision feature position in document d, sample a vision feature v according to s_v,φ_t, $v \sim Multinomial(s_v,\varphi_t)$

The problem of mmLDA is that it can only handle one hidden variable in the whole process. Nevertheless, an image might need to be explained from multiple perspectives in order to get a comprehensive understanding. In LDA extended models, a new latent variable tend to be introduced to cope with the multi-hidden-variable problem. For example, JST has added one "latent sentiment variable" to help in sentiment detection. The graphic model of JST is shown as in Figure 2 (c). JST assumes each document will have a distribution on both topic and sentiment, which are two hidden variables in this model. The generative process of JST is shown as follows:

(1) For each topic $t\epsilon\{1,2,3 \dots T\}$, sample a sentiment distribution according to a Dirichlet distribution parameterized by β, $\varphi_t \sim Dir(\beta)$

(2) For each document d, sample a sentiment distribution according to a Dirichlet distribution parameterized by γ, $\pi_d \sim Dir(\gamma)$

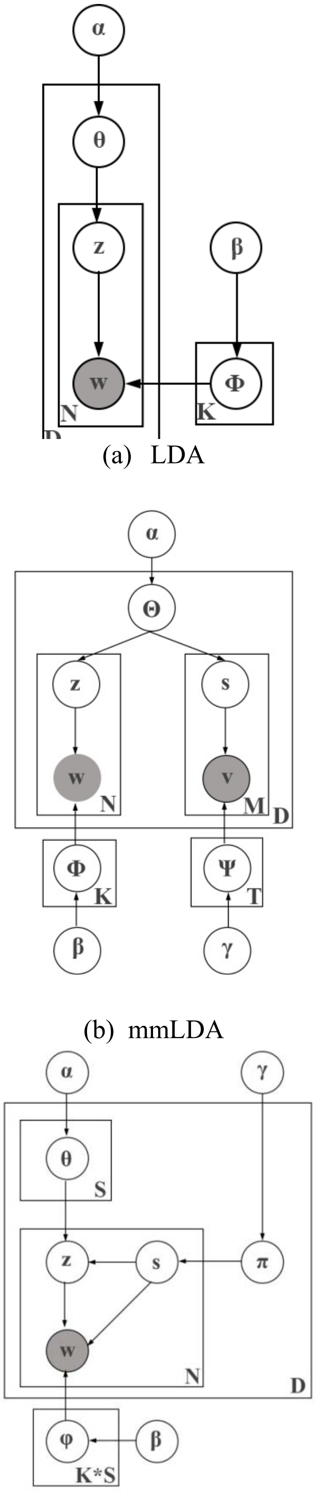
(3) For each sentiment label $s\epsilon\{1,2,3 \dots S\}$ under document d, sample a distribution according to a Dirichlet distribution parameterized by α, $\theta_{d,s} \sim Dir(\alpha)$

(4) For each word w in document d

(4.1) Sample a sentiment label $s \sim Multinomial\ (\pi_d)$

(4.2) Sample a topic $z \sim Multinomial(\theta_{d,s})$

(4.3) Sample a word $w \sim Multinomial\ (\varphi_z, s)$



(a) LDA

(b) mmLDA

(c) JST

Fig. 2. Graphic models of foundation models

Another problem of mmLDA is that the topics are not shared by the document collection, just as shown in Fig.2 (b),each document has its own topic distribution θ_d.

Private topic distribution will make the final result messy and hard to interpret. In order to solve this problem, JTE-MMHLDA utilizes a hierarchical structure. For each document, its own topic distribution θ_d is generated by a global shared topic distribution θ. In this case, we can interpret each image's topic via the shared θ, making the results more comparable and flexible.

2. Principles of JTE-MMHLDA

JTE-MMHLDA extends mmLDA by adding a distribution level and introducing a new hidden variable. Even though a lot of sentiment models assume topics and sentiments are dependent, either topic is dependent on the sentiment orsentiment is dependent on the topic, JTE-MMHLDA assumes them independent. Most sentiment models are developed for a specific entity, such as a specific restaurant. In this case, under each topic, the sentiment distributions change very little among reviews because of the existing real status of the restaurant. If a restaurant does have a very awful service, most people will put forward a negative sentiment about its service in their reviews, so the sentiment distribution under the topic "service" about this restaurant tends have a negative bias. In the contrast, online images' emotions under a specific object might change a lot from images to images. Even for the same object, different images might present very different emotions. Taking the images in Figure 1 as examples, Figure 1(a) and Figure 1(b) share the same topic, but they have very opposite emotions. So, it is more reasonable to deal topic and emotion as two independent variables in image emotion analysis.

The basic idea of JTE-MMHLDA is that each image has two independent distributions, one is about topic, and the other one is about emotion. There are two distribution levels for each of them respectively. The first level distribution is shared by all the images in the collection, and the second level distributions are private to each of the images. But the second level distributions are controlled by the first level ones. Both of the topics and emotions are denoted by the features extracted from the image entity. Two kinds of features will be used, text features and vision features.

