Predicting Personalized Image Emotion Perceptions in Social Networks

Sicheng Zhao, Hongxun Yao, Yue Gao, Senior Member, IEEE, Guiguang Ding and Tat-Seng Chua

Abstract—Images can convey rich semantics and induce various emotions to viewers. Most existing works on affective image analysis focused on predicting the dominant emotions for the majority of viewers. However, such dominant emotion is often insufficient in real-world applications, as the emotions that are induced by an image are highly subjective and different with respect to different viewers. In this paper, we propose to predict the personalized emotion perceptions of images for each individual viewer. Different types of factors that may affect personalized image emotion perceptions, including visual content, social context, temporal evolution, and location influence, are jointly investigated. Rolling multi-task hypergraph learning (RMTHG) is presented to consistently combine these factors and a learning algorithm is designed for automatic optimization. For evaluation, we set up a large scale image emotion dataset from Flickr, named Image-Emotion-Social-Net, on both dimensional and categorical emotion representations with over 1 million images and about 8,000 users. Experiments conducted on this dataset demonstrate that the proposed method can achieve significant performance gains on personalized emotion classification, as compared to several state-of-the-art approaches.

Index Terms—Personalized image emotion, social context, temporal evolution, location influence, hypergraph learning.

1 INTRODUCTION

With the rapid development of digital photography technology and wide-spread popularity of social networks, people have become used to sharing their lives and expressing their opinions using images and videos together with text. The explosively growing volume of online social data have greatly motivated and promoted the research on large-scale multimedia analysis. As what people feel may directly determine their decision making, the understanding of these data at the emotional level is of great importance, which can benefit social communication and enable wide applications [1], [2], ranging from marketing [3] to political voting forecasts [4].

Despite the promising progress of textual sentiment analysis [5], emotion analysis of social images remains an open problem [6]. The main challenges that limit the development of image emotion analysis lie in the so-called affective gap, as well as the subjective evaluation [2], [7]. Trying to find features that can express emotions better to bridge the affective gap, previous works mainly focused on predicting the dominant emotions for the majority of viewers, without considering the subjective evaluation. However, predicting personalized emotion is usually more practical, as the emotions that are induced in viewers by an image are highly subjective and different, due to the influence of social, educational and cultural backgrounds [8], [9], [10], [11], [12], as shown in Figure 1. Under such a circumstance, traditional dominant emotion based methods may not work well for personalized emotion prediction [13]. So far, little progress has been made on predicting personalized emotion perceptions (termed PEP), mainly due to two key challenges:

The Lack of Benchmarks. To the best of our knowledge, there is no public dataset on PEP of images, though a few works have been done on emotion or sentiment analysis of social multimedia data [1], [13], [14], [15], [16], [17], [18]. To avoid the tedious manual labeling of large scale social images, existing works on emotion analysis mainly used two methods to obtain the emotion labels based on different emotion representation models. First, from the dimensional emotion perspective, traditional lexicon-based methods are used to find out the polarity values of the comments with a predefined word dictionary. Yang et al. [13], [17] adopted this method for Chinese microblog analysis.

---

**Title:** Light in Darkness

**Tags:** London, Stormy, Dramatic weather, ...

**Description:** ...there was a break in the clouds such that a strip of the far bank and bridge were lit up while the sky behind remained somewhat foreboding. I thought it made for a pretty intense scene....

**Expected emotion:** Emotion category: fear

**Emotion:** fear, Sentiment: negative

V: 2.625 A: 5.805 D: 3.625

---

**Comments to this image from different viewers**

<table>
<thead>
<tr>
<th>Comments</th>
<th>Expected emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wow, that is fantastic... it looks so incredible; like a painting... slightly unreal. That sky is amazing.</td>
<td>Sentiment: positive</td>
</tr>
<tr>
<td>Yup a fave for me as well. Exciting drama at its best.</td>
<td>Sentiment: positive</td>
</tr>
<tr>
<td>Hey, it really frightened me! My little daughter just looked scared.</td>
<td>Sentiment: negative</td>
</tr>
</tbody>
</table>

**Personalized emotion labels**

- Wow, that is fantastic... it looks so incredible; like a painting... slightly unreal. That sky is amazing.: Emotion: awe, Sentiment: positive
  - V: 7.121 A: 4.479 D: 6.635
- Yup a fave for me as well. Exciting drama at its best.: Emotion: excitement, Sentiment: positive
- Hey, it really frightened me! My little daughter just looked scared.: Emotion: fear, Sentiment: negative
  - V: 5.305 A: 3.625

---

**Fig. 1. Illustration of personalized image emotion perceptions in social networks.** (a) The uploaded image to Flickr. (b) The title, tags and description given by the uploader to the image. (c) The expected emotion labels that we assign to the uploader using the keywords in (b) in red with both categorical and dimensional representations. (d) The comments to this image from different viewers and the personalized emotion labels that we obtain using the keywords in red. We can find that the emotions perceived by the viewers are truly subjective.
with emotions represented by one dimension. However, they just classified the emotions into two discrete categories by taking threshold of the continuous polarities. No large scale dataset on social image emotions using dimensional representation based on English dictionary has been released. Second, from the categorical emotion perspective, keywords or synonyms based retrieval of specified emotions are used. The works in [14], [15], [16], [18] adopted this strategy. The emotion categories used for classification are ad hoc and arbitrary with categories numbers ranging from six to tens [10], [19]. Besides, the numbers of positive and negative emotions are unbalanced [15], [18].

