A Mixed Bag of Emotions: Model, Predict, and Transfer Emotion Distributions

Kuan-Chuan Peng, Tsuhan Chen
Cornell University
{kp388, tsuhan}@cornell.edu

Amir Sadovnik
Lafayette College
sadovnia@lafayette.edu

Andrew Gallagher
Google Inc.
agallagher@google.com

Abstract

This paper explores two new aspects of photos and human emotions. First, we show through psychovisual studies that different people have different emotional reactions to the same image, which is a strong and novel departure from previous work that only records and predicts a single dominant emotion for each image. Our studies also show that the same person may have multiple emotional reactions to one image. Predicting emotions in "distributions" instead of a single dominant emotion is important for many applications. Second, we show not only that we can often change the evoked emotion of an image by adjusting color tone and texture related features but also that we can choose in which "emotional direction" this change occurs by selecting a target image. In addition, we present a new database, Emotion6, containing distributions of emotions.

1. Introduction

Images are emotionally powerful. An image can evoke a strong emotion in the viewer. In fact, photographers often construct images to elicit a specific response by the viewer. By using different filters and photographing techniques, photographs of the same scene may elicit very different emotions. Motivated by this fact, we aim to mimic this process after the image was taken. That is, we wish to change an image's original evoked emotion to a new one by changing its low-level properties.

Further, the viewer's emotion may be sometimes affected in a way that was unexpected by the photographer. For example, an image of a hot air balloon may evoke feelings of joy to some observers (who crave adventure), but fear in others (who have fear of heights). We address the fact that people have different evoked emotions by collecting and predicting the distributions of emotional responses when an image is viewed by a large population. We also address the fact that the same person may have multiple emotions evoked by one image by allowing the subjects to record multiple emotional responses to one image.

This paper proposes a framework for transferring the distribution of evoked emotion of an input image without severely interfering with the high-level semantic content of the original image. Since we work with emotion distributions, we propose a novel approach for emotion transfer that includes choosing an image representing the target emotion distribution. Using a target image for emotion transfer is intuitive and allows an operator to change multiple emotion dimensions simultaneously. Figure 1 shows an example of transferring evoked emotion distribution, where the transformed image evokes (versus the source) emotions more similar to those of the target image. Further, we build a model to predict the evoked emotion that a population of observers has when viewing a particular image.

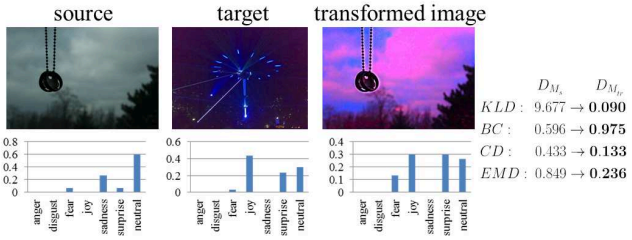

Figure 1. An example of transferring evoked emotion distribution. We transform the color tone and texture related features of the source to those of the target. The ground truth probability distribution of the evoked emotion is shown under each image, supporting that the color modification makes the source image more joyful. A quantitative evaluation measuring the similarity of two probability distributions with four metrics M ($M \in \{KLD, BC, CD, EMD\}$) (see Sec. 5) is shown on the right, where D_{M_s} is the distance between source and target distributions, and $D_{M_{tr}}$ is the distance between transformed and target distributions. For each metric, the better number is displayed in bold. By any of the 4 measures, the transformed image evokes more similar emotions to the target image versus the source image.

We make the following contributions: 1) We show that different people have different emotional reactions to an image and that the same person may have multiple emotional reactions to an image. Our proposed database, Emotion6, addresses both findings by modeling emotion distributions. 2) We use a convolutional neural network (CNN) to predict emotion distributions, rather than simply

predicting a single dominant emotion, evoked by an image. Our predictor of emotion distributions for Emotion6 can serve as a better baseline than using support vector regression (SVR) with the features from previous works [20, 28, 30] for future researchers. We also predict emotions in the traditional setting of affective image classification, showing that CNN outperforms Wang's method [30] on Artphoto dataset [20]. 3) This paper introduces the application of transferring the evoked emotion distribution from one image to another. With the support of a large-scale (600 images) user study, we successfully adjust the evoked emotion distribution of an image toward that of a target image without changing the high-level semantics.

2. Prior Work

In computer vision, image classification based on abstract concepts has recently received a great deal of focus. Aesthetic quality estimation [14] and affective image classification [20, 28, 30] are two typical examples. However, these two abstract concepts are fundamentally different because the evoked emotion of an image is not equivalent to aesthetic quality. For example, one may feel joyful after viewing either amateur or expert photos, and aesthetically ideal images may elicit either happy or sad emotions. Moreover, aesthetic quality is a one-dimensional attribute, whereas emotions are not [10].