3. Model Description

We treat topic/emotion detection task as a two-level problem. Locally, there are four kinds of groups: the text topic, the text emotion, the vision topic and the vision emotion for each document. Globally, the text topic and vision topic should be mapped into the same topic space which is shared by all the documents, and similarly, the text emotion and vision emotion should be mapped to the shared emotion space too. Thus, in the first level, we need assign the features to different groups, and in the second level, we need to map the local groups to the global ones.

Assume we have a corpus of D image entities, and each image entity contains one image and its descriptive text, including title, description and tags. Each image entity can be represented by an N-dimension feature vector, which includes *Nt* dimensions of text features and *Nv* dimensions of vision features. If one feature is a text feature, it will be decided as either a text topic feature or a text emotion feature. Similarly, a vision feature will be decided as either a vision topic feature or a vision emotion feature. Such decision will be made beforehand according to prior knowledge, which will be detailed in the Experiment part.

A text topic feature will be used to infer text topics, a text emotion feature will be used to infer text emotion, a vision topic feature will be used to infer vision topics, and a vision emotion feature will be used to infer the vision emotions. In addition, the local text and vision topics will be mapped to the global topics and the local text and vision emotions will be mapped to the global emotions. Figure 3 shows the graphic model of JTE-MMHLDA

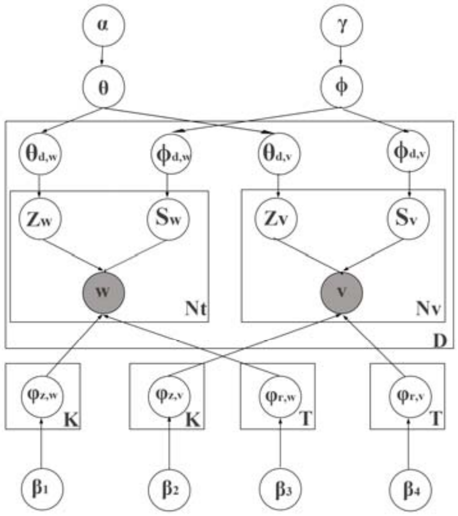


Fig. 3. Graphic model of JTE-MMHLDA

The generative process of JTE-MMHLDA can be described as follows:

(1) For each text topic $tt\epsilon\{1,2,3 \ldots TT\}$, sample a distribution according to a Dirichlet distribution parameterized by β_1, $\varphi_{tt}{\sim}Dir(\beta_1)$

(2) For each text emotion $te\epsilon\{1,2,3\ldots TE\}$, sample a distribution according to a Dirichlet distribution parameterized by β_2,$\varphi_{te}\sim Dir(\beta_2)$

(3) For each vision topic $vt\epsilon\{1,2,3\ldots VT\}$, sample a distribution according to a Dirichlet distribution parameterized by β_3, $\varphi_{vt}\sim Dir(\beta_3)$

(4) For each vision emotion $ve\epsilon\{1,2,3\ldots VE\}$, sample a distribution according to a Dirichlet distribution parameterized by β_4, $\varphi_{ve}\sim Dir(\beta_4)$

(5) Draw a global topic distribution θ according to hyper-parameter α, $\theta\sim Dir(\alpha)$

(6) Draw a global emotion distribution □ according to hyper-parameter γ, $\phi\sim Dir(\gamma)$

(7) For each document d:

(7.1) Draw a text topic distribution $\theta_{d,w}$ according to θ

(7.2) Draw a vision topic distribution $\theta_{d,v}$ according to θ

(7.3) Draw a text emotion distribution $\phi_{d,w}$ according to □

(7.4) Draw a vision emotion distribution $\phi_{d,v}$ according to □

(7.5) For each feature position in document d:

i. If this feature is a text topic feature

- draw a topic z_w according to θ_d, $z_w\sim Multinomial(\theta_{d,w})$
- draw a text topic feature according to z_w, φ_{tt}, $w\sim Multinomial(z_w, \varphi_{tt})$

ii. If this feature is a text emotion feature, draw an emotion r_w according to ϕ_d

- draw an emotion s_w according to ϕ_d, $s_w\sim Multinomial(S_d)$
- draw a text emotion feature according to s_w, φ_{te}, $v\sim Multinomial(s_w, \varphi_{te})$

iii. If this feature is a vision topic feature

- draw a topic z_v according to θ_d, $z_v\sim Multinomial(\theta_d)$
- draw a text topic feature according to z_v, φ_{vt},$w\sim Multinomial(z_v, \varphi_{vt})$

iv. If this feature is a vision emotion feature

- draw a topic s_v according to ϕ_d, $s_v\sim Multinomial(\phi_d)$
- draw a text topic feature according to r_v, φ_{ve}, $w\sim Multinomial(s_v, \varphi_{ve})$

Table 1 explains the annotation in Figure 3.