**Multi-factor influence.** Besides the visual content of images, there are many other factors that may influence the personalized perception of image emotions. Personal interest may directly influence the emotion perceptions [20]. Viewers’ emotions are often temporally influenced by their recent past emotions [21]. In social networks, the emotions are widely influenced by the social connections. Whittfield [22] showed that how happy you are is influenced by your social links to people even you’ve never heard of and never met. How a viewer’s emotion is influenced by their friends on social networks is quantitatively studied in [15], [18]. The locations of social images, if known, can also be used for visual data analysis [23]. How to consistently and effectively combine these factors for robust personalized emotion analysis is a challenging problem.

In this paper, we make the first attempt in predicting PEP of social images to tackle the two challenges above. For the first challenge, we set up a large scale image dataset of personalized emotions, named Image-Emotion-Social-Net, which are represented by both dimensional and categorical emotion models. We use 8 categories as the categorical emotion states which are defined in a rigorous psychological study [24], including anger, disgust, fear, sadness as negative emotions, and amusement, awe, contentment, excitement as positive emotions. Tens of keywords for each emotion category are used to search the titles, tags and descriptions of images or the comments to the images to obtain the personalized emotion labels. Then we computed the average value of valence, arousal and dominance (VAD) for dimensional emotion space using recently published VAD norms of 13,915 English lemmas [25]. Besides, we give the sentiment category (positive or negative). For the second challenge, we take different types of factors into account, including visual content, social context, temporal evolution and location influence. We propose a rolling multi-task hypergraph learning (RMTHG) to formalize the personalized emotion perception prediction problem by modelling these various factors. Experiments are conducted on our collected dataset and the results demonstrate that by incorporating the various factors, the proposed model can significantly improve the prediction performance compared with the state-of-the-art approaches. We also design one novel application based on the personalized emotion prediction results. Please note that in this paper we do not clearly distinguish expressed, perceived and induced emotions as in music [26]. We use “perceive (perception)” and “induce (induction)” from the perspective of emotion subjects, such as “User A perceives fear from image B” and “Image B induces fear in user A”.

The contributions of this paper can be summarized in three aspects as follows:

1) We propose to distinguish personalized emotions from traditional dominant emotions. What’s more, we also consider different types of factors, including visual content, social context, temporal evolution and location influence.

2) We present a compound vertex hypergraph to model all the different factors in a consistent and expandable way. To enable efficient inference in this framework, we devise a rolling multi-task hypergraph learning algorithm, which can simultaneously predict individual emotions of different users.

3) We set up a large scale personalized emotion dataset of social images from Flickr with over 1 million images and 1.4 million labels for about 8,000 users. The dataset, containing images, metadata, social connections and personalized emotions, will be released along with this work.

One preliminary conference version on personalized image emotion perception prediction has been accepted by ACM Multimedia [27]. Our new improvement compared with the conference version lies in the following aspects. First, we add a data analysis section to inspire the motivation of modelling personalized emotion perceptions. Second, we extend the hyperedge construction in more detail. Finally, we conduct more comparative experiments and enrich the analysis of the results.

The rest of this paper is organized as follows. Section 2 introduces related work of image emotion and sentiment analysis, social media analysis and (hyper)graph based learning. Section 3 presents data analysis based justification to motivate the proposed idea. Section 4 introduces the constructed Image-Emotion-Social-Net dataset. Section 5 gives an overview of the proposed method. Hypergraph construction and rolling multi-task hypergraph learning are described in Section 6 and Section 7, respectively. Experimental evaluation and analysis are presented in Section 8, followed by conclusion and future work in Section 9.

## 2 Related Work

**Image emotion and sentiment analysis.** Some research efforts have been dedicated to improving the accuracy of image emotion prediction. Related works can be divided into different types, according to the adopted emotion models, the required tasks, the extracted features and the models used.

There are two kinds of emotion representation models: categorical emotion states (CES) and dimensional emotion space (DES). CES models model emotions as one of a few basic categories [14], [15], [18], [19], [30], [31], [32], [33], while DES methods employ 3-D or 2-D space to represent emotions, such as valence-arousal-dominance [34], natural-temporal-energetic [35] and valence-arousal [2], [7], [32]. Accordingly, related works on image emotion analysis can be classified into three different tasks: affective image classification [2], [13], [14], [15], [16], [17], [18], [19], [31], [32], [33], regression [2], [32] and retrieval [6], [30]. We model image emotions using both representation models.

From a feature’s view point, different levels of visual features are extracted for image emotion analysis. Low level holistic image features including Wiccest features and Gabor features are extracted to classify image emotions in [31]. Lu et al. [32] investigated the computability of emotion through shape features. Machajdik et al. [19] extracted features inspired from psychology and art theory, including color, texture as low level features, composition as mid level features while face and skin as high level features. Zhao et al. [2] proposed to extract mid level features.

1. Specifically, image emotion is often called image sentiment for binary classification (positive or negative) [1], [16], [28], [29].
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2628787, IEEE Transactions on Affective Computing
The keyword examples of each emotion category. '#' indicates the total keyword numbers.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>#</th>
<th>Keyword examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>amusement</td>
<td>24</td>
<td>amused, amusement, cheer, delight, funny, pleasing</td>
</tr>
<tr>
<td>anger</td>
<td>79</td>
<td>angry, annoyed, enraged, hateful, offended, provoked</td>
</tr>
<tr>
<td>awe</td>
<td>36</td>
<td>amazing, astonishment, awesome, impressive, wonderful</td>
</tr>
<tr>
<td>contentment</td>
<td>28</td>
<td>comfortable, fulfilled, gladness, happy, pleasure, satisfied</td>
</tr>
<tr>
<td>disgust</td>
<td>35</td>
<td>dejection, disgusted, nauseous, queasy, revolt, weary</td>
</tr>
<tr>
<td>excitement</td>
<td>49</td>
<td>adventure, enthusiastic, inspired, stimulation, thrilled</td>
</tr>
<tr>
<td>fear</td>
<td>71</td>
<td>afraid, frightened, nightmare, horror, scared, timorous</td>
</tr>
<tr>
<td>sadness</td>
<td>72</td>
<td>bereaved, heartbroken, pessimistic, sadness, unhappy</td>
</tr>
</tbody>
</table>

