Jesse S. Jin · Changsheng Xu · Min Xu

The Era of Interactive Media

Springer

The Era of Interactive Media

Jesse S. Jin • Changsheng Xu • Min Xu

The Era of Interactive Media

Springer

Jesse S. Jin
University of Newcastle
Callaghan, NSW, Australia

Changsheng Xu
Chinese Academy of Science
Beijing, P.R. China

Min Xu
University of Technology
Sydney Broadway, NSW, Australia

ISBN 978-1-4614-3500-6 ISBN 978-1-4614-3501-3 (eBook)
DOI 10.1007/978-1-4614-3501-3
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012945429

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Pacific-Rim Conference on Multimedia (PCM) is a major annual international conference organized as a forum for the dissemination of state-of-the-art technological advances and research results in the fields of theoretical, experimental, and applied multimedia analysis and processing. It brings together researchers, developers, and educators in the field of multimedia from around the world. Since the first PCM was held in Sydney in 2000, it had been held successfully around the Pacific Rim, including Beijing in 2001, Hsinchu in 2002, Singapore in 2003, Tokyo in 2004, Jeju in 2005, Zhejiang in 2006, Hong Kong in 2007, Tainan in 2008, Bangkok in 2009, and Shanghai in 2010. After 10 years, PCM came back to Sydney in 2011.

PCM 2011 was the 12th conference in this highly successful and increasingly influential series and was held from 20 to 22 December 2011 at University of Technology, Sydney, Australia. The technical program featured opening keynote addresses, invited plenary talks, tutorials, and technical presentations of refereed papers. This year, we received 113 submissions and accepted 59 papers for oral presentations. The papers were carefully peer-reviewed. The accept rate of PCM 2011 was 52%. Papers in this volume covered a range of pertinent topics in the field including face detection, recognition, and synthesis; video coding and transmission; audio, image, and video quality assessment; audio and image classification; stereo image and video analysis; object detection, action recognition, and surveillance; visual analysis and retrieval; watermarking and image processing and applications.

PCM 2011 could never have been successful without the support of ARC Network in Enterprise Information Infrastructure (EII) and University of Technology, Sydney (UTS). We would also like to thank all the committee members, the keynote speakers, and the tutorial speakers. Our thanks must go to the reviewers who generously spent their time and provided valuable comments. And at the end, we would like to thank all the authors for submitting their work to PCM 2011.

NSW, Australia — Jesse S. Jin
Beijing, China — Changsheng Xu
NSW, Australia — Min Xu

Organization

General Chairs

- Jesse Jin, Newcastle, Australia
- Yong Rui, MSRA, Microsoft, USA
- Xiaofang Zhou, UQld, Australia

Local Organising Chairs

- Massimo Piccardi, UTS, Australia
- Sean He, UTS, Australia
- Maolin Huang, UTS, Australia

Honorary Chairs

- Thomas Huang, UIUC, USA
- Gerard Medioni, USC, USA
- Shih-Fu Chang, Columbia, USA

Program Chairs

- Ernest Edmonds, UTS, Australia
- Heng Tao Shen, UQld, Australia
- Changsheng Xu, CAS, China

Organising Committee

Panel Chairs

- Svetha Venkatesh, Curtin
- Anthony Maeder, UWS
- David Taubman, UNSW

Tutorial Chairs

- Mohan Kankanhalli, NUS
- Manzur Murshed, Monash
- Phoebe Chen, Deakin

Publication Chairs

- Suhuai Luo, Newcastle
- David Tien, CSU

Special Session Chairs

- Ishwar Sethi, Oakland
- Tao Mei, MSRA, Microsoft
- Richard Xu, CSU

Publicity Chairs

- Edward Chang, Google, USA
- Sameer Singh, Loughborough
- Mark Liao, ISAS, Taiwan
- Shuqiang Jiang, CAS, China
- Stephan Chalup, Newcastle
- Zheng-Jun Zha, NUS, S'pore
- Supot Nitsuwat, KMU, Thailand

Technical Committee Members

Bo Geng, Peking University, China
Cees Snoek, University of Amsterdam, Netherlands
Dong Liu, Columbia University, USA
Fei Wu, Zhejiang University, China
Gang Hua, IBM T.J. Watson Research Center, USA
Homer Chen, National Taiwan University, Taiwan
Hong Lu, Fudan University, China
Jialie Shen, Singapore Management University, Singapore
Jian Cheng, Institute of Automation, Chinese Academy of Sciences, China
Jie Shao, University of Melbourne, Australia
Jing Liu, Institute of Automation, Chinese Academy of Sciences, China
Jinhui Tang, Nanjing University of Science and Technology, China
Jinqiao Wang, Institute of Automation, Chinese Academy of Sciences, China
Jizheng Xu, Microsoft.com
Kemal Ugur, Nokia Research
Kuan-Ta Chen, Academia Sinica, Taiwan
Lexing Xie, Australian National University, Australia