4. Model Inference

We use Gibbs Sampling as the method to infer the distribution. Following the model described as above, the full joint distribution for the model can be represented as follows:

$$p(w,v,\theta,\phi,\boldsymbol{\varphi}|\alpha,\boldsymbol{\beta},\gamma) = p(\theta|\alpha) * p(\phi|\gamma) * p(\boldsymbol{\varphi}|\boldsymbol{\beta}) * p(\boldsymbol{\theta_d}|\theta) * p(z_w|\theta_{d,w}) * p(z_v|\theta_{d,v}) * p(w^z|z_w,\varphi_{z_w}) * p(v^z|z_v,\varphi_{z_v}) * p(\boldsymbol{\phi_d}|\phi) * p(s_w|\phi_{d,w}) * p(s_v|\phi_{d,v}) * p(w^s|s_v,\varphi_{s_w}) * p(v^s|s_v,\varphi_{s_v}) \quad (1)$$

where w^z denotes to a text topic word, v^z denotes to a vision topic word, w^s denotes to a text emotion word, and v^s denotes to a vision emotion word.

Table 1. The Annotations of graphical model

α	The hyper parameter to generate the global topic distribution
β	The hyper parameter to control the word distributions
γ	The hyper parameter to generate the global emotion distribution
θ; θ_d	The topic distribution; θ is the global topic distribution, $\theta_d$is the local distribution
□; □$_d$	The emotion distribution; □ is the global emotion distribution, □$_d$ is the local emotion distribution
z	A topic
s	An Emotion
w	A text feature
v	A vision feature
N	The length of features within a document
D	The number of document
T	The number of emotions
K	The number of topics

Note that one word can only be one of the four categories, namely the text topic word, the text emotion word, the vision topic word, and the vision emotion word, so the above formula can be simplified. Taking the text topic word as an example, if a word is defined as a text topic word, the above formula can be simplified as follows:

$$p(w^z,\theta_{d,w},\varphi_{tt}|\alpha,\beta_1) = p(\theta|\alpha) * p(\theta_{d,w}|\theta) * p(z_w|\theta_{d,w}) * p(w^z|z_w,\varphi_{z^w}) * p(\varphi_{tt}|\beta_1) \quad (2)$$

Just as mentioned above, the JTE-MMHLDA is a two level model. Still using a text topic assignment as an example, we will infer the local text topic first, and then map this topic to the global one. The local text topic inference process can be calculated as follows:

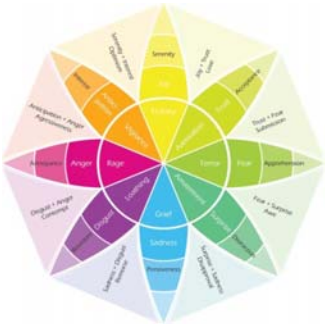
$$p(z_w|\theta_{d,w}) * p(w^z|z_w,\varphi_{z^w}) * p(\varphi_{tt}|\beta_1) \propto \quad (3)$$
$$\frac{N_d^{z_w}+\eta}{\sum(N_d^{z_w}+\eta)} * \frac{N_{z_w}^{w^z}+\beta_1}{\sum(N_{z_w}^{w^z}+\beta_1)}$$

The local topic matching process can be calculated as follows:

$$p(\theta|\alpha) * p(\theta_{d,w}|\theta) \propto \quad (4)$$
$$\frac{N_z^{z_w}+\zeta}{\sum N_z^{z_w}+\zeta} * \frac{N_z^{w^z}+\alpha}{\sum N_z^{w^z}+\alpha}$$

where $N_d^{z_w}$ indicates the number words in document d has been assigned to local text topic z_w, $N_{z_w}^{w^z}$ indicates how many times the word w^z has been assigned to topic z_w, $N_z^{z_w}$ indicates how many times local text topic z_w has been assigned to global topic z, and $N_{z_w}^{w^z}$ indicates how many times the word z_w has been assigned to the global topic z.

By iteration the process as shown above, we can infer the real distribution of text topic words over topics. Other distributions, namely text emotion word→text emotion distribution→emotion distribution, vision topic word →vision topic distribution → topic distribution, vision emotion word →vision emotion distribution→ emotion distribution can be infer by similar process. To be simplification, this paper will not expand detail here.



Keyword List:
Ecstasy→ Joy→ Serenity
Admiration→ Trust→ Acceptance
Terror→ Fear→ Apprehension
Amazement→ Surprise→ Distraction
Grief → Sadness→ Pensiveness
Loath → Disgust→ Boredom
Rage→ Anger→ Annoyance
Vigilance → Anticipation→ Interest

Fig. 4. Plutchnik's wheel of emotion.[*]

IV. EXPERIMENT

1. Datasets and Data Process

The test data we want to use needs to be emotional images containing both image and text information. In order to get a decent dataset, we built a program to help extract images from Flickr. We appliy Plutchnik's Wheel of Emotion[33]as our theory foundation to obtain the images with emotions. According to Plutchnik's theory, human will basically have 8 emotions, and each has 3 valences, just as shown in Figure 4. Thus, we use all the 24 emotion words as our keywords to visit corresponding images in Flickr. The 24 words are shown Figure 4. Totally we have collected 36765 images as our dataset. The dataset is available from the authors upon request.