In most previous works on affective image classification [3, 15, 20, 28, 30], image features are used solely for classification. Relatively few works address manipulating an image's evoked emotion by editing the image. Wang et al. [29] associate color themes with emotion keywords using art theory and transform the color theme of the input image to the desired one. However, most results in their work deal with cartoon-like images, and they only use few examples. Also showing few examples, Peng et al. [24] propose a framework for changing the image emotion by randomly sampling from a set of possible target images and using an emotion predictor. In contrast with Peng's method [24], our proposed method of emotion transfer does not need random sampling or a predefined set of target images, and performance on a large set (600 images) is deeply analyzed. Jun et al. [16] show that pleasure and excitement are affected by changes in brightness and contrast. However, changing these two features can only produce limited variation of the input image. Here, we modify the color tone and texture related features of an image to transfer the evoked emotion distribution. We also show that the change of the evoked emotion distribution is consistent with the ground truth from our user study.

To modify the evoked emotion of an image, one may predict how the image transformation affects image features. This is not a trivial task. Some image attributes have well-established transformation methods, such as color histogram reshaping [12, 25], texture transfer [8], edge histogram specification [22], etc. However, most of these works belong to the domain of image processing, and have not considered the affective interpretation of the image edits. This is an important part of our goal. We use these image editing tools and focus on building connections between the processed images and the evoked emotions, and we show that it is often possible to change the evoked emotion distribution of an image towards that of a target image.

In predicting the emotion responses evoked by an image, researchers conduct experiments on various types of images. Wang et al. [30] perform affective image classification on artistic photos or abstract paintings, but Solli and Lenz [28] use Internet images. Machajdik and Hanbury [20] use both artistic and realistic images in their experiment. The fact that groups of people seldom agree on the evoked emotion [14] and that even a person may have multiple emotions evoked by one image, are ignored by previous works. According to the statistics of Emotion6, the image database we collect in Sec. 3, more than half of the subjects have emotion responses different from the dominant emotion. Our statistics also show that $\sim$22% of all the subjects' responses select $\geq$2 emotion keywords to describe one subject's evoked emotions. Both of these observations support our assertion that emotion should be represented as a distribution rather than as a single dominant emotion. Further, predicting emotions by distribution rather than as a single dominant emotion is important for practical applications. For example, a company has two possible ads arousing different emotion distributions (`ad1`: 60% joy and 40% surprise; `ad2`: 70% joy and 30% fear). Though `ad2` elicits joy with higher probability than `ad1` does, the company may choose `ad1` instead of `ad2` because `ad2` arouses negative emotion in some part of the population.

In psychology, researchers have been interested in emotions for decades, leading to three major approaches – "basic emotion", "appraisal", and "psychological constructionist" traditions [11]. With debate on these approaches, psychologists designed different kinds of models for explaining fundamental emotions based on various criteria. Ortony and Turner [23] summarized some past theories of basic emotions, and even now, there is not complete consensus among psychologists. One of the most popular frameworks in the emotion field proposed by Russell [26], the valence–arousal (VA) model, characterizes emotions in VA dimensions, where valence describes the emotion in the scale of positive to negative emotion, while arousal indicates the degree of stimulation or excitement. We adopt VA model as part of emotion prediction. In terms of emotion categories, we adopt Ekman's six basic emotions [9], which details are explained in Sec. 3.

Issues of previous databases	Explanation and how Emotion6 solves the issues
Ad-hoc categories	Previous databases select emotion categories without psychological background, but Emotion6 uses Ekman's 6 basic emotions [9] as categories.
Unbalanced categories	Previous databases have unbalanced proportion of images from each category, but Emotion6 has balanced categories with 330 images per category.
Single category per image	Assigning each image to only one category (dominant emotion), previous databases ignore that the evoked emotion can vary between observers [14]. Emotion6 expresses the emotion associated with each image in probability distribution.

Table 1. The issues of previous emotion image databases and how our proposed database, Emotion6, solves these issues.

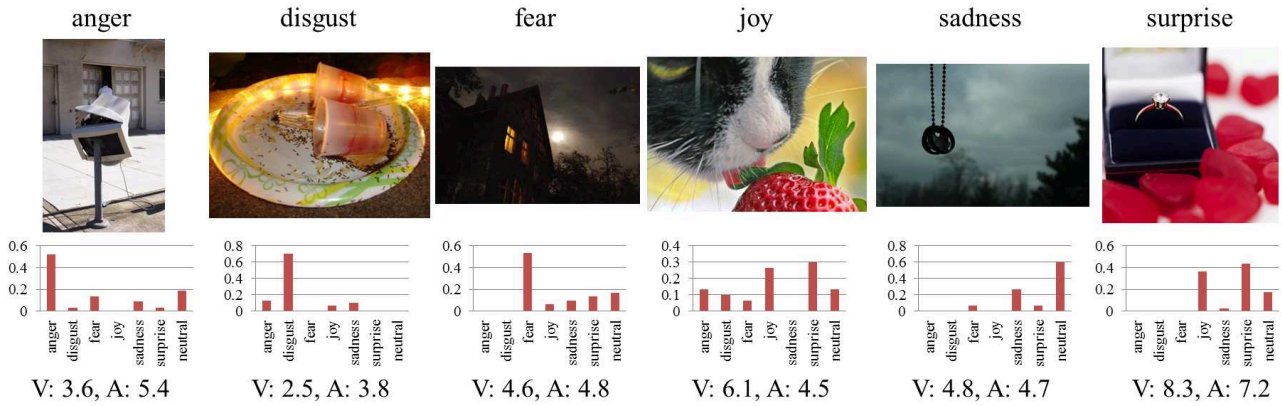

Figure 2. Example images of Emotion6 with the corresponding ground truth. The emotion keyword used to search each image is displayed on the top. The graph below each image shows the probability distribution of evoked emotions of that image. The bottom two numbers are valence–arousal (VA) scores in SAM 9-point scale [1].