Ling-Yu Duan, Peking University, China
Marco Bertini, University of Florence, Italy
Margrit Gelautz, Vienna University of Technology
Meng Wang, National University of Singapore
Nicu Sebe, Universtiy of Trento Italy
Qi Tian, University of Texas at San Antonio, USA
Qingshan Liu, Rutgers University, USA
Qionghai Dai, Tsinghua University, China
Ravindra Guntur, National University of Singapore
Richang Hong, Hefei University of Technology, China
Roger Zimmermann, National University of Singapore
Ruigang Yang, University of Kentucky
Shin'ichi Satoh, National Institue of Informatics, Japan
Shuicheng Yan, National University of Singapore
Shuqiang Jiang, Chinese Academy of Sciences
Tao Mei, Microsoft Research Asia, China
Xiangyang Xue, Fudan University, China
Xiao Wu, Southwest Jiaotong University, China
Xiaokang Yang, Shanghai Jiaotong University
Xinbo Gao, Xidian University, China
Yan Liu, Hong Kong Polytechnic University, China
Yantao Zheng, Institute for Infocomm Research, Singapore
Yao Zhao, Beijing Jiaotong University, China
Yi Yang, Carnegie Mellon University, USA
Yugang Jiang, Fudan University
Zhengjun Zha, National University of Singapore
Zhong Wu, Microsoft, USA
Zi Huang, University of Queensland, Australia

Contents

Audio, Image and Video Quality Assessment

Audio and Image Classification

Stereo Image and Video Analysis

Visual Analysis and Retrieval

Watermarking and Image Processing

Applications

Best Papers and Runner-ups

Image Re-Emotionalizing

Mengdi Xu, Bingbing Ni, Jinhui Tang, and Shuicheng Yan

Abstract In this work, we develop a novel system for synthesizing user specified emotional affection onto arbitrary input images. To tackle the subjectivity and complexity issue of the image affection generation process, we propose a learning framework which discovers emotion-related knowledge, such as image local appearance distributions, from a set of emotion annotated images. First, emotion-specific generative models are constructed from color features of the image super-pixels within each emotion-specific scene subgroup. Then, a piece-wise linear transformation is defined for aligning the feature distribution of the target image to the statistical model constructed from the given emotion-specific scene subgroup. Finally, a framework is developed by further incorporation of a regularization term enforcing the spatial smoothness and edge preservation for the derived transformation, and the optimal solution of the objective function is sought via standard non-linear optimization. Intensive user studies demonstrate that the proposed image emotion synthesis framework can yield effective and natural effects.

Keywords Re-emotionalizing • Linear piece-wise transformation • GMM

M. Xu (✉) • S. Yan
National University of Singapore, Vision and Machine Learning Lab, Block E4, #08-27, Engineering Drive 3, National University of Singapore, 117576 Singapore
e-mail: g0900224@nus.edu.sg; eleyans@nus.edu.sg

B. Ni
Advanced Digital Sciences Center, 1 Fusionopolis Way, Connexis North Tower 08-10, 138632 Singapore
e-mail: bingbing.ni@adsc.com.sg

J. Tang
Nanjing University of Science and Technology, School of Computer Science, Xiao Ling Wei 200, Nanjing, 210094 China
e-mail: jinhuitang@mail.njust.edu.cn

J.S. Jin et al., *The Era of Interactive Media*,
DOI 10.1007/978-1-4614-3501-3_1,

1 Introduction

Images may affect people into different emotions. For example, a photo taken in a rainy day looking at a dark street will usually give one a feeling of sadness; while a picture of a sunshine beach will mostly make people delighted.

Throughout the decade, the multimedia research community has shown great interest in affective retrieval and classification of visual signals (digital media). Bianchi-Berthouze [2] proposed an early Emotional Semantic Image Retrieval (i.e., ESIR) system known as *K-DIME*. In *K-DIME*, individual models for different users are built using neural network. In [8], Itten's color contrast theory [16] is applied for feature extraction. Wang et al. [23] also developed emotion semantic based features for affective image retrieval, while other works, such as [14] and [24], used generic image processing features (e.g., color histograms) for image emotion classification. In [25], Yanulevskaya et al. applied Gabor and Wiccest features, which are combined with machine learning techniques, to perform emotional valence categorization. Cho [6] developed a human–computer interface for interactive architecture, art, and music design. The studies [13] and [22] focused on affective content analysis for movies clips. More recently, some affective image data sets [17] were proposed for affective image classification. Inspired by the empirical concepts from psychology and art theory, low-level image features, such as color, texture, and high-level features (composition and content), are extracted and combined to represent the emotional content of an image for classification tasks. The authors also constructed an image dataset which contains a set of artistic photographs from a photo sharing site and a set of peer rated abstract paintings.