2. Text Features

To simplify the process, we just use each word from the image's title, description and tags as a text feature. Generally speaking, prior knowledge will help to improve the final accuracy. Thus, we predefine the character of each feature before they are put into the model. We utilized the sentiment dictionary MPQA[34] to help tell whether a word conveys emotion. MPQA is a dictionary recording most of the words with sentiment. We simply assume that all words in MPQA as text emotion words, otherwise as text topic words.

3. Vision Features

We use the toolkit provided by Jianxiong Xiao el at[35] to extract the vision features. The features we use are the Geo-Color, Geo-Map, Geo-Texton, GIST, LBP, LBPHF, Texton and Tiny_Image provided by this toolkit.

As mentioned in Jana Machiajdi el at's work[14], color and texture are two import features in image emotion representation. Thus, we extract these two kinds of features as the vision emotion features. Geo-Color and Geo-Texton are two groups of features to deal with color and texton histograms for each geometric class (ground, vertical porous, and sky). Texton is a traditional and powerful local image descriptor is to convolve the image with Gabor-like filter bank[36], and this group of features is also used to indicate the emotion of the image. LBP and LBPHF are also deal with the texture things, and are also defined as vision emotion features.

Geo-Map is to compute the geometric class probabilities for image regions. GIST is used to compute wavelet image decomposition. Tiny_Image is used to reduce the image dimensions for object recognition and scene classification. All these three groups of features are used as vision topic features.

4. Experiment Results

We test our algorithm on our Flickr dataset. For all the hyper-parameters, we employ the standard and out-box settings without any tune to our data. We set all the βs as 0.01, and other hyper parameters as 0.02. The results shown in Table 2 are generated under the situation that

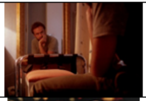
[*] image is from https://www.pinterest.com/pin/259801472228430213/

topic number is 20, and emotion number is 8. Table 2 gives a quick view of our experiment results.

In Table 2, Column "Original Tags" presents the tags attached with this image from Flickr. Because some images have too many tags, considering the limitation of space, we just present some of them, and use "…" to indicate there are more tags. Column "Topical Words" presents the representative words generated by our model to describe this image's topic. Column "Emotion Words" presents the representative words generated by our model to describe this image's emotion.

Comparing column "Original Tags" and "Topic Words", we find these two groups of words are highly correlated. For image 1, our model has detected the topic words "Church", "Cathedral", "Liverpool", "Christian", which exactly appear in its original tags. For image 3, our model has detected the word "Face" and "Head", which do not appear in its original text, but definitely can be used describe this image.

Table 2. Examples of the experiment results.

ID	Example Image	Original Tags	Topical Words	Emotion Words
1		Liverpool Cathedral Church, cathedral, church, house of worship, house of god; Christian…	Church, Cathedral, Liverpool, Selvage, House, Christian, Place	Interesting, Like, Beautiful, Hope, Angel, Fantasy, Love Support
2		Flowers, Daisy, Birthday, Surprise, Gardens, Nature…	Color, Birthday, Flower, Blooms, Display, Art, Photo, Thanks	Love, Nature, Joy,Surprise, Exhilaration, Happiness, Delightfulness, Pleasure
3		Enfant, Adolescent, Surprise, Child, Girl, Eyes, Marie Bousquet, Photo Marie Bousquet…	Face, Portrait, Soul Woman, Head	Surprise, Motionless, Funny, Restlessness
4		Hotel, Room, Portrait, Glass, Mirror, Bored, Boredom, Paris …	People, Man, Portrait, Face, Monochrome, Bored, Life, Hotel, Hanging	Boredom, Dark, Loathing, Bad, Close, Isolation, Lonely
5		Statue , Art Sculpture Grave, Tombstone, Tomb…	Art, Sculpture, Statue, Person, Monument, , Tombstone, Display, Angel	Sadness, Sad, Lost, Tragic, Mourning, Burial
6		Woman, Sad, Color, Mask, Look, Face, Red, Orange, White, Sadness, Colored, Sadness…	Face, Portrait, Soul, Woman, Head	Sadness, Twisted, Sad, Lost, Calm, Worst, Cold

Comparing column "Original Tags" and "Emotion Words", we find that our model can help to find more rich word collections to describe an image's emotion. Most tags given to an image are noun words, conveying very limited emotions. Taking Image 4 as an example, it describes a man sitting lonely and bored in a room. Even though the original text give words like "Boredom" and "Bored" to describe its emotion, our model has helped to find words like "Loathing", "Isolation" and "Lonely", which also are very appropriate words to describe this image's emotion.

Comparing Image 3 and Image 6, we find that these two images belong to the same topic, and they all portrait a female face, but our model has detected that they belong to very different emotions. The emotion in Image 3 is about "surprise" and "motionless", while emotion in Image 6 is about "sadness" and "twisted". This example also indicates that emotion detection, besides topic detection, is very important for an image retrieval system.

