User-Centric Affective Computing of Image Emotion Perceptions

Sicheng Zhao, Hongxun Yao, Wenlong Xie, Xiaolei Jiang
School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China.
zsc@hit.edu.cn; h.yao@hit.edu.cn; wlxie@hit.edu.cn; xljiang@hit.edu.cn
Appendix: https://sites.google.com/site/schzhao/

Abstract

We propose to predict the personalized emotion perceptions of images for each viewer. Different factors that may influence emotion perceptions, including visual content, social context, temporal evolution, and location influence are jointly investigated via the presented rolling multi-task hypergraph learning. For evaluation, we set up a large scale image emotion dataset from Flickr, named Image-Emotion-Social-Net, with over 1 million images and about 8,000 users. Experiments conducted on this dataset demonstrate the superiority of the proposed method, as compared to state-of-the-art.

Motivation

Images can convey rich semantics and evoke strong emotions in viewers. Most existing works on image emotion analysis tried to find features that can express emotions better to bridge the affective gap (Zhao et al. 2014). These methods are mainly image centric, focusing on the dominated emotions for the majority of viewers.

However, the emotions that are evoked in viewers by an image are highly subjective and different (Zhao et al. 2015), as shown in Figure 1. Therefore, predicting the personalized emotion perceptions for each viewer is more reasonable and important. In such cases, the emotion prediction task becomes user centric. We build a large-scale dataset for this task and classify personalized emotion perceptions.

The Image-Emotion-Social-Net Dataset

We set up the first large-scale dataset on personalized image emotion perception, named Image-Emotion-Social-Net, with over 1 million images downloaded from Flickr. To get the personalized emotion labels, firstly we use traditional lexicon-based methods to obtain the text segmentation results of the title, tags and descriptions from uploaders for expected emotions and the comments from viewers for actual emotions. Then we compute the average value of valence, arousal and dominance (VAD) of the segmentation results as ground truth for dimensional emotion representation based on the VAD norms of 13,915 English lemmas (Warriner, Kuperman, and Brysbaert 2013). See Appendix A1 for details.

Title: Light in Darkness
Tags: london, stormy, dramatic weather, …
Description: …there was a break in the clouds … while the sky behind remained somewhat foreboding. I thought it made for a pretty intense scene….

Expected emotion:
Emotion category: fear
Sentiment: negative
Valence: 4.1956
Arousal: 4.4989
Dominance: 4.8378

(a) Original image (b) Image metadata (c) Expected emotion

Comments to this image from different viewers	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 V: 7.950 A: 6.950 D: 7.210
Hey, it really frightened me! My little daughter just looked scared.	Emotion: fear, Sentiment: negative V: 2.625 A: 5.805 D: 3.625

(d) Personalized emotion labels

Figure 1: Illustration of personalized image emotion perceptions in social networks. Related emotions are obtained using the keywords in red.

Problem Definition

Formally, a user u_i in social networks observes an image x_{it} at time t, and her perceived emotion after viewing the image is y_{it}. Before viewing x_{it}, the user u_i may have seen many other images. Among them we select the recent past ones, which are believed to influence the current emotion. These selected images comprise a set S_i. The emotional social network is formulized as a hybrid hypergraph $\mathcal{G} = \langle \{\mathcal{U}, \mathcal{X}, \mathcal{S}\}, \mathcal{E}, \mathbf{W} \rangle$. Each vertex v in vertex set $\mathcal{V} = \{\mathcal{U}, \mathcal{X}, \mathcal{S}\}$ is a compound triple (u, x, S), where u represents user, x and S are the current image and the recent past images, named as 'Target Image' and 'History Image Set', respectively. Note that in this triple, both x and S are viewed by user u. $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the hyperedge set. Each hyperedge e of $\mathcal{E}$ represents a link between two vertexes based on one component of the triple and is assigned with a weight $w(e)$. $\mathbf{W}$ is the diagonal matrix of the edge weights.

Mathematically, the task of personalized emotion prediction is to find the appropriate mapping for each user u_i

$$f : (\mathcal{G}, y_{i1}, ..., y_{i(t-1)}) \rightarrow y_{it}. \quad (1)$$

Rolling Multi-Task Hypergraph Learning

Rolling multi-task hypergraph learning (RMTHG) is presented to jointly combine the various factors that may influence emotion perception: visual content, social context, temporal evolution, and location influence. The framework

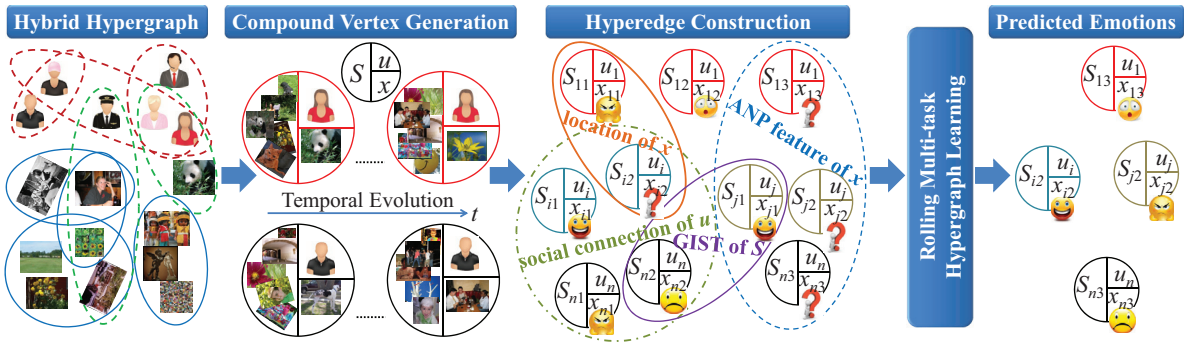

Figure 2: The framework of the proposed method for personalized image emotion prediction.

is shown in Figure 2. The compound vertex formulation enables our system to model all factors. Different types of hyperedges can be constructed for the three components of the compound vertex. Please see Appendix A2 for details.