(a) Expected emotion

(b) Personalized emotion

**Fig. 4. Dataset validation results. >= n means at least n Yes’s.**

<table>
<thead>
<tr>
<th>emotion</th>
<th>#</th>
<th>keyword numbers</th>
<th>270,748</th>
<th>328,303</th>
<th>181,431</th>
<th>115,065</th>
<th>1,016,186</th>
</tr>
</thead>
<tbody>
<tr>
<td>anger</td>
<td>29,844</td>
<td>disgust</td>
<td>20,962</td>
<td>55,802</td>
<td>57,476</td>
<td>362,400</td>
<td></td>
</tr>
<tr>
<td>disgust</td>
<td></td>
<td>fear</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fear</td>
<td></td>
<td>sadness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sadness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 2**

Image numbers of categorical emotions.

4. [https://en.wikipedia.org/wiki/Fleiss%27_kappa](https://en.wikipedia.org/wiki/Fleiss%27_kappa)


4.3 Statistics of Dataset

The distribution of images per emotion category is shown in Table 2, where the first four columns represent the number of images in each of the 8 emotions; while the last column is the number of images with binary sentiments. We can find that the number of negative emotions is relatively small. The distribution of valence, arousal (without showing dominance here) is illustrated in Figure 7(a), which looks like a petal or heart, similar to the emotion space in [63]. The user distribution based on the involved

Dataset Construction and Quality Validation

We downloaded 21,066,920 images from Flickr with 2,060,357 users belonging to 264,683 groups. Each image is associated with the metadata, such as the title, tags, taken time and location if available. Each user is associated with the personal information, the contact list and the group list they joined in. As how to measure emotions is still far from consensus in research community [62], we defined emotions using both categorical and dimensional representations. For CES, we used the 8 categories rigorously defined in psychology [24], including 4 negative and 4 positive emotions. To get the ground truth labels, we adopted keywords based searching strategy as in [14], [16], [18]. Tens of keywords for each emotion category are obtained from a public synonym searching site and are manually verified, with examples shown in Table 1. Expected emotions of the image uploaders are firstly considered. The keywords are searched from the title, tags and descriptions given by the uploaders. The emotion category with the most frequent keywords is set as the ground truth of expected emotions from the uploaders.

As we are focusing on PEP, we then searched from all the comments of the images tackled above to get the personalized emotion labels of each viewer. As in [14], [18], we removed the images if the searched title, tags or descriptions contain negation adjacent and prior to the target keywords, such as "I am not fear". Similarly, we also removed the comments with negation adjacent and prior to the target keywords. This is because the antonym of an emotion is not so clear. Note that the labels of an image for a specific user are allowed to have different emotion categories (such as fear, disgust) but must have only one sentiment (positive or negative). Then we computed the average value of valence, arousal and dominance of the segmented text (metadata or comments) as ground truth for dimensional emotion representation based on recently published VAD norms of 13,915 English lemmas [25]. Besides, we also gave the sentiment categories (positive or negative). We combined the expected emotions and actual emotions of all involved images for each user. This process resulted in a dataset containing 1,012,901 images uploaded by 11,347 users; and 1,060,636 comments on these images commented by 106,688 users. We chose 7,723 active users with more than 50 involved images. Finally we obtained 1,434,080 emotion labels of three types, including 8 emotion categories, 2 sentiment categories and continuous VAD values. Note that all the involved images of one user are labelled with sentiment categories and VAD values, while a tiny proportion of them are not assigned with the emotion categories if no keyword is found.

If one user is the uploader of an image, then the emotion of the metadata text (title, tags and descriptions) is the personalized emotion of this user, which is also the expected emotion that is expected to induce in other viewers by this user. If one user is a viewer of an image, then the emotion of the comment is the personalized emotion of this user.

Dataset Validation

To validate the quality of the dataset, we did a crowdsourcing experiment on discrete emotions. For each emotion category, we randomly selected 200 images with associated titles, tags and descriptions for expected emotions, and 200 comments with corresponding images for personalized emotions. 5 graduate students (3 males, 2 females) were invited to judge whether the text was used to express the assigned emotions of related images. To facilitate this judgement, they were asked simple question like “Do you think that the text is used to express excitement for this image?”, and they just needed to choose YES or NO. Each image was judged by all the 5 annotators. The result is shown in Figure 4. We can find that for both expected and personalized emotions, on average more than 88% of emotion labels receive at least 3 Yeses, which verifies the quality of the constructed dataset. In such cases, the expected emotion labels are 3.5% more accurately assigned than personalized emotions. To assess the inter-rater agreement, we also calculate the Fleiss’ kappa^4 of the 5 annotators. The average Fleiss’ kappa (the standard deviation) for the 8 emotion categories of expected emotions and personalized emotions are 0.2297 (0.0748) and 0.3224 (0.1411), respectively.

4.3 Statistics of Dataset

The distribution of images per emotion category is shown in Table 2, where the first four columns represent the number of images in each of the 8 emotions; while the last column is the number of images with binary sentiments. We can find that the number of negative emotions is relatively small. The distribution of valence, arousal (without showing dominance here) is illustrated in Figure 7(a), which looks like a petal or heart, similar to the emotion space in [63]. The user distribution based on the involved
Fig. 5. Image examples of personalized emotion categories in the Image-Emotion-Social-Net dataset. From top to bottom are: original images, VA emotion labels and the distribution of 8 emotion labels.