Beyond these emotion analysis efforts, one question naturally rises: could we endow a photo (image) with user specified emotions? An effective solution to this problem will lead to many potential interesting multimedia applications such as instant online messengers and photo editing softwares. This new function, illustrated in Fig. 1, can help the inexperienced users create professional emotion-specific photos, even though they have little knowledge about photographic techniques. Nevertheless, this problem has rarely been studied. Not surprising, image emotion synthesis is a difficult problem, given that: (1) the mechanism of how image affect the human being's feeling evolves complex biological and psychological processes and the modern biology and psychology studies have very limited knowledge on it. Thus, mathematical modeling of this mechanism is intractable; and (2) human being's affection process is highly subjective, i.e., the same image could affect different people into different emotions. Although to develop an expert system like computational model is intractable, we believe that these problems could be alleviated by a learning-based approach. It is fortunate that we can obtain a large number of emotion-annotated images from photo sharing websites such as *Flickr.com*. From a statistical point of view, images within each emotional group must convey some information and common structures which determine its affective property. Therefore, if the underlying

Fig. 1 Objective of the proposed work: emotion synthesis. Given an input image, our system can synthesize any user specific emotion on it automatically

cues that constitute an emotion-specific image can be mined by a learning framework, they can be further utilized for automatic image emotion synthesis.

Our proposed solution is motivated by the recent advances in utilizing web data (images, videos, meta-data) for multimedia applications [18, 5]. First, an emotion-specific image dataset is constructed by collecting Internet images and annotating them with emotion tags by Amazon's Mechanic Turk [1]. Training images within each emotion group are clustered into different scene subgroups according to their color and texture features. Then these images are decomposed into over-segmented patch (super-pixel) representations and for each emotion + scene group, a generative model (e.g., Gaussian Mixture Models) based on the color distribution of the image segments is constructed. To synthesize some specific emotion onto an input image, a piece-wise linear transformation is defined for aligning the feature distribution of the target image with the statistical model constructed from the source emotion + scene image subgroup. Finally, a framework is developed by further incorporation of a regularization term enforcing the spatial smoothness and edge preservation for the derived transformation, and the objective function is solved by gradient descent method. Extensive user studies are performed to evaluate the validity and performance of the proposed system.

2 Related Works

Several works have been done for image color transformation [4, 12, 19, 21]. In [19], Reinhard et al. presented a system that transfers color by example via aligning the mean and standard deviation of the color channels in both input and reference images. However, user input is required to perform the preferred color transformation. Other works focused on non-photorealistic rendering (i.e., image stylization) which communicates the main context of an image and explores the rendering effect of the scene with the artistic styles, such as painting [11, 26], cartoon [15] etc. Typically, the target exemplar style image is selected manually [4].

Our work is distinctive with these works: first, most of the previous works focused on only color transformation without any semantic knowledge transfer, however, our work directly synthesizes affective property onto arbitrary images,

which is hardly investigated throughout literature; second, our proposed system is fully automatic which requires no human interactions, however, most of the previous methods require either users' manual selection of certain painting parameters [11] or users' specification of specific example images [4].

3 Learning to Emotionalize Images

In this section, we first discuss our emotion-specific image dataset construction; then we introduce the statistical modeling of the image emotion related features and propose an emotion transfer model for synthesizing any user specified emotion onto the input images.

3.1 Dataset Construction

A training image dataset that contains emotion specific images is constructed. In this work, we mainly consider landscape images (for other categories of images, the same method applies). In [27], the International Affective Picture System (IAPS) was developed and used as emotional stimuli for emotion and attention investigations. Note that we do not use the dataset provided in [17] since most of the images in [17] are artistic photographs or abstract paints, which are not appropriate for training emotion-specific models for real images such as landscape photos. A subset from the NUS-WIDE [7] image dataset, which is collected from web images, is selected as our training dataset. To obtain emotion annotations, we adopt the interactive annotation scheme by Amazon *Mechanical Turk*. The web users are asked to annotate the images into *eight* emotions, including *Awe, Anger, Amusement, Contentment, Sad, Fear, Disgust, Excitement* by Mechanical Turk. We only accept the annotations which are at least agreed by *three* (out of *five*) users, resulting about 5, 000 emotion specific images. As mentioned, we only choose landscape photos, e.g., beach, autumn, mountain, etc. as our training set. Exemplar images are shown in Fig. 2 and the statistics of the resulting image dataset are shown in Table 1. From Table 1, we can observe that only a few landscape images are labeled as disgust or fear, thus we only consider *four* types of emotions, including two positive emotions (i.e., *Awe* and *Contentment*) and two negative emotions (e.g., *Fear* and *Sad*).