Given N users $u_1, \ldots, u_N$ and related images, our objective is to explore the correlation among all involved images and the user relations. Suppose the training vertexes are $\{(u_1, x_{1j}, S_{1j})\}_{j=1}^{m_1}, \cdots, \{(u_N, x_{Nj}, S_{Nj})\}_{j=1}^{m_N}$, the training labels are $\mathbf{Y}_1 = [y_{11}, \cdots, y_{1m_1}]^{\mathrm{T}}, \ldots, \mathbf{Y}_N = [y_{N1}, \cdots, y_{Nm_N}]^{\mathrm{T}}$, the to-be-estimated relevance values of all images related to specified users are $\mathbf{R}_1 = [R_{11}, \cdots, R_{1m_1}]^{\mathrm{T}}, \ldots, R_N = [R_{N1}, \cdots, \mathbf{R}_{Nm_N}]^{\mathrm{T}}$. Let

$$\mathbf{Y} = [\mathbf{Y}_1^{\mathrm{T}}, \cdots, \mathbf{Y}_N^{\mathrm{T}}]^{\mathrm{T}}, \mathbf{R} = [\mathbf{R}_1^{\mathrm{T}}, \cdots, \mathbf{R}_N^{\mathrm{T}}]^{\mathrm{T}}. \quad (2)$$

The proposed RMTHG can be conducted as a semi-supervised learning to minimize the empirical loss and the regularizer on the hypergraph structure simultaneously by

$$\arg\min_{\mathbf{R}} \{\Psi + \lambda\Gamma\}, \quad (3)$$

where λ is a trade-off parameter, $\Gamma = ||\mathbf{R} - \mathbf{Y}||^2$ is the empirical loss and $\Psi = \mathbf{R}^{\mathrm{T}}\big(\mathbf{I} - \mathbf{D}_v^{-1/2}\mathbf{H}\mathbf{W}\mathbf{D}_e^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{D}_v^{-1/2}\big)\mathbf{R}$ is the regularizer on the hypergraph structure. $\mathbf{H}$ is the incidence matrix, $\mathbf{D}_v$ and $\mathbf{D}_e$ are two diagonal matrices with the diagonal elements denoting the vertex and edge degrees.

By setting the derivative of Equ. (3) with respect to $\mathbf{R}$ to zero, the solution of Equ. (3) can be achieved by

$$\mathbf{R} = \big(\mathbf{I} + \frac{1}{\lambda}\Delta\big)^{-1}\mathbf{Y}, \quad (4)$$

where $\Delta = \mathbf{I} - \mathbf{D}_v^{-1/2}\mathbf{H}\mathbf{W}\mathbf{D}_e^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{D}_v^{-1/2}$ is the hypergraph Laplacian. 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 $\hat{\mathbf{E}} = F(\mathbf{R})$, we can iteratively update Equ. (4) based on the emotion of history image set until convergence.

Experiments

We use precision, recall and F1-Measure to evaluate the performance of different methods. The first involved 50% images of each viewer in the Image-Emotion-Social-Net dataset are used for training and the rest are used for test. Naive Bayes (NB), Support Vector Machine (SVM) with RBF kernel and Graph Model (GM) are adopted as baseline methods. Please see Appendix A3 for detailed results.

We firstly test the performances using different visual features and learning models for the 8 emotion categories. From the results, we can observe that: (1) Generally, the fusion of different features outperforms single feature based method, possibly because it can utilize the advantages from different aspects; (2) The proposed RMTHG greatly outperforms the baselines on almost all features.

We also evaluate the influence of different factors on emotion classification, by comparing the performance with all factors and that without one factor. Here all the visual features are considered. From the results, we can see that (1) By incorporating the various factors, the classification performance is greatly improved, which indicates that the social emotions are affected by these factors; (2) Even without using the visual content, we can still get competitive results, which demonstrates that the social features and the temporal evolution play an important role in emotion perception.

Conclusion

In this paper, we proposed to predict personalized perceptions of image emotions by incorporating various factors 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 demonstrated that the performance of the proposed method is superior over the state-of-the-art approaches.

This work was supported by the National Natural Science Foundation of China (No. 61472103 and No. 61133003).

References

Warriner, A. B.; Kuperman, V.; and Brysbaert, M. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods* 45(4):1191–1207.

Zhao, S.; Gao, Y.; Jiang, X.; Yao, H.; Chua, T.-S.; and Sun, X. 2014. Exploring principles-of-art features for image emotion recognition. In *ACM MM*, 47–56.

Zhao, S.; Yao, H.; Jiang, X.; and Sun, X. 2015. Predicting discrete probability distribution of image emotions. In *IEEE ICIP*.