5. Evaluation

We use text information as the hint for accuracy. Taking topic assignment as an example, for each document, we collect the top 3 topics this document has been assigned to, and gather all the top 50 words in each of them as topic words. We compare this document's text word set with the topic word set. If their union size is larger than the threshold we defined, we deem this document has been rightly classified. We set the threshold as 5 words. But, if the document's length is smaller than 10, we assume if half number of words are found in the topic word set, the document has been rightly assigned.

First, we measure how the accuracy changes with different settings of the topic/emotion number in JTE-MMHLDA. At the beginning, we fix the number of topics, and change the number of emotions (shown in Figure 5(a)), and find that the emotion accuracy falls as the emotion number increases. When the emotion number set as 3, the accuracy can hit 98%, but when the emotion number increase to 6 and above, the accuracy can only keep around 70%. Then, we fix the number of emotion, and change the number of topics (shown in Figure5 (b)), and find even though there is fluctuations, the accuracy of topic stays relatively stable, keeping between 64% and 71%.

The possible reasons contributing to such results might root from the unbalance in the dataset. Common sense tells us that the more classes exist, the more possible a word could be misclassified, and the less accurate the final result might be. In our dataset, most words are topic words, and much fewer words are emotion words. Thus, the emotion accuracy is more sensitive because of the small sized emotion word set, even if only a small subset of words have been misclassified, the final accuracy will be affected a lot. Comparing to the emotion word set, the topic word set is much larger, and thus although a reasonable words have misclassified, the final result will not be impacted so much.

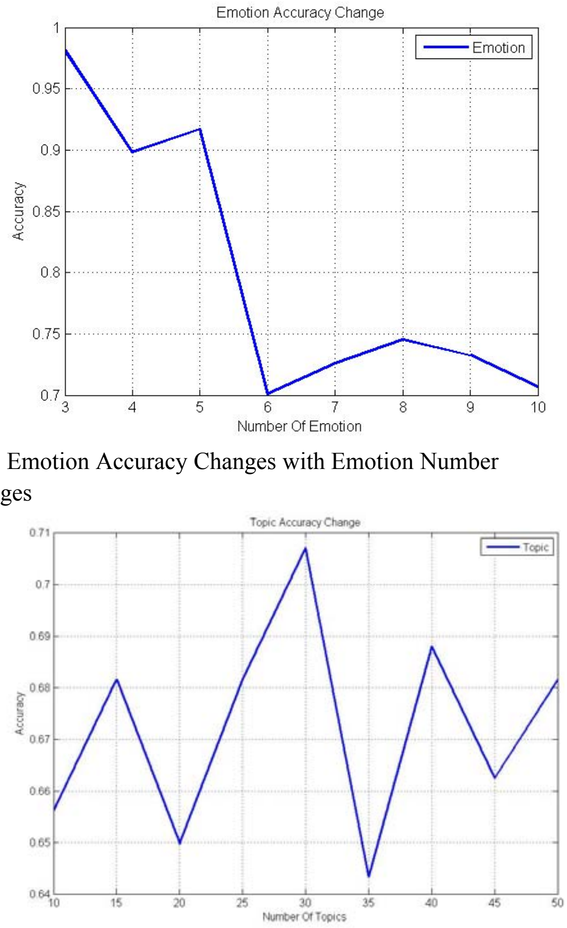
Then, we decompose our model into three parts, model with only text features(t-JTE), model with only vision features (v-JTE), and model with both text and vision features(j-JTE). We want to explore whether the combination of text features and vision features will improve the model's performance. Figure 6 shows the experiment results.

Figure 6(a) shows the emotion accuracy changes with different emotion numbers but fixed topic number. Both j-JTE and t-JTE outperform the v-JTE, and this indicates that text is an important factor to reveal an image's emotion, and vision features alone are hardly to be competent. Comparing j-JTE and t-JTE, we find the results of these two are very close, but j-JTE is a little better. So, vision features alone might not be good for images' emotion detection, but when combined with text features, it can help to improve the final result.

Figure 6(b) shows the topic accuracy changes with different topic numbers but fixed emotion number. Again, both j-JTE and t-JTE beat v-JTE in topic detection. Comparing j-JTE and t-JTE, j-JTE still edges out t-JTE. Once more, this result shows that vision features alone are not good for topic detection comparing to the text features, but they help to improve the performance when combined with text features.

Figure6 shows a big difference between the accuracies of emotion detection and topic detection. The emotion detection accuracy can be higher than 70%, but topic detection accuracy stays below 70%. The main reason contribute to this phenomenon is still the data unbalance problem. For many images, they have limited number, if not none, of emotion words, so it is very easy for one of the top3 emotions to contain the certain words, and thus the average accuracy could not be very low. While, the topic collection has very many words, and a topic might rank many of one word's synonyms ahead, but not exactly that word. In this case, even the document has been rightly assigned, but the evaluation algorithm could not find the exact word, and will judge it as a mismatch. So, the topic detection accuracy is relatively lower.