3.2 Emotion-Specific Image Grouping

One can observe that even within each emotion group, image appearances may have large variations. Therefore, to develop a single model for each emotion image class is not reasonable. To cope with this problem, we first divide each

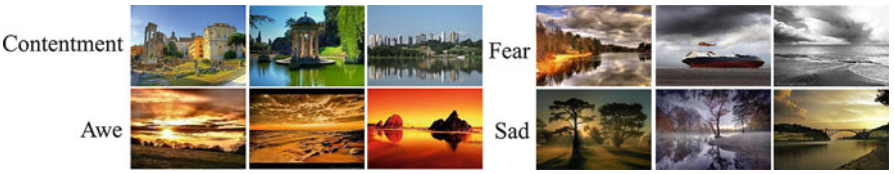

Fig. 2 Exemplar emotion-specific images of our dataset. The exemplar images are from *Contentment, Awe, Fear and Sad*, respectively

Table 1 Statistics of our constructed emotion annotated image dataset

	Amuse.	Anger	Awe	Content.	Disgust	Excite.	Fear	Sad	Sum
NUS-WIDE-Subset	115	199	**1,819**	**1,643**	24	201	**238**	**627**	4,866

Values in bold face show the size of chosen emotion sets.

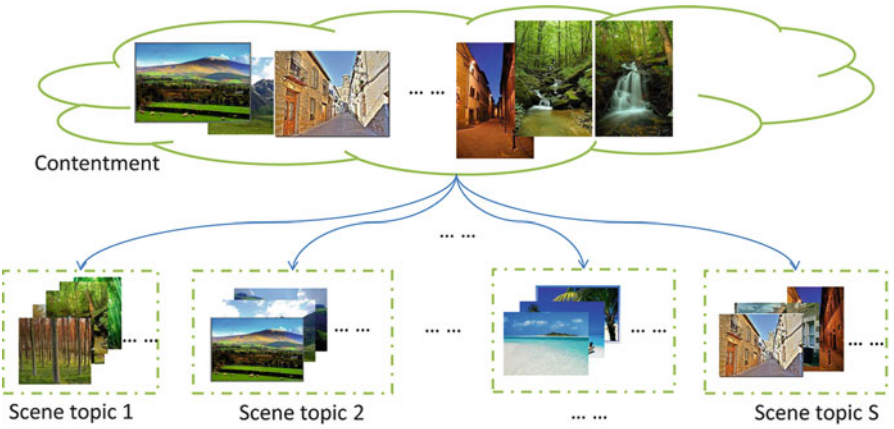

Fig. 3 Example results of the image grouping process. The image set annotated with the emotion *contentment* is grouped into several scene subgroups

emotion-specific image set into several subsets such that the images within the same subgroup share similar appearances and structures. Then computational model is constructed for each of these image sub-groups. Similar with [5], first we decompose each image into a set of over-segmented image patches (i.e., super-pixels) by [10], then color (color moment) [20] and texture features (HOG) [9] are extracted and quantized by the bag-of-words model. Note that color and texture are complementary to each other in measuring image patch characteristics. Finally we cluster the images into several scene subgroups by K-means. An illustration of the image grouping result is given in Fig. 3. One can observe that within each scene subgroup, the images' appearances are quite similar. We can also note that different scene subgroups belong to different landscape types such as *beach, autumn, mountain*, etc.

3.3 Image Emotion Modeling

Emotion specific information is implicit within each emotion + scene subgroup. To uncover this information for our emotion synthesis task, we use generative models, i.e., Gaussian mixture models (GMM), for modeling the statistics of the image patch (segment) appearances within each emotion + scene image subgroup. We denote $\mathbf{x}$ as the appearance feature (i.e., a $3D$ vector of R, G, B values) of an image patch segmented by [10]. Then each image is regarded as a bag of image segments. The reason for using this simple image features (i.e., RGB color space) is that it is simple and direct for color transformation, which has been extensively demonstrated by previous works such as [19, 21]. We further denote that there are C emotion + scene image subgroups.