Fig. 6. Image examples of different dominant emotion categories in the Image-Emotion-Social-Net dataset.

Fig. 7. Statistical distributions of the dataset.

images is shown in Figure 7(b). The distribution of emotion numbers for the images with more than 20 labels each is shown in Figure 7(c). Some image examples with assigned emotions in both CES and DES forms are given in Figure 5. We can find that the emotion perceptions of different users are truly subjective and personalized. More image examples with dominant emotions are given in Figure 6.

We also analyze the relation between the expected and personalized emotions. For each of the images with more than 20 labels, we compute the Euclidean distances between personalized emotions and expected emotion in VA space, and average all the distances. The histogram of the average VA distance is shown in Figure 8(a). For CES, we count the proportion of personalized emotions that are different from expected emotion for each image. The histogram of different emotion proportions is illustrated in Figure 8(b). It is clear that there exists great inconsistency between expected and personalized emotions.

The average and standard deviation of friend numbers among the 8k users are 45.7 and 21.4, respectively. Besides, users can be correlated by joining the same groups. So there exist rich social connections in the dataset. The time of the collected dataset ranges from Oct. 5, 2012 to Mar. 29, 2013, lasting about 6 months.

4.4 Challenging Tasks
The challenging tasks that can be performed by researchers on this dataset include, but are not limited to, the followings:

1) Image-centric emotion analysis. For each image, we can predict the dominant emotion category like the traditional affective image classification. Besides, we can also predict the emotion distribution of each image, taking the normalized emotion proportion as the ground truth.

2) User-centric emotion prediction. For each user, we can predict their personalized emotion perception of some specific images. The above two tasks can be extended to regression and retrieval tasks, all of which can be done using visual, social, temporal and the combination of all features.
3) Emotion related data mining and applications. This dataset contains visual and social information to support research on emotion influence mining, social advertising, affective image retrieval, emotional outbreak prediction and affective recommender systems, etc.

For different tasks, the roles of expected and personalized emotions are different. For image-centric expected emotion analysis, only the expected emotions can be used. For image-centric dominant emotion analysis or emotion distribution analysis, the expected emotions can be viewed as one type of personalized emotions. For user-centric emotion prediction, the expected emotions can also be viewed as one type of personalized emotions, but only for the uploaders of the related images. In this paper, we focus on the second task, trying to combine social, temporal and other factors with traditional visual features to predict PEP for each viewer.

5 OVERVIEW

Our goal is to predict the emotions of a specified user after viewing an image, associated with online social networks. Different types of factors can influence the emotion perception and can be exploited for emotion prediction. The following factors are hypothesized to contribute to emotion perceptions:

**Visual Content.** The visual content of an image can directly influence the emotion perception of viewers. Different from traditional emotion prediction, both the images viewed in the recent past and the current image are taken into account in our system. This point is also mentioned in the factor of temporal evolution. The visual descriptors used in this paper include low, middle, and high level features.

**Social Context.** One viewer’s emotion may be easily and largely affected by the social environment that they live in, e.g., their friends, so social context is highly helpful to make our prediction more accurate. Specifically, we consider the following social contexts: whether two users are in common interest groups, have common contact lists and the similarity of comments to same images.

**Temporal Evolution.** One viewer’s emotional variation within a short time is not so obvious, i.e., the current emotion is not independent on the recent past emotion, so temporal evolution gives additional information with respect to emotion prediction. Our system can take the recent past emotion into consideration.

**Location Influence.** Where and when a picture is taken is another factor which may contribute to emotional variation. One example is that photographs taken in entertainment venues usually lead people to feel happy. We encapsulate location information (if available) in our framework to improve the prediction performance.

We present rolling multi-task hypergraph learning (RMTHG) to jointly combine these factors. Formally, a user \( u_i \) in social networks observes an image \( x_{it} \) at time \( t \) and their perceived emotion after viewing the image is \( y_{it} \). Before viewing \( x_{it} \), the user \( u_i \) may have seen many other images. Among these images we select the recent past ones, which we believe to affect the current emotion. These selected images comprise a set \( S_i \). The emotional social network is formalized as a hybrid hypergraph \( \mathcal{G} = (\{u, x, S\}, \mathcal{E}, \mathcal{W}) \). Each vertex \( v = (u, x, S) \) in vertex set \( \mathcal{V} = \{u, x, S\} \) is a compound triple \( (u, x, S) \), where \( u \) represents user, \( x \) and \( S \) are the current image and the recent past images, which are named as ‘Target Image’ and ‘History Image Set’, respectively. It should be noted that in this triple, both \( x \) and \( S \) are viewed by user \( u \). \( \mathcal{E} \) is the hyperedge set. Each hyperedge \( e \) of \( \mathcal{E} \) represents a link between two or more vertices based on one component of the triple and is assigned with a weight \( w(e) \). \( \mathcal{W} \) is the diagonal matrix of the edge weights.

Mathematically, the task of personalized emotion prediction is to find the appropriate mapping

\[
f : (\mathcal{G}, y_{i1}, \ldots, y_{i(t-1)}) \rightarrow y_{it},
\]

for each user \( u_i \).

The framework of the proposed method is shown in Figure 9. First, we generate the compound triple vertex for each viewer based on the time of related images which they uploaded or commented. Second, the hypergraphs are constructed for each component of the triple based on different factors. Finally, we obtain the personalized emotion prediction results after the rolling learning of the multi-task hypergraphs.