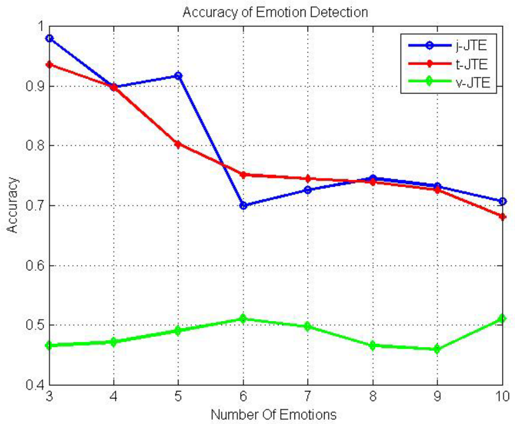
Finally, we compare j-JTE, t-JTE and v-JTE with different numbers of topics and the emotions. Figure 7(a) shows the result on emotion analysis, and Figure 7(b) shows the result on topic analysis. In Figure 7, the mark "10/3" indicates this result is generated under the settings of topic number is 10, and emotion number is 3.



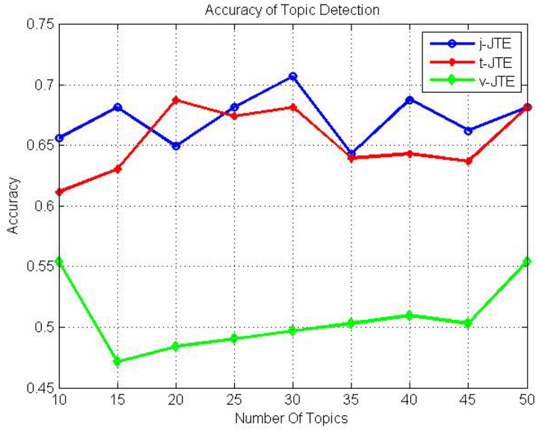
(a) Emotion Accuracy Changes with Emotion Number Changes

(b) Topic Accuracy Changes with Topic Number Changes

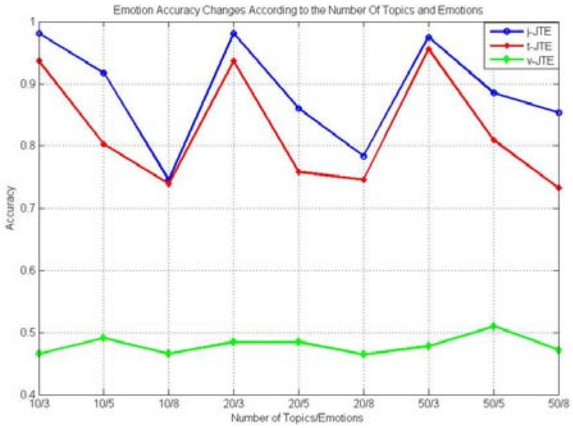
Fig. 5. Accuracy Changes with Number Changes



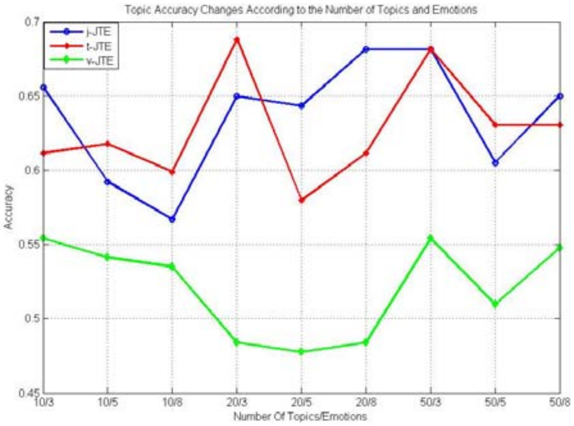
(a) Emotion Accuracy Measurement When Topic Fixed



(b) Topic Accuracy Measurement When Emotion Fixed
Fig. 6. Accuracy Comparison among j-JTE, t-JTE, and v-JTE



(a) Emotion Measurement with Different Topic/Emotion Number



(b) Topic Measurement with Different Topic/Emotion Number
Fig. 7. Accuracy Changes with Different Topic/Emotions

Figure7 (a) shows j-JTE and t-JTE are always outperform v-JTE. The accuracy of j-JTE and v-JTE can be higher than 75%, while v-JTE can only stay below 50%. But, v-JTE is not as sensitive as the other two models are. J-JTE and t-JTE's accuracy drops dramatically when the number of emotions increases.

Figure 7(b) shows that j-JTE and t-JTE are better than v-JTE in topic detection too. The j-JTE and v-JTE have some crosses in the topic detection part, but generally j-JTE is a little better than t-JTE (j-JTE has 4 points are above 65%, while t-JTE only has 2). Topic accuracy is not very sensitive to the topic number change, especially when the topic number is larger enough, namely to be 20 or 50, j-JTE keeps a relatively stable accuracy in topic detection.