To recognize and classify different emotions, scientists build connections between emotions and various types of input data including text [10], speech [7, 17], facial expressions [7, 6], music [31], and gestures [7]. Among the research related to emotions, we are interested in emotions evoked by consumer photographs (not just artworks or abstract images as in [30]). Unfortunately, the number of related databases is relatively few compared to other areas mentioned previously. These databases, such as IAPS [19], GAPED [4], and emodb [28] have a few clear limitations. We propose a new emotion database, Emotion6, which the paper is mainly based on. Table 1 summarizes how Emotion6 solves the limitations of previous databases. Sec. 3 describes the details of Emotion6.

3. The Emotion6 Database

For each image in Emotion6, the following information is collected by a user study: 1) The ground truth of VA scores for evoked emotion. 2) The ground truth of emotion distribution for evoked emotion. Consisting of 7 bins, Ekman's 6 basic emotions [9] and neutral, each emotion distribution represents the probability that an image will be classified into each bin by a subject. For both VA scores, we adopt the Self-Assessment Manikin (SAM) 9-point scale [1], which is also used in [19]. For the valence scores, 1, 5, and 9 mean very negative, neutral, and very positive emotions respectively. For the arousal scores, 1 (9) means the emotion has low (high) stimulating effect. Figure 2 shows some images from Emotion6 with the corresponding ground truth. The details about the selection of emotion model/categories/images and the user study are described in the following paragraphs. More statistics are shown in the supplementary material. We will release the database upon publication.

Emotion model and category selection: According to the list of different theories of basic emotions [23], we use Ekman's six basic emotions [9] (anger, disgust, joy, fear, sadness, and surprise) as the categories of Emotion6. Each of these six emotions is adopted by at least three psychological theorists in [23], which provides a consensus for the importance of each of these six emotions. We adopt the valence–arousal (VA) model, in addition to using emotion keywords as categories because we want to be consistent with the previous databases where ground truth VA scores are provided.

Image collection and user study: We collect the images of Emotion6 from Flickr by using the 6 category keywords and synonyms as search terms. High-level semantic content of an image, including strong facial expressions, posed humans, and text, influences the evoked emotion of an image. However, one of our goals is to modify the image at a low level (rather than modifying text or facial expressions) to manipulate the evoked emotion. One could argue that Emotion6 should not contain images with high-level semantic concepts. However, this is not trivial because the definition of high-level semantic contents is debatable.

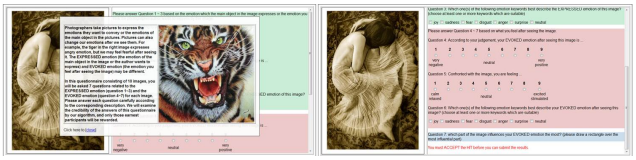

Figure 3. Two screenshots of the interface of our user study on AMT. Before the subject answers the questions (right image), we provide instructions and an example (left image) explaining how to answer the questions to the subject.

Therefore, we only remove the images containing apparent human facial expressions or text directly related to the evoked emotion because these two contents are shown to have strong relationship to the emotion [7, 10]. In contrast to the database emodb [28], that has no human moderation, we examine each image in Emotion6 to remove erroneous images. A total of 1980 images are collected, 330 for each category, comparable to previous databases. Each image is scaled to approximately VGA resolution while keeping the original aspect ratio.

We use Amazon Mechanical Turk (AMT) to collect emotional responses from subjects. For each image, each subject rates the evoked emotion in terms of VA scores, and chooses the keyword(s) best describing the evoked emotion. We provide 7 emotion keywords (Ekman's 6 basic emotions [9] and neutral), and the subject can select multiple keywords for each image. Instead of directly asking the subject to give VA scores, we rephrase the questions to be similar to GAPED [4]. Figure 3 shows two snapshots of the interface. To compare with previous databases, we randomly extract a subset S containing 220 images from GAPED [4] such that the proportion of each category in S is the same as that of GAPED. We rejected the responses from a few subjects who failed to demonstrate consistency or provided a constant score for all images.

Each HIT on AMT contains 10 images, and we offer 10 cents to reward the subject's completion of each HIT. In the instructions, we inform the subject that the answers will be examined by an algorithm that detects lazy or fraudulent workers and only workers that pass will be paid. In each HIT, the last image is from S, and the other 9 images are from Emotion6. We create 220 different HITs for AMT such that the following constraints are satisfied: 1) Each HIT contains at least one image from each of 6 categories (by keyword). 2) Images are ordered in such a way that the frequency of an image from category i appearing after category j is equal for all i, j. 3) Each image or HIT cannot be rated more than once by the same subject, and each subject cannot rate more than 55 different HITs. 4) Each image is scored by 15 subjects.