For each image subgroup $c \in \{1, 2, \ldots C\}$, we utilize GMM to describe the patch feature distribution, which is given as follows:

$$p(\mathbf{x}|\Theta^c) = \sum_{k=1}^{k} \omega_k \mathcal{N}(\mathbf{x}|\mu_k^c, \Sigma_k^c), \tag{1}$$

where $\Theta^c = \{\mu_1^c, \Sigma_k^c, \omega_1^c, \ldots, \mu_k^c, \Sigma_k^c, \omega_k^c\}$. K denotes the number of Gaussian components. μ_k^c, Σ_k^c and ω_k^c are mean, covariance matrix and weight of the kth Gaussian component, respectively. For notational simplicity, we drop the superscript c for the rest of this subsection, while all the equations are presented for each emotion + scene subgroup. $\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$ denotes the uni-modal Gaussian density, namely,

$$\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x} - \mu_k)^T \Sigma_k^{-1}(\mathbf{x} - \mu_k)\}. \tag{2}$$

The parameters of GMM can be obtained by applying Expectation-Maximization (EM) approach.

After EM, we can obtain the estimated GMM parameters $\{\Theta^1, \Theta^2, \cdots, \Theta^C\}$, where each Θ^c characterizes the feature distribution of a specific emotion + scene subgroup.

3.4 Learning-Based Emotion Synthesis

We first classify the input image into the image subgroup within the target emotion group. This can be achieved by first over-segmenting the input image and forming bag-of-words representation based on the color and texture features; then the nearest neighbor image in the target emotion group is found by computing the Euclidean distance of the histogram representations between the input image and the training images, and the scene label of the nearest database image is selected to be the scene label of the input image, denoted as c.

As studied in [21, 19], color (contrast, saturation, hue, etc.) can convey emotion related information. We therefore perform emotion synthesis via applying linear mapping on the RGB color space for the target image. Instead of performing global mapping for the entire image as in [21], we propose the following piece-wise linear mapping for each segment (super-pixel or patch) of the target image as,

$$f_i(\mathbf{x}) = P_i\mathbf{x} + \Delta\mathbf{x}. \tag{3}$$

Equivalently, we can augment P, $\mathbf{x}$ with an additional constant values, i.e., $\widetilde{\mathbf{x}} = [\mathbf{x}^T, 1]^T$, $\widetilde{P} = [P, \Delta\mathbf{x}]$ as,

$$f_i(\mathbf{x}) = \widetilde{P}_i\widetilde{\mathbf{x}}. \tag{4}$$

For notational simplicity, we use P, $\mathbf{x}$ to represent $\widetilde{P}, \widetilde{\mathbf{x}}$ for the rest of this subsection. Here, $\mathbf{x}$ denotes the appearance feature of one super-pixel (image segment). P_i denotes the mapping function for operating the i-th image segment (super-pixel). These image patches are super-pixels which are obtained by using [10]. Note that every pixel within the same super-pixel (image segment) shares the same mapping function f_i. The goal of our synthesis process is to obtain the set of appropriate linear mapping functions for the entire target image (suppose we have M image segments), namely, $\mathcal{P} = \{P_1, ..., P_M\}$. The objective of emotion synthesis can be expressed as,

$$\max_{\mathcal{P}} (\mathcal{F}_1 + \mathcal{F}_2), \tag{5}$$

The objective formulation contains two parts. The first part is a regularization term, which enforces the smoothness of the transformation and also maintains the edges of the original image. $\mathcal{F}_1$ can be expressed as:

$$\begin{aligned}\mathcal{F}_1 = &- \sum_{i,j\in N(a)} \omega_{ij}^a \|P_i\mathbf{x}_i - P_j\mathbf{x}_j\|_2^2 + \lambda \sum_{i,j\in N(s)} \omega_{ij}^s \|P_i\mathbf{x}_i - P_j\mathbf{x}_j\|_2^2 \\ &- \sum_{i,j\in N(c)} \|P_i - P_j\|_F^2,\end{aligned} \tag{6}$$

where

$$\begin{aligned}N(a) &= \{i,j|i,j \in N(c), \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \leq \theta_1\}, N(s) \\ &= \{i,j|i,j \in N(c), \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \geq \theta_2\}.\end{aligned} \tag{7}$$

Here, $N(c)$ denotes the spatial neighborhood, i.e., two super-pixels i and j are adjacent. θ_1 and θ_2 are the color difference thresholds. ω_{ij}^a and ω_{ij}^s are the weighting coefficients, which are defined as follows:

$$\omega_{ij}^a \propto \exp(-\|\mathbf{x_i} - \mathbf{x_j}\|_2^2/a), \omega_{ij}^s \propto 1 - \exp(-\|\mathbf{x_i} - \mathbf{x_j}\|_2^2/a). \tag{8}$$

Here, θ_1, θ_2, λ and a are set to be optimal empirically. We can note from this prior that: (1) The first term ensures that original contours in the target image will be preserved by enforcing that originally distinctive neighborhood segments present distinctive color values in the transformed image; (2) The second term encourages smooth transition from image segments to near-by segments.