6 HYPERGRAPH CONSTRUCTION

As stated in Section 5, a vertex of the proposed RMTHG is a compound one, which consists of three components. It should be emphasized that the images in such a vertex is a general concept, which not only refers to the pixel array itself, but also includes some additional information associated with the image, such as location, time, and emotional labels (if any) with respect to the specified user. For conventional emotion prediction, a vertex containing a pixel array is sufficient. In contrast, the compound vertex formulation enables our system to model all the four kinds of factors in Section 3: visual descriptors both in the target image and history image set can be extracted to represent visual content;
user relationship can be exploited from the user component to
take social context into consideration; past emotion can be inferred
from history image set to reveal temporal evolution; location influence
is embedded in the associated information with target image
and history image set. Consequently, such a vertex generation
mechanism is flexible and extensible.

In our hypergraph framework, hyperedges are used to indicate
some types of similarity among vertices, meaning that vertices
sharing a common hyperedge tend to have similar emotions.
Based on the vertex components, we construct different types
of hyperedges, including target image centric, history image set
centric, and user centric hyperedges.

6.1 Target Image Centric Hyperedges

6.1.1 Visual Content

As demonstrated in [6], the features that determine the emotions
of an image are different for various kinds of images. Similar
to [6], we extract commonly used visual features of different
levels and generalities for each image.

Low-level Features

Low-level features suffer from the difficulty of interpretation
and weak link to emotions [6]. However, they are useful as global
descriptors of the overall image content. Here we extract two
classes of low-level visual features. The first is the generic GIST
feature, which is one of the most commonly used features, for its
relatively powerful description ability of visual phenomena in a
scene perspective [64] [65].

The second class includes the special features derived from
elements of art, including color and texture [19]. Here the
color features include mean saturation and brightness, vector
based mean hue, emotional coordinates (pleasure, arousal and
dominance) based on brightness and saturation, colorfulness and
color names; while the texture features include Tamura texture,
Wavelet textures and gray-level co-occurrence matrix (GLCM)
based texture [19].

Mid-level Features

Mid-level features are more semantic, interpretable and have
stronger link to emotions than low-level features [2]. Here we
extract two classes of mid-level features. The first is attribute based
representation that has been widely studied in recent years for its
intuitive interpretation and cross-category generalization property
in visual recognition domain [65], [66], [67], [68]. We extract
102 dimensional attributes which are commonly used by humans
to describe scenes as mid-level generic features. As in [65], the
attributes can be classified into five types: materials (mental),
surface properties (dirty), functions or affordances (reading), s-
patial envelop attributes (cluttered) and object presence (flowers).
GIST features and SVM implemented in Liblinear toolbox5 are
used to train attribute classifiers based on 14,340 images in SUN
database [64].

The second class is the specific mid-level features inspired
from the principles-of-art, including balance, contrast, harmony,
variety, gradation, and movement [2]. As the guidelines and
tools, these artistic principles are used to arrange and orchestrate
artistic elements in art theory for describing specific semantics
and emotions; they are found to have stronger link to emotions


6.1.2 Location Influence

Besides the image itself, there are often other metadata of social
images, such as the taken time, the location, etc. We consider the
location influence here, as the images taken around the similar
place within a short time tend to describe similar events and
express similar emotions. The geographical similarity between
\(v_{\text{t1}}\) and \(v_{\text{t2}}\) with locations \(l(x_{\text{t1}}) = (\text{lat}_{\text{t1}}, \text{lon}_{\text{t1}})\) and
\(l(x_{\text{t2}}) = (\text{lat}_{\text{t2}}, \text{lon}_{\text{t2}})\) (if available) is measured by the

![Fig. 10. Illustration of target image centric hyperedge construction.](image-url)
Harversion formula [49], which is computed by
\[
s_H^T(v_{it_1}, v_{jt_2}) = 1 - 2 \arctan^2(\sqrt{\phi}, \sqrt{1 - \phi}),
\]
\[
\phi = \sin^2\left(\frac{\Delta \text{lat}}{2}\right) + \cos(\text{lat}_{it_1}) \cos(\text{lat}_{jt_2}) \sin^2\left(\frac{\Delta \text{lon}}{2}\right),
\]
\[
\Delta \text{lat} = \text{lat}_{jt_2} - \text{lat}_{it_1}, \Delta \text{lon} = \text{lon}_{jt_2} - \text{lon}_{it_1}.
\]

Figure 10(b) illustrates the construction procedure of location based hyperedges.

6.2 History Image Set Centric Hyperedges

Similar to target images, we can construct hyperedges for the history image sets based on the visual features and known locations. Besides, the emotion labels of the history image sets are known. So we can also construct hyperedges based on this information by using the normalized emotion distribution to explore the temporal influence of emotions.

As each history image set is composed of different numbers of sequential images, we use dynamic time warping (DTW) [72] to measure the distance between two history image sets. For pairwise images from two history image sets, the Euclidean distance is used to measure the visual distance and the visual similarity can be obtained using Eq. (2). Pairwise location similarity is computed by the Harversion formula. Similar to Plutchik’s wheel [73], pairwise emotion distance is defined as \(1 + \frac{\text{the number of steps}}{\text{the number of emotion groups}}\).

6.3 User Centric Hyperedges

In social networks, the emotions perceived by one user can be easily influenced by their friends. In our dataset, there are various kinds of social connections. The social similarity between two users \((u_i, u_j)\) is measured by
\[
s_S(v_{it_1}, v_{jt_2}) = \begin{cases} 1, & \text{if } u_i = u_j, \\ \frac{1}{3}(g_S + c_S + f_S), & \text{otherwise}, \end{cases}
\]
where \(g_S(u_i, u_j)\) is defined as the ratio between the common groups that both users join, and the groups that either of them joins,
\[
g_S(u_i, u_j) = \frac{|g(u_i) \cap g(u_j)|}{|g(u_i) \cup g(u_j)|},
\]
where \(g(u_i)\) is the set group that \(u_i\) joins, \(|x|\) is the element number of a set. Similar to \(g_S\), \(c_S(u_i, u_j)\) is defined as the ratio between the common contact lists that both users follow, and the contact lists that either of them follows.

\[
f_S(u_i, u_j) = \frac{1}{M} \sum_{k=1}^{M} s(BoW^i_k, BoW^j_k),
\]
where \(M\) is the number of images that both users comment on, \(BoW^i_k\) is the BoW feature of the \(k\)th comment from user \(u_i\), \(s(., .)\) is the cosine function that computes the similarity of two BoW features.