Overall, j-JTE has a best performance comparing to t-JTE and v-JTE. Specifically, j-JTE outperforms t-JTE by 4.4% and v-JTE by 18.1% in topic detection, and outperforms t-JTE by 7.1% and v-JTE by 39.7% in emotion detection. This proves that jointly using text features and vision features can help to improve the topic/emotion detection results. Besides, j-JTE has relatively high emotion detection accuracy (above 90% when emotion number is small, and above 70% when the number is large), and reasonable topic detection accuracy (above 60% in most cases). Thus, the JTE-MMHLDA is promising in solving the corresponding problem.

V. CONCLUSION AND FUTURE WORK

This paper has proposed a JTE-MMHLDA model to detect images' topic and emotion distributions in a unified framework. JTE-MMHLDA has extended the classical mmLDA, but expanded it to two levels, which can facilitate the topic/emotion sharing among images, and introduced an extra hidden variable to infer the emotion distribution for each images. The experiment and evaluation results show that JTE-MMHLDA is a promising model for simultaneously topic/emotion detection from images.

However, something still needs to be strengthened in the future. First, we use Gibbs Sampling method to make the inference. The advantages of Gibbs Sampling are including easy to use and more accurate, but it is slower comparing to other inference methods. We will conduct some accelerate works in the future. Second, currently, the number of topics and emotions need to be manually determined, and we find sometimes it is hard to make precise estimations about these numbers. We are considering incorporating some non-parameter methods to help deal with the number determinations.

REFERENCES

[1] A. Mojsilović, J. Gomes, and B. Rogowitz, "Semantic-friendly indexing and quering of images based on the extraction of the objective semantic cues," *International Journal of Computer Vision,* vol. 56, pp. 79-107, 2004.

[2] C. Cusano, G. Ciocca, and R. Schettini, "Image annotation using SVM," in *Electronic Imaging 2004*, pp. 330-338, 2003.

[3] G. Carneiro and N. Vasconcelos, "Formulating semantic image annotation as a supervised learning problem," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, pp. 163-168. 2005

[4] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 25, pp. 1075-1088, 2003.

[5] D. Putthividhy, H. T. Attias, and S. S. Nagarajan, "Topic regression multi-modal latent dirichlet allocation for image annotation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 3408-3415, 2010.

[6] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, pp. 524-531, 2005

[7] C. Wang, D. Blei, and F.-F. Li, "Simultaneous image classification and annotation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1903-1910, 2009.

[8] D. Putthividhya, H. T. Attias, and S. S. Nagarajan, "Supervised topic model for automatic image annotation," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 1894-1897, 2010

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research,* vol. 3, pp. 993-1022, May 15 2003.

[10] V. Vonikakis and S. Winkler, "Emotion-based sequence of family photos," in *Proceedings of the 20th ACM international conference on Multimedia*, pp. 1371-1372, 2012.

[11] V. Yanulevskaya, J. Van Gemert, K. Roth, A.-K. Herbold, N. Sebe, and J.-M. Geusebroek, "Emotional valence categorization using holistic image features," in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pp. 101-104, 2008.

[12] J. Jia, S. Wu, X. Wang, P. Hu, L. Cai, and J. Tang, "Can we understand van gogh's mood?: learning to infer affects from images in social networks," in *Proceedings of the 20th ACM international conference on Multimedia*, pp. 857-860, 2012.

[13] B. Li, S. Feng, W. Xiong, and W. Hu, "Scaring or pleasing: exploit emotional impact of an image," in *Proceedings of the 20th ACM international conference on Multimedia*, pp. 1365-1366, 2012.

[14] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the international conference on Multimedia*, pp. 83-92, 2010.

[15] V. Yanulevskaya, J. Uijlings, E. Bruni, A. Sartori, E. Zamboni, F. Bacci, *et al.*, "In the eye of the beholder: employing statistical analysis and eye tracking for analyzing abstract paintings," in *Proceedings of the 20th ACM international conference on Multimedia*, pp. 349-358, 2012.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research,* vol. 3, pp. 993-1022, 2003.

[17] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, pp. 2169-2178, 2006.

[18] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, pp. 1800-1807, 2005.

[19] Z.-H. Zhou and M.-L. Zhang, "Multi-instance multi-label learning with application to scene classification," in *Advances in Neural Information Processing Systems*, pp. 1609-1616, 2006.

[20] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 127-134, 2003.

[21] C.-T. Nguyen, D.-C. Zhan, and Z.-H. Zhou, "Multi-modal image annotation with multi-instance multi-label LDA," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pp. 1558-1564, 2013.

[22] J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *Advances in neural information processing systems*, pp. 121-128, 2008.

[23] Y. Shang, Y. An, X. T. Hu, M. Zhang, and X. Lin, "Enhancing Entity Annotation using Web Service and Ontology Hierarchy in Biomedical Domains," in *2013 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 465-468, 2013.

[24] M. Liu, Y. An, X. Hu, D. Langer, C. Newschaffer, and L. Shea, "An Evaluation of Identification of Suspected Autism Spectrum Disorder (ASD) Cases in Early Intervention (EI) Records."