The second part of the framework is the emotion fitting term, which is expressed as:

$$\mathcal{F}_2 = \log\left(\prod_{i=1}^{M} p(\mathcal{I}|P_i)\right) = \log\left(\prod_{i=1}^{M} p(\mathbf{x}_i|\Theta^c)\right). \tag{9}$$

Here $p(\mathbf{x} \mid \Theta^c)$ is the trained GMM model for emotion+scene subgroup c, $\mathbf{x}_i$ is the color vector of the ith image segment. We can note that this term encourages the distributions of the target image to move towards the statistical model of the training data. Finally the cost function is denoted as:

$$\begin{aligned}\mathcal{F} &= \mathcal{F}_1 + \mathcal{F}_2 \\ &= -\sum_{i,j\in N(a)} \omega_{ij}^a \|P_i\mathbf{x}_i - P_j\mathbf{x}_j\|_2^2 + \lambda \sum_{i,j\in N(s)} \omega_{ij}^s \|P_i\mathbf{x}_i - P_j\mathbf{x}_j\|_2^2 \\ &\quad - \sum_{i,j\in N(c)} \|P_i - P_j\|_F^2 + \sum_i \log\left(\sum_k \mathcal{N}_k\right).\end{aligned} \tag{10}$$

Note that Eq. (10) is nonlinear and complex. Therefore, to optimize the cost function, we adopt Newton's method with linear constraints, which can guarantee local optimum [3], as:

$$\max_{\mathcal{P}} \mathcal{F}, s.t. 0 \preceq P_i\mathbf{x} \preceq 255, \forall i = 1, 2, \cdots, M, \tag{11}$$

where $\preceq$ denotes component-wise inequality constraints. These constraints ensure that the resulting color value is within appropriate range. Our method is schematically illustrated in Fig. 4.

4 Experiments

In this section, we will introduce our experimental settings, user studies along with discussions. As mentioned in the previous sections, during the pre-processing stage, training images within each emotion group are segregated into several scene subgroups (subclasses) based on the distributions of the image super-pixels' HOG and color moment features. For each subclass, we train a GMM to describe the

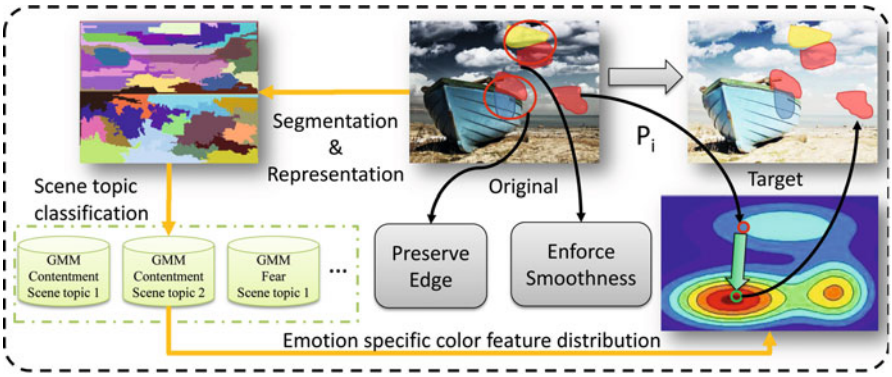

Fig. 4 The learning-based emotion synthesis scheme

distribution of super-pixels' color information. Given an arbitrary input image, in the emotion synthesis phase, we first assign it to the nearest scene subclass based on the HOG and color moment bag-of-words representation. Then our task is to obtain a mapping function which can optimize Eq. (11). Since our probability function is non-convex, we can easily get trapped in a local optimum. Therefore, a good initial mapping matrix is crucial. To get a proper initialization, we firstly assign patch (super-pixel) j to the nearest Gaussian component center μ_i. After that, a pseudo inverse is performed as $P^i_{inv} = x_j^{\dagger} u_i$, here $\mathbf{x}_j$ denotes mean color feature value of image patch j. The linear multiplier transformation part of the initial mapping matrix becomes, $P^i_{ini} = \lambda I + (1 - \lambda) P^i_{inv}$. Here I is the identity matrix. In our experiment, we set $\lambda = 0.8$ empirically. With a good initialization, we can mostly obtain a good mapping matrix using standard non-linear optimization algorithms such as Newton's method.

In our experiment, we choose 55 images from the NUS-WIDE dataset which serve as the testing images while the others construct the training image set. We compare our proposed method with the color transfer method proposed in [19], which directly aligns the mean and standard deviation of the color distribution between the source (reference) and the target image. The target image is mapped with the reference image chosen from the emotion + scene subclass by nearest neighbor assignment (in terms of the HOG and color moment based bag-of-words representation).