7 ROLLING MULTI-TASK HYPERGRAPH LEARNING

Given the emotional hybrid hypergraph \(\mathcal{H} = (\mathcal{U}, \mathcal{X}, \mathcal{S}, \mathcal{E}, \mathcal{W})\), we obtain the incidence matrix \(\mathbf{H}\) by computing each entry as,
\[
h(v, e) = \begin{cases} 1, & \text{if } v \in e, \\ 0, & \text{otherwise}. \end{cases}
\]

The vertex degree of vertex \(v \in \mathcal{V}\) and the edge degree of hyperedge \(e \in \mathcal{E}\) are defined as \(d(v) = \sum_{e \in \mathcal{E}} w(e) h(v, e), \delta(e) = \sum_{v \in \mathcal{V}} h(v, e)\). According to \(d(v)\) and \(\delta(e)\), we define two diagonal matrices \(\mathbf{D}_v\) and \(\mathbf{D}_e\) as \(\mathbf{D}_v(i, i) = d(v_i)\) and \(\mathbf{D}_e(i, i) = \delta(e_i)\).

Given users \(u_1, \ldots, u_N\) and the related images, our objective is to explore the correlation among all involved images and the user relations. Suppose the training vertices and the training labels are \(\{(u_1, x_{1j}, S_{1j})\}_{j=1}^m = \{(v_{uN}, x_{Nj}, S_{Nj})\}_{j=1}^m\) and \(\mathbf{Y}_1 = [y_{11}, \ldots, y_{1m}]^T, \ldots, \mathbf{Y}_N = [y_{N1}, \ldots, y_{Nm}]^T\), and the to-be-estimated relevance values of all images related to the specified users are \(\mathbf{R}_1 = [R_{11}, \ldots, R_{1m}], \ldots, \mathbf{R}_N = [R_{N1}, \ldots, R_{Nm}]^T\). We denote \(\mathbf{Y}\) and \(\mathbf{R}\) as
\[
\mathbf{Y} = [\mathbf{Y}_1^T, \ldots, \mathbf{Y}_N^T]^T, \mathbf{R} = [\mathbf{R}_1^T, \ldots, \mathbf{R}_N^T]^T.
\]

The proposed rolling multi-task learning can be conducted as a semi-supervised learning to minimize the empirical loss and the regularizer on the hypergraph structure simultaneously by
\[
\arg \min_{\mathbf{R}} \{\Gamma + \lambda \Psi\},
\]
where \(\lambda\) is a trade-off parameter, \(\Gamma\) is the empirical loss defined by
\[
\Gamma = ||\mathbf{R} - \mathbf{Y}||^2,
\]
and \(\Psi\) is the regularizer on the hypergraph structure defined by
\[
\Psi = \frac{1}{2} \sum_{e \in \mathcal{E}, \mu, \nu \in \mathcal{V}} w(e) h(\mu, e) h(\nu, e) \left(\frac{\mathbf{R}(\mu)}{\sqrt{\mathbf{D}(\mu, \mu)}} - \frac{\mathbf{R}(\nu)}{\sqrt{\mathbf{D}(\nu, \nu)}}\right)^2
\]
\[
= \mathbf{R}^T (I - D_v^{-1/2} \text{HWD}_v^{-1} \text{H}^T D_v^{-1/2}) \mathbf{R}.
\]
By setting the derivative of Eq. (9) with respect to \(\mathbf{R}\) to zero, we can derive the solution for the objective function Eq. (9)
\[
\mathbf{R} = (I + \frac{1}{\lambda} \Delta)^{-1} \mathbf{Y},
\]
where \(\Delta\) is the hypergraph Laplacian [54], [56], defined as
\[
\Delta = I - D_v^{-1/2} \text{HWD}_v^{-1} \text{H}^T D_v^{-1/2}.
\]

By using the relevance scores in \(\mathbf{R}\), we can rank the related images for each user. The top results with high relevance scores are assigned with related emotion category. Suppose the predicted results of the test images are \(\hat{\mathbf{E}} = F(\mathbf{R})\) based on relevance score
Algorithm 1: Learning procedure for RMTHG

Input: Error threshold $\varepsilon$, regularization $\lambda$, max-epochs $K$, training labels $Y$

Output: Predicted emotion labels $\hat{E}$

1. Initialization: $\hat{E}^{(0)} \leftarrow \emptyset$; Construct hypergraph $G^{(0)} = \langle \{U, X, S\}, E^{(0)}, W^{(0)} \rangle$;
2. for $k \leftarrow 1$ to $K$ do
   3. Compute $H^{(k-1)}$, $D^{(k-1)}$, $D^{(k-1)}$ from $G^{(k-1)}$;
   4. Compute $\Delta^{(k-1)}$ by Eq.(13) using $H^{(k-1)}$, $D^{(k-1)}$, $D^{(k-1)}$;
   5. $R^{(k)} \leftarrow (I + \frac{1}{\lambda} \Delta^{(k-1)})^{-1}Y$;
   6. $\hat{E}^{(k)} \leftarrow F(R^{(k)});$
   7. if $||\hat{E}^{(k)} - E^{(k-1)}||_2 < \varepsilon$ then
      break;
   8. end
9. Update the hypergraph $G^{(k)} = \langle \{U, X, S\}, E^{(k)}, W^{(k)} \rangle$ based on the emotions of history image set $s_H^{(k)}$;
10. end
11. return $\hat{E}^{(k)}$.

threshold function $F$, which is simply ‘of top M results’ [23], [56], where M is selected by the average F1. We can then iteratively update Eq. (12) until convergence, as shown in Algorithm 1. We call this method rolling multi-task hypergraph learning, because it simultaneously classifies the emotions of a specified category for $N$ users in a hypergraph framework and the learning process is iteratively updated until convergence.

8 EXPERIMENTS

To evaluate the effectiveness of the proposed RMTHG method for personalized image emotion prediction, we carried out emotion classification experiments on the IESN dataset. We also designed one novel application using the predicted emotions.

8.1 Experimental Settings

As users upload or comment on images in chronological order and the perceived emotions can be influenced temporally, we split the dataset into a training set and a test set based on the uploading time and the commenting time of related images. The first set covering 50% of images of each viewer is used for training and the rest is used for test. As there are about 8,000 users in the dataset, we randomly split them into 80 groups to facilitate fast computation and save memory. Each time we conduct experiments on one group and eventually we computed the average performance and the standard deviation of all experiments.

8.1.1 Baselines

For emotion classification, we used four state-of-the-art classifiers as baselines: (1) Naive Bayes (NB), (2) Support Vector Machine (SVM) with RBF kernel, which are commonly used for traditional affective image classification in [19] and [2], [32], respectively; (3) Graph Model (GM), which is used for personalized emotion prediction [13]; and (4) Hypergraph learning (HG), which is used for brand data gathering [23]. In GM, the social factors in [13] and our visual features are combined. In HG, only the “Target Image” component of RMTHG is modelled using visual and location features. The social context features makes no sense since for one user all the images are connected together with the same weight.

For emotion regression, we tested the performances of Support Vector Regression (SVR) as in [2], [32] and multiple linear regression (MLR). It is difficult for these methods to model features like social context, so we just used visual features. Different kernels were tested for emotion regression using SVR. How to effectively combine the different factors for emotion analysis in a regression framework remains our future work.

8.1.2 Evaluation Criteria

For emotion classification, we used precision, recall and F1-Measure to evaluate the performance of different methods:

- **Precision (Pre)**. Precision measures the accuracy of emotion classification and is computed by:
  \[
  \text{Pre} = \frac{\# \text{Correctly Classified Relevant Images}}{\# \text{All Classified Relevant Images}},
  \]
  where $\# \text{Correctly Classified Relevant Images}$ is the number of relevant images that are correctly classified, and $\# \text{All Classified Relevant Images}$ is the total number of images that are classified as relevant ones.

- **Recall (Rec)**. Recall measures the data coverage of relevant images for a specified emotion category, and is calculated by:
  \[
  \text{Rec} = \frac{\# \text{Correctly Classified Relevant Images}}{\# \text{All Relevant Images}},
  \]
  where $\# \text{All Relevant Images}$ is the number of relevant images in the dataset for a specified emotion category.

- **F1-Measure (F1)**. F1-Measure jointly considers Recall and Precision, which is defined as:
  \[
  F1 = \frac{2 \times \text{Pre} \times \text{Rec}}{\text{Pre} + \text{Rec}}.
  \]

For emotion regression, we used mean squared error\(^7\) as the evaluation criteria:

- **Mean squared error (MSE)**. Mean squared error measures the difference of the predicted values and the real values of VAD, and is calculated by:
  \[
  \text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{E}_i - E_i)^2,
  \]
  where $n$ is the number of test images, $E_i$ and $\hat{E}_i$ are the real values and estimated values of valence, arousal or dominance, respectively.

All the four measurements above range from 0 to 1. Higher values of Pre, Rec and F1 represent better performances for emotion classification, while lower MSE indicates better performance for emotion regression.

8.2 Personalized Emotion Classification

8.2.1 On Visual Features

First, we conducted experiments to compare the performance of different visual features for personalized emotion classification; and used SVM and NB as baselines. For comparison, we used a simple version of GM, HG and RMTHG that consider only visual features, abbreviated as GM(V), HG(V) and RMTHG(V). The

average performances on emotion classification using different visual features and methods are given in Figure 12. The performances of F1 for every emotion category are shown in Figure 13.

From the results, we can observe that: (1) generally, the fusion of different kinds of visual features outperforms the use of only single feature, possibly because it can utilize the advantages from different aspects; (2) the proposed hypergraph learning method greatly outperforms the baselines on almost all features; for the fusion of different features, the proposed RMTHG(V) method achieves 65.2%, 62.2%, 22.3% and 16.0% performance gains on F1 as compared to SVM, NB, GM(V) and HG(V), respectively; (3) for the 8 emotion categories, the most discriminative features are different; but overall the high-level and mid-level features have stronger discriminability than low-level ones, which is consistent with the conclusions in [6]; (4) the positive emotions are more accurately modelled than the negative ones by almost all the four methods; and (5) the overall precision, recall and F1 are still very low, indicating that only using visual features to classify personalized image emotions is not sufficient.

8.2.2 On Different Factors

Besides, we evaluated the influence of different factors in the proposed method. We computed the performance of the proposed method without visual content (RMTHG-V), without social context (RMTHG-S), without temporal evolution (RMTHG-T), and without location influence (RMTHG-L). RMTHG-T means that the history image sets of all vertices are empty. Here all the visual features are considered. Meanwhile, we compared the performances of GM(V) and GM (fused visual features and social factors [13]), HG(V) and HG (fused visual features and location information of the “Target Image” component of RMTHG). Figure 14 shows the average performance on emotion classification, while Figures 15, 16 and 17 present the performances of Precision, Recall and F1 for every emotion category.

From the results, we can observe that: (1) for the GM method, the combination of the social factors in [13] can improve the performance, though the improvement is not that significant; (2) for the HG method, the performance is improved with the help of location information; (3) by incorporating the various factors, the classification performance of RMTHG is greatly improved; compared to RMTHG(V), the improvement of RMTHG on precision, recall and F1 are 49.4%, 9.5% and 33.2%, respectively, which indicates that the social emotions are jointly affected by these factors; (4) the proposed RMTHG method significantly outperforms the four baselines, achieving 120.0%, 116.0%, 54.9% and 49.2% performances gains on F1 for SVM, NB, GM and HG, respectively; (5) the contributions of the various factors in the proposed RMTHG to emotion perceptions are different; on average, the discriminability order of these factors is visual content > social context > temporal evolution > location influence; and (6) even without using the visual content, we can still get competitive results (by comparing RMTHG(V) and RMTHG-V), which demonstrates that the social features and the temporal evolution play an important role in emotion perception.

To better illustrate the help of different factors in predicting the correct emotions, we give some visual examples, as shown in Figure 18. From this example, we can easily understand how these factors work in personalized emotion perception prediction.
8.2.3 Case Studies

Here we show some interesting cases to demonstrate the effectiveness of the proposed method. By leveraging the statistics on the performance improvement and different factor influences, we have the following discoveries.

The performance improvement in the prediction of negative emotions is higher than positive ones. Comparing the prediction results of RMTHG(V) and RMTHG in Figure 17, we can observe that when taking temporal evolution, social context and location influence into account, the performance of negative emotion classification improves more significantly than position ones. This observation means that these context factors play a more important role in personalized perception of negative emotions, indicating that the influence of negative emotions is likely to be stronger than positive ones, which is consistent with [74]. Similar conclusion can be obtained when comparing GM(V) and GM.

Sadness is one special category that the influence of temporal evolution is larger than social context. From Figure 17, it is easy to see that the performance gain of social context is larger than temporal evolution for all the 4 positive emotions and the 3 negative emotions of anger, disgust and fear, while it is the opposite for sad emotion. This is probably because sadness tends to be a long lasting emotion [75], [76], which may last up to 240 times longer than surprise, fear, etc., as what spurred the feeling is often of greater importance [76]. So sadness requires one to make meaning of the event and cope with the new situation which takes time, while fear, disgust or awe following event offset often lacks purpose [77].

Stronger social connections tend to have more influences on performance improvement. We randomly selected 100 users from the dataset and collected the images they uploaded or commented together with various metadata. For each of the 100 users, we first obtained the average social similarity in the constructed hypergraph and then computed the performance gains of F1 with the help of social context. The relation is depicted in Figure 19. It seems that with the increase of social connections, the overall trend of performance gains is growing, which indicates that stronger social connections correlate with higher performance gains. The results accord with the findings in [78], which concluded that emotions are influenced by social context and that the stronger the connection is, the larger the influence is. This fact also demonstrates the basic hypothesis of social context’s contribution to emotion perceptions.

8.3 Personalized Emotion Regression

MLR performs the worst with MSE much greater than 10, which indicates that the linear regression is not appropriate for emotion regression. The results of average MSE on valence, arousal and dominance using SVR are shown in Figure 20. It is clear that valence is regressed the worst, while arousal and dominance are relatively better predicted. SVR performs much better than MLR, which demonstrates its effectiveness on emotion regression. Besides, the RBF kernel works slightly better than Sigmoid kernel.

8.4 Application: Affective Curve Visualization

Given a specified user, we can predict the personalized emotion perceptions of the current image based on the proposed method,
which simultaneously considers various factors, including visual content, social context, temporal evolution and location influence. We can then draw an affective curve for each user to learn about the process of their emotion changes. We use VAD models to represent both different dimensions of emotions for better visualization. By adding related images to the affective curve, we can clearly understand what emotion is perceived by the specified user and what kind of image causes this perception. Besides, by comparing the affective curves, we can also analyze whether this emotion is influenced by their friends and to what extent this emotion is influenced. Figure 21 shows two examples of visualized affective curves. We can observe that the perceived emotions of both users are relatively stable within a short time, which corroborates the hypothesis that the perceived emotions are temporally evolved. In such cases, we can construct emotion based image storyline for these users and concentrate on the turning point images to understand what causes each variation in emotion. Figure 22 shows the image storyline of User A together with the turning point images and corresponding behaviors. For clarity, we just use valence to represent the pleasantness of emotions. From this figure, we can clearly see the emotional status change of a soldier before, in and after war.

9 Conclusion and Future Work

In this paper, we proposed to predict personalized perceptions of image emotions by incorporating various factors (social context, temporal evolution, and location influence) with visual content. Rolling multi-task hypergraph learning was presented to jointly combine these factors. A large-scale personalized emotion dataset of social images was constructed and some baselines were provided. Experimental results on personalized emotion classification demonstrated that the performance of the proposed method is superior over the state-of-the-art approaches. The predicted personalized emotions can be used to develop various applications, such as affective curve visualization.

For further studies, we will try to mine group emotions on the Image-Emotion-Social-Net. Modelling social connections of users dynamically and considering interest prior by mining related personal profile may improve the performance of emotion prediction. In addition, we can model the different emotions in a multi-task learning framework to explore the emotion relations. How to extend the proposed method for personalized affective image regression is also worth studying.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 61472103, 61133003, 61571269, 61271394, 61671267), and partially supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office. The authors would also like to thank Wenlong Xie and Xiaolei Jiang both from Harbin Institute of Technology for their valuable suggestion and help.

REFERENCES