[25] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 375-384 , 2009.

[26] W. X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 56-65, 2010.

[27] Y. Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 815-824, 2011.

[28] X. Xu, S. Tan, Y. Liu, X. Cheng, and Z. Lin, "Towards jointly extracting aspects and aspect-specific sentiment knowledge," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 1895-1899, 2012.

[29] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 223-232, 2013.

[30] W. Wei-ning, Y. Ying-lin, and J. Sheng-ming, "Image retrieval by emotional semantics: A study of emotional space and feature extraction," in *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*, pp. 3534-3539, 2006.

[31] Y. Guo and H. Gao, "Emotion recognition system in images based on fuzzy neural network and HMM," in *Cognitive Informatics, 2006. ICCI 2006. 5th IEEE International Conference on*, pp. 73-78, 2006.

[32] S. Liu, D. Xu, and S. Feng, "Emotion categorization using affective-plsa model," *Optical Engineering*, vol. 49, No.12, pp. 127201, 2010.

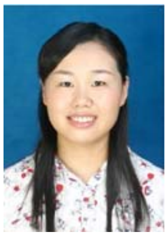
[33] R. Plutchik, *Emotion: A psychoevolutionary synthesis*: Harper & Row New York, 1980.

[34] J. Wiebe, E. Breck, C. Buckley, C. Cardie, P. Davis, B. Fraser, *et al.*, "MPQA: Multi-Perspective Question Answering Opinion Corpus Version 1.2", 2006.

[35] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva, "SUN Database: Exploring a Large Collection of Scene Categories," *International Journal of Computer Vision,* pp. 1-20, 2014.

[36] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Computer Vision–ECCV 2006*, pp.1-15, 2006.

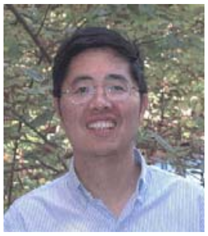
Authors



Wanying Ding is a PhD student and Research Assistant in College of Computing and Informatics in Drexel University. She received her Bachelor Degree from Wuhan University, and Master Degree from Peking University. She is a member of the NSF Center of Visual and Decision Informatics (CVDI). She takes the responsibility to develop algorithms, analyze and visualize data from social media in this group. Her research interests include data mining and data visualization for social media.

Junhuan Zhu is a PhD student in computer science at the University of Rochester. He received his BEng degree in electronics and information engineering from Huazhong University of Science and Technology in 2010, and his MS degree in electrical and computer engineering from the University of Rochester in 2013. His current research interests include digital video analysis and social media data mining.

Dr. Lifan Guo is a research scientist at TCL Research America. He often takes the leadership role in big data architect in various challenges of text mining, information retrieval and social media analysis, where he develops best practices for management and analysis of big data. Prior to that, he graduated from Drexel University, majored in information science.



Dr. Xiaohua Hu is a full professor and the founding director of the data mining and bioinformatics lab at the College of Computing and Informatics, Drexel University. He is also serving as the founding Co-Director of the NSF Center (I/U CRC) on Visual and Decision Informatics, IEEE Computer Society Bioinformatics and Biomedicine Steering Committee Chair, and IEEE Computer Society Big Data Steering Committee Chair. Besides, he also serves as the founding editor-in-chief or associate editor of many international journals. Till now, he has published more than 240 peer-reviewed research papers and co-edited 20 books/proceedings.

Dr. Jiebo Luo joined the University of Rochester in fall 2011 after over 15 years at Kodak Research Laboratories as a senior principal scientist. He has been involved in numerous technical confer-ences, including serving as the program co-chair of ACM Multimedia 2010 and IEEE CVPR 2012. He is the editor-in-chief of the Journal of Multimedia and has served on the editorial boards of numerous international journals. He has authored over 200 technical papers and 80 U.S. patents. He is a fellow of SPIE, IEEE, and IAPR.

Dr. Haohong Wang is the General Manager of TCL Research America at San Jose, California. Prior to joining TCL, he held technical and management positions at AT&T, Catapult, Qualcomm, Marvell and Cisco. He is the inventor of 50+ patents and pending applications, and co-author of 5 books and 50+ articles in journals and conferences. He is the Editor-in-Chief of the Journal of Communications, and has been a member of the Steering Committee of the IEEE Transactions on Multimedia. He co-chairs the IEEE Technical Committee on Human Perception and Multimedia Computing, and has chaired the IEEE Multimedia Communications Technical Committee. He chairs the Steering Committee of ICNC conference, and has served as the General Chair of IEEE ICME 2011 and IEEE VCIP 2014, and as the TPC Chair of IEEE GLOBECOM 2010. He is the recipient of the IEEE MMTC Distinguished Service Award (2013), Manager of the Year Award (2012), Distinguished Inventor Award (2013), and Technology Innovation Award (2014) by TCL Corporation. He received his Ph.D. from Northwestern University, Evanston, USA.