The comparative user studies are conducted as follows. Firstly, the transformed images of both methods are presented to the participants in pairs (with the left-right order randomly shuffled). Participants are asked to decide whether these image can express the specified target emotion. We also consider the naturalness of the synthesized images, since the naturalness will significantly affect the image quality. In this sense, the participants are also asked to compare which image of the same pair is more natural. In particular, participants are asked to give a judgement that whether the left image is Much Better, Better, Same, Worse, Much Worse than right one.

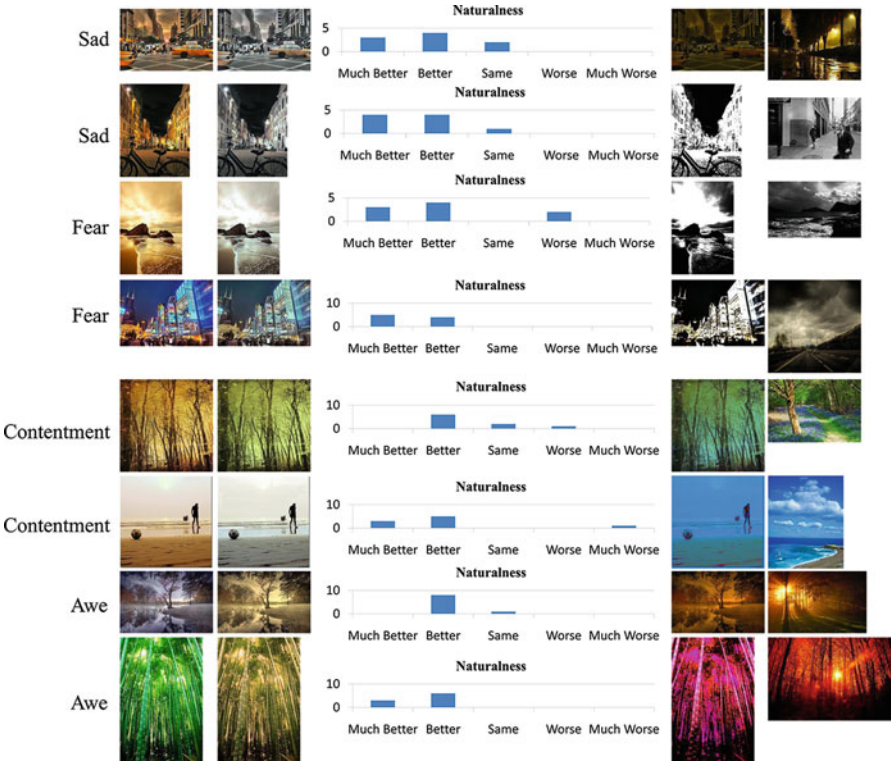

Fig. 5 Example results of the image emotion synthesis. Each row, from *left* to *right*, show the original image, synthesized image using our method, naturalness evaluation bar, color transfer result and reference image in color transfer. The *middle bars* show statistics of user's responses which indicate based on naturalness whether synthesizing result (*left*) is Much Better, Better, Same, Worse, Much Worse than the result from color transfer method (*right*). For better viewing, please see in x2 resolution and in color pdf file

In our user study, 9 participants are asked to judge the image's naturalness, and 20 participants with ages ranging from 20 to 35 years old are asked to judge whether these images can express the target emotion. The statistics of the results for the user study are illustrated in Fig. 6 in terms of the naturalness. We also show several example results in Fig. 5 for both our method and the color transfer method.

In Fig. 6, yellow bars show the number of participants voting for each type of the ratings. We can observe that the images resulting from our method are more natural to the audience than the ones from the color transfer method. This could be explained by the fact that the statistic modeling using GMM is more generative and robust, while the exemplar image based color transfer might sometimes lead to over-fitting. Figure 6 and Table 2 show that in most cases images which are synthesized using our method outperform the color transfer results in terms of the accuracies of emotion synthesis. Figure 5 further shows that our results are more

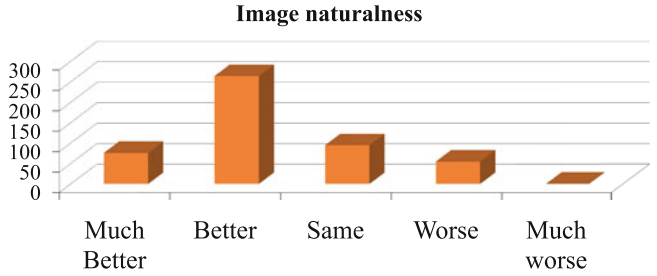

Fig. 6 The statistics from our user studies. The *bar* shows the summation of user's feedback based on naturalness, i.e. whether the result of our method is Much Better, Better, Same, Worse, Much Worse than the result of color transfer

Table 2 Perceptibility comparison of each emotion set

	Awe	Contentment	Fear	Sad	Average
Baseline	0.4375	0.3600	0.4833	0.3788	0.4082
Our method	0.6250	0.6550	0.6100	0.7904	0.7045

natural than color transfer based results. As can be seen, color transfer based results rely on reference images. Therefore, if the color distribution of reference image is too far from the target image, the transformed result will be unnatural, e.g., trees in the last example look red which are not realistic. However, our statistical learning based method do not have such a problem.