Mean and standard deviation, in seconds, on each HIT are 450 and 390 respectively. The minimum time spent on 1 HIT is 127 seconds, which is still reasonable. 432 unique subjects took part in the experiment, rating 76.4 images on average. After collecting the answers from the subjects, we sort the VA scores, and average the middle 9 scores (to remove outliers) to serve as ground truth. For emotion category distribution, the ground truth of each category is the average vote of that category across subjects. To provide grounding for Emotion6, we compute the VA scores of the images from S using the above method and compare them with the ground truth provided by GAPED [4], where the original scale 0∼100 is converted linearly to 1∼9 to be consistent with our scale. The average of absolute difference of V (A) scores for these images is 1.006 (1.362) in SAM 9-point scale [1], which is comparable in this highly subjective domain.

4. Predicting Emotion Distributions

Randomly splitting Emotion6 into training and testing sets with the proportion of 7:3, we propose three methods– SVR, CNN, CNNR and compare their performance with those of the three baselines. The details of the proposed three methods are explained below.

SVR: Inspired by previous works on affective image classification [20, 28, 30], we adopt features related to color, edge, texture, saliency, and shape to create a normalized 759-dimensional feature set shown in Table 2. To verify the affective classification ability of this feature set, we perform the exact experiment from [20], using their database. The average true positive per class is ∼60% for each category, comparable to the results presented in [20].

We train one model for each emotion category using the ground truth of the category in Emotion6 with Support Vector Regression (SVR) provided in LIBSVM [2] with the parameters of SVR learned by performing 5-fold cross validation on the training set. In the predicting phase, the probabilities of all emotion categories are normalized such that they sum up to 1. To assess the performance of SVR in emotion classification, we compare the emotion with the greatest prediction with the dominant emotion of the ground truth. The accuracy of our model in this multi-class classification setting is 38.9%, which is about 2.7 times that of random guessing (14.3%).

CNN and CNNR: In CNN, we use the exact convolutional neural network in [18] except that the number of output nodes is changed to 7 to represent the probability of the input image being classified as each emotion category in Emotion6. In CNNR, we train a regressor for each emotion category in Emotion6 with the exact convolutional neural network in [18] except that the number of output nodes is changed to 1 to predict a real value and that the softmax loss layer is replaced with the Euclidean loss layer. In the predicting phase, the probabilities of all emotion categories are normalized to sum to 1. Using the Caffe implementation [13] and its default parameters for training the ImageNet [5] model, we pre-train with the Caffe reference

Feature Type	Dimension	Description
Texture	24	Features from Gray-Level Co-occurrence Matrix (GLCM) including the mean, variance, energy, entropy, contrast, and inverse difference moment [20].
	3	Tamura features (coarseness, contrast and directionality) [20].
Composition	2	Rule of third (distance between salient regions and power points/lines) [30].
	1	Diagonal dominance (distance between prominent lines and two diagonals) [30].
	2	Symmetry (sum of intensity differences between pixels symmetric with respect to the vertical/horizontal central line) [30].
	3	Visual balance (distances of the center of the most salient region from the center of the image, the vertical and horizontal central lines) [30].
Saliency	1	Difference of areas of the most/least saliency regions.
	1	Color difference of the most/least saliency regions.
	2	Difference of the sum of edge magnitude of the most/least saliency regions.
Color	80	Cascaded CIECAM02 color histograms (lightness, chroma, hue, brightness, and saturation) in the most/least saliency regions.
Edge	512	Cascaded edge histograms (8 (8)-bin edge direction (magnitude) in RGB and gray channels) in the most/least saliency regions.
Shape	128	Fit an ellipse for every segment from color segmentation and compute the histogram of fit ellipses in terms of angle (4 bins), the ratio of major and minor axes (4 bins), and area (4 bins) in the most/least saliency regions.

Table 2. The feature set we use for SVR in predicting emotion distributions.

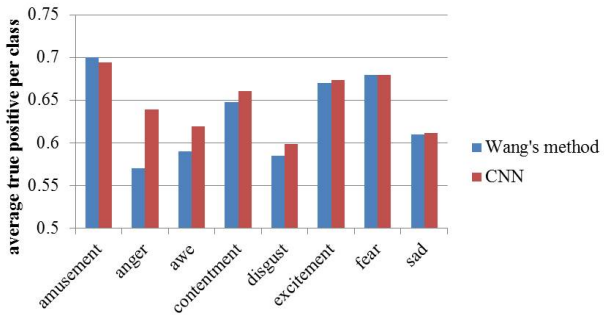

Figure 4. Classification performance of CNN and Wang's method [30] with Artphoto dataset [20]. In 6 out of 8 emotion categories, CNN outperforms Wang's method [30].

model [13] and fine-tune the convolutional neural network with our training set in both CNN and CNNR.

To show the efficacy of classification with the convolutional neural network, we use CNN to perform binary emotion classification with Artphoto dataset [20] under the same experimental setting of Wang's method [30]. In this experiment, we change the number of output nodes to 2 and train one binary classifier for each emotion under 1-vs-all setting. We repeat the positive examples in the training set such that the number of positive examples is the same as that of the negative ones. Figure 4 shows that CNN outperforms Wang's method [30] in 6 out of 8 emotion categories. In terms of the average of average true positive per class of all 8 emotion categories, CNN (64.724%) also outperforms Wang's method [30] (63.163%).

The preceding experiment shows that CNN achieves state-of-art performance for emotion classification of images. However, what we are really interested in is the prediction of emotion distributions, which better capture the range of human responses to an image. For this task, we use CNNR as previously described, and show that its performance is state-of-art for emotion distribution prediction.

We compare the predictions of our proposed three

Method 1	Method 2	P_{KLD}	P_{BC}	P_{CD}	P_{EMD}
CNNR	Uniform	0.742	0.783	0.692	0.756
CNNR	Random	0.815	0.819	0.747	0.802
CNNR	OD	0.997	0.840	0.857	0.759
CNNR	SVR	0.625	0.660	0.571	0.620
CNNR	CNN	0.934	0.810	0.842	0.805
Uniform	OD	0.997	0.667	0.736	0.593

Method	$\overline{KLD}$	$\overline{BC}$	$\overline{CD}$	$\overline{EMD}$
Uniform	0.697	0.762	0.348	0.667
Random	0.978	0.721	0.367	0.727
OD	10.500	0.692	0.510	0.722
SVR	0.577	0.820	0.294	0.560
CNN	2.338	0.692	0.497	0.773
CNNR	**0.480**	**0.847**	**0.265**	**0.503**

Table 3. The performance of different methods for predicting emotion distributions compared using P_M and $\overline{M}$ ($M \in \{KLD, BC, CD, EMD\}$). The upper table shows P_M, the probability that Method 1 outperforms Method 2 with distance metric M. Each row in the upper table shows that Method 1 outperforms Method 2 in all M. The lower table lists $\overline{M}$, the mean of M, of each method, showing that CNNR achieves better $\overline{M}$ than the other methods listed here. CNNR performs the best out of all the listed methods in terms of all P_Ms with better $\overline{M}$.

methods with the following three baselines: 1) A uniform distribution across all emotion categories. 2) A random probability distribution. 3) Optimally dominant (OD) distribution, a winner-take-all strategy where the emotion category with highest probability in ground truth is set to 1, and other emotion categories have zero probability. The first two baselines represent chance guesses while the third represents a best case scenario for *any* (prior art) multiclass emotion classifier that outputs a single emotion.

We use four different distance metrics to evaluate the similarity between two emotion distributions – KL-Divergence (KLD), Bhattacharyya coefficient (BC), Chebyshev distance (CD), and earth mover's distance (EMD) [21, 27]. Since KLD is not well defined when a bin has value 0, we use a small value $\varepsilon = 10^{-10}$ to approximate the values in such bins. In computing EMD in our

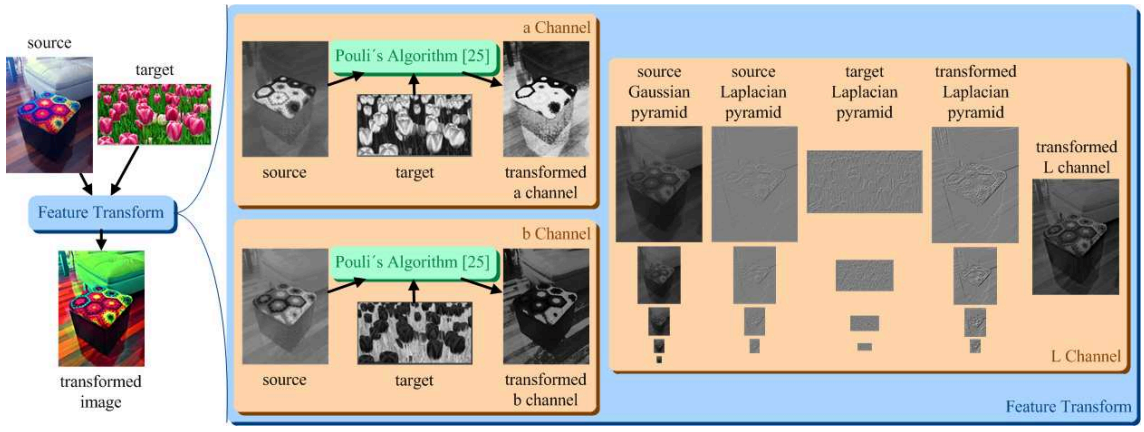

Figure 5. The framework of transferring evoked emotion distribution by changing color and texture related features.

paper, we assume that each of the 7 dimensions (Ekman's 6 basic emotions [9] and neutral) is such that the distance between any two dimensions is the same. For KLD, CD and EMD, lower is better. For BC, higher is better.

For each distance metric M, we use $\overline{M}$ and P_M to evaluate the ranking between two algorithms, where $\overline{M}$ is the mean of M, and P_M in Table 3 (upper table) is the proportion of images where Method 1 matches the ground truth distribution more accurately than Method 2 according to distance metric M. Method 1 is superior to Method 2 when P_M exceeds 0.5. For the random distribution baseline, we repeat 100000 times and report the average P_M. The results are in Table 3. CNNR outperforms SVR, CNN, and the three baselines in both P_M and $\overline{M}$, and should be considered as a standard baseline for future emotion distribution research. Table 3 also shows that OD performs even worse than uniform baseline. This shows that predicting only one single emotion category like [20, 28, 30] does not well model the fact that people have different emotional responses to the same image and that the same person may have multiple emotional responses to one image.

We also use CNNR to predict VA scores. CNNR outperforms the two baselines – guessing VA scores as the mode of all VA scores and guessing VA scores uniformly, and has comparable performance with respect to SVR. The detailed results are in the supplementary material.

5. Transferring Evoked Emotion Distributions

In emotion transfer, the goal is to modify the evoked emotion distribution of the source towards that of the target image. We believe that selecting a target image is more intuitive than specifying the numerical change of each bin of evoked emotion distributions because the quantization of emotion change may be unnatural for humans. Setting up source and target images, we examine the differences between the distributions before and after adjusting the color tone and texture related features. We only change low-level features because we do not want to severely change the high-level semantics of the source image. The framework of transferring evoked emotion distribution is illustrated in Figure 5. For each pair of source and target images, we decompose the images into CIELab color space, and modify the color tone in the ab- channels, and texture related features in the L channel. For the color tone adjustment, we adopt Pouli's algorithm [25] with full transformation. For the adjustment of texture related features, we first create the Laplacian pyramids L_s and L_t for source and target images, respectively. Second, $L_s(i)$ are scaled such that the average of the absolute value of $L_s(i)$ is the same as that of $L_t(i)$, where $L(i)$ is the i-th level of the Laplacian pyramid L. Finally, the modified L_s and the Gaussian pyramid of the source image are used to reconstruct. Figure 1 and 6 show the adjustment of the color tone and texture related features.

To investigate whether the evoked emotion distributions can be pushed towards any of the six directions via the transferring method, we experiment by moving neutral images towards each of the other emotion categories. We construct a set of source images S_s consisting of the 100 most neutral images in Emotion6 in terms of the ground truth of evoked emotion, and we use Emotion6 as the set of target images S_t. For an image in S_s, each image in S_t takes a turn as the target image and a corresponding transformed image is produced using the method of Figure 5. For each source image, 6 transformed-target pairs are chosen (one for each of Ekman's 6 basic emotions [9] e_i ($i \in \{1, 2, \cdots, 6\}$)) such that the transformed image tr_{e_i} has the highest predicted probability in e_i. These probabilities are predicted by our classifier (Sec. 4) which takes an image as input and outputs its evoked emotion distribution. This results in 600 source-target-transformed triplets. We put the transformed images of these 600 triplets on AMT and collect the ground truth of the transformed images using the same method when building Emotion6.

We use the metrics $M \in \{KLD, BC, CD, EMD\}$

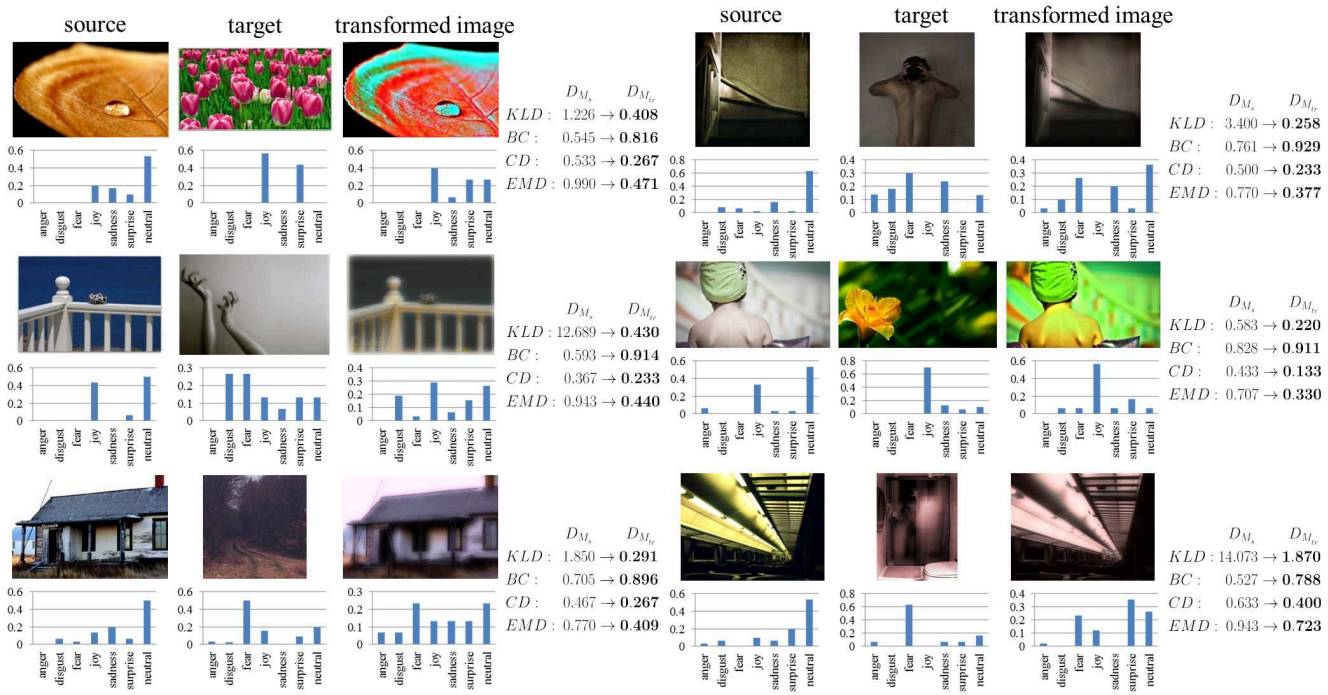

Figure 6. Examples showing the feature transform in transferring evoked emotion distributions. For each example, D_{M_s} and $D_{M_{tr}}$ are provided ($M \in \{KLD, BC, CD, EMD\}$) with better scores marked in bold. The ground truth of evoked emotion distribution from AMT is provided under each image. In each example, the transformed image has closer evoked emotion distribution to that of the target compared to the source in all 4 metrics.

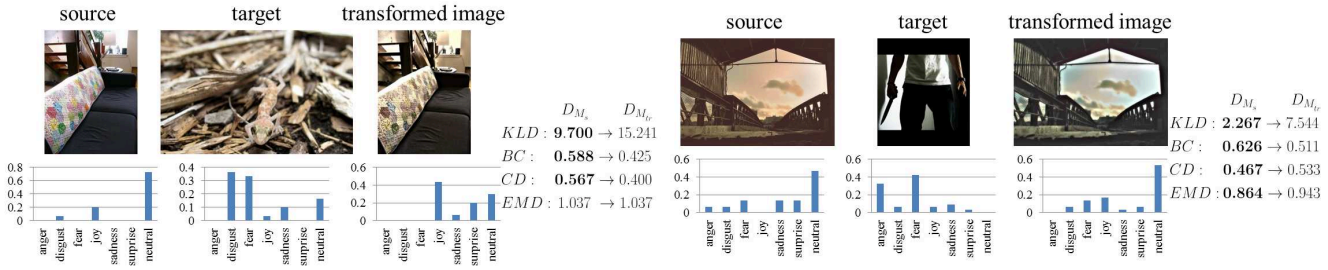

Figure 7. Failure examples of transferring evoked emotion distributions. The ground truth of evoked emotion distribution from AMT is provided under each image. For each example, D_{M_s} and $D_{M_{tr}}$ are provided ($M \in \{KLD, BC, CD, EMD\}$) with better scores marked in bold. The results show that the evoked emotion distribution of the source does not move toward that of the target in these examples.

from Sec. 4 to evaluate the distance between two emotion distributions. For each distance metric M ($M \in \{KLD, BC, CD, EMD\}$), we compute the distances between: 1) source and target images $D_{M_s} = M(d_s, d_t)$. 2) transformed and target images $D_{M_{tr}} = M(d_{tr}, d_t)$, where d_s, d_t, and d_{tr} are the ground truth probability distributions of evoked emotion of the source, target, and transformed images respectively. The results are reported in terms of D_{M_s}, $D_{M_{tr}}$, and P_M, where P_M represents the probability that d_{tr} is closer to d_t than d_s is, using metric M. Table 4 shows that we shape d_s toward d_t successfully in about 60∼70% of the cases in each e_i. Figure 6 depicts some examples with D_{M_s} and $D_{M_{tr}}$ computed from the ground truth given by the user study, showing that our feature transformation moves d_s closer to d_t in terms of all 4 metrics.

We also show some typical failure modes of emotion transferring in Figure 7. There are two main reasons for such failure cases: 1) d_s may be mostly caused by the high level semantics such that the modification in low-level features can hardly shape d_s closer to d_t. 2) d_t may be also mostly caused by the high level semantics such that copying the low-level features of the target cannot totally replicate its emotional stimuli. More examples of transferring evoked emotion and their emotion distributions are provided in the supplementary material.

In an additional experiment, randomly selecting 6 target images from S_t for each of 100 source images in S_s, we

Category	Anger	Disgust	Fear	Joy	Sadness	Surprise
P_{KLD}	0.74	0.64	0.70	0.79	0.68	0.70
P_{BC}	0.65	0.61	0.68	0.68	0.58	0.66
P_{CD}	0.69	0.61	0.56	0.78	0.70	0.66
P_{EMD}	0.64	0.69	0.72	0.79	0.63	0.80

Table 4. The results of transferring evoked emotion in terms of P_M ($M \in \{KLD, BC, CD, EMD\}$) in each category, which shows that in more than a half cases, the transformed image has a closer emotion distribution to the target (versus the source).

generate the corresponding 600 transformed images and collect their emotion distributions judged by AMT with the same setting as that of the previous experiment. The resulting P_Ms are 0.67, 0.56, 0.62, and 0.61 for $M = KLD, BC, CD, EMD$ respectively, which shows that the transformation produces an evoked emotion distribution closer to the target (versus the source image) with 99% confidence from binomial test.

To show that our framework of emotion transferring actually transforms the evoked emotion distribution of the source image towards that of the "correct" target, we perform cross evaluation with the ground truth of 600 triplets. Assume t_{e_i} is the corresponding target image of tr_{e_i} given a source image. For each source image in S_s, we compute all $D_{M_{ij}} = M\left(d_{t_{e_i}}, d_{tr_{e_j}}\right), \forall i, j \in \{1, 2, \cdots, 6\}$, where $d_{t_{e_i}}$ and $d_{tr_{e_j}}$ are the ground truth probability distributions of evoked emotion of t_{e_i} and tr_{e_j} respectively and $M \in \{KLD, BC, CD, EMD\}$. For each source image in S_s, each t_{e_i}, and each M, we compare all 6 $D_{M_{ij}}$ ($j \in \{1, 2, \cdots, 6\}$) and compute $P_{1/6\_M}$ which is defined as the probability that the following condition is true:

$$i = \begin{cases} \arg\min_j D_{M_{ij}}, \text{if } M \neq BC \\ \arg\max_j D_{M_{ij}}, \text{if } M = BC \end{cases} \quad (1)$$

In other words, $P_{1/6\_M}$ is the probability that a transformed image's emotion distribution matches its target's emotion distribution more closely than the other five transformed images from the other five targets. Table 5 lists all $P_{1/6_M}$s, comparing them with 16.67%, the probability of randomly selecting 1 transformed image out of 6. Table 5 shows our strategy moves the evoked emotion distribution of the source image closer towards that of the desired target (versus other transformed images) with the probability higher than chance in most cases. Considering all transform categories as a whole for $M \in \{KLD, BC, EMD\}$, we achieve a confidence level higher than 95% using binomial test.

Inspired by the user study by Wang et al. [29], we perform an additional user study comparing pairs of images. Wang's algorithm [29] outputs a color-adjusted image given an input image and an emotion keyword. Photoshop experts were hired to produce an output image which represents

Transform Category	Anger	Disgust	Fear	Joy	Sadness	Surprise	All
$P_{1/6\_KLD}$	**0.23**	0.15	**0.22**	**0.27**	0.13	**0.21**	**0.20**
$P_{1/6\_BC}$	**0.26**	0.14	**0.21**	**0.26**	0.12	**0.19**	**0.20**
$P_{1/6\_CD}$	**0.19**	**0.22**	0.15	**0.18**	**0.18**	0.15	**0.18**
$P_{1/6\_EMD}$	**0.26**	**0.22**	**0.19**	**0.19**	**0.18**	**0.21**	**0.21**

Table 5. Cross evaluation results in terms of $P_{1/6\_M}$ ($M \in \{KLD, BC, CD, EMD\}$) in each transform category (the numbers larger than $1/6$ are marked in bold), which shows that our framework of emotion transferring moves the evoked emotion distribution of the source closer towards that of the desired target (versus other transformed images) with the probability higher than chance in most cases.

the same emotion given the same input. In Wang's experiment [29], they ask subjects to select one image better corresponding to the emotion keyword out of two images: the outputs of their algorithm and the Photoshop expert. However, there are two major shortfalls in Wang's experiment [29]: 1) Only 20 pairs of images were studied, a small sample size. 2) The output of their algorithm is not compared directly with the input image, neglecting the possibility that both the outputs of their algorithm and the Photoshop expert are worse than the input image.

We improve Wang's experiment [29] by making the following two adjustments: 1) We use the 600 neutral and transformed image pairs for the user study. 2) For each pair, we upload the source and transformed images in random order to AMT and ask 15 subjects to choose the one image (of two) that better corresponds to the emotion keyword. Out of all 600×15 evaluations, 66.53% selections indicate that our transformed image better corresponds to the associated emotion keyword, roughly comparable to the 69.70% reported by Wang et al. [29]. In 76.50% of the pairs, more subjects think our transformed image better matches the emotion keyword than the source image. This user study shows that our framework performs well when targeting a specific dominant emotion. Further, as shown in Table 4, our framework can transfer emotion in distributions, more general than previous work [29].

6. Conclusion

This work introduces the idea of representing the emotional responses of observers to an image as a distribution of emotions. We describe methods for estimating the emotion distribution for an image, and describe a method for modifying an image to push its evoked emotion distribution towards a target image. Further, our proposed emotion predictor, CNNR, outperforms other methods including using SVR with the features from previous work and the optimal dominant emotion baseline, the upper-bound of the emotion predictors that predict a single emotion. Finally, we