5 Conclusions

In this work, we developed a learning based image emotion synthesis framework which can transfer the learnt emotion related statistical information onto arbitrary input images. Extensive user studies well demonstrated that the proposed method is effective and the re-emotionalized images are natural and realistic.

Acknowledgements This research is done for CSIDM Project No. CSIDM- 200803 partially funded by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore. This work is partially supported by Human Sixth Sense Project, Illinois@Singapore Pte Ltd.

References

1. https://www.mturk.com/mturk/welcome
2. Bianchi-Berthouze, N.: K-dime: an affective image filtering system. TMM 10, 103–106 (2003)
3. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)

4. Chang, Y., Saito, S., Nakajima, M.: Example-based color transformation of image and video using basic color categories. TIP 16, 329–336 (2007)
5. Cheng, B., Ni, B., Yan, S., Tian, Q.: Learning to photograph. In: ACM MM. pp. 291–300 (2010)
6. Cho, S.B.: Emotional image and musical information retrieval with interactive genetic algorithm. In: Proceedings of the IEEE. pp. 702–711 (2004)
7. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.T.: Nus-wide: A real-world web image database from national university of singapore. In: CIVR (2009)
8. Colombo, C., Bimbo, A.D., Pala, P.: Semantics in visual information retrieval. TMM 6, 38–53 (1999)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. pp. 886–893 (2005)
10. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. IJCV 59, 167–181 (2004)
11. Guo, Y., Yu, J., Xu, X., Wang, J., Peng, Q.: Example based painting generation. CGI 7(7), 1152–1159 (2006)
12. Gupta, M.R., Upton, S., Bowen, J.: Simulating the effect of illumination using color transformation. SPIE CCI 111, 248–258 (2005)
13. Hanjalic, A.: Extracting moods from pictures and sounds: towards truly personalized tv. IEEE Signal Processing Magazine 23, 90–100 (2006)
14. Hayashi, T., Hagiwara, M.: Image query by impression words-the iqi system. TCE 44, 347–352 (1998)
15. Hong, R., Yuan, X., Xu, M., Wang, M., Yan, S., Chua, T.S.: Movie2comics: A feast of multimedia artwork. In: ACM MM. pp. 611–614 (2010)
16. Itten, J.: The art of color: the subjective experience and objective rationale of color. John Wiley, New York (1973)
17. Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: ACM MM. pp. 83–92 (2010)
18. Ni, B., Song, Z., Yan, S.: Web image mining towards universal age estimator. In: ACM MM. pp. 85–94 (2009)
19. Reinhard, E., Ashikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. CGA 21 (2001)
20. Stricker, M., Orengo, M.: Similarity of color images. In: SPIE. pp. 381–392 (1995)
21. Thompson, W.B., Shirley, P., Ferwerda, J.A.: A spatial post-processing algorithm for images of night scenes. Journal of Graphics Tools 7, 1–12 (2002)
22. Wang, H.L., Cheong, L.F.: Affective understanding in film. TCSVT 16, 689–704 (2006)
23. Wang, W.N., Yu, Y.L., Jiang, S.M.: Image retrieval by emotional semantics: A study of emotional space and feature extraction. In: IEEE International Conference on Systems, Man and Cybernetics. pp. 3534–3539 (2006)
24. Wu, Q., Zhou, C., Wang, C.: Content-based affective image classification and retrieval using support vector machines. Affective Computing and Intelligent Interaction (2005)
25. Yanulevskaya, V., van Gemert, J.C., Roth, K., Herbold, A.K., Sebe, N., Geusebroek, J.M.: Emotional valence categorization using holistic image features. In: ICIP (2008)
26. Zhang, X., Constable, M., He, Y.: On the transfer of painting style to photographic images through attention to colour contrast. In: Pacific-Rim Symposium on Image and Video Technology. pp. 414–421 (2010)
27. Lang, P.J., Bradley, M.M., Cuthbert, B.N.: International affective picture system (IAPS): Affective ratings of pictures and instruction manual. In: Technical Report A-8. University of Florida, Gainesville, FL.(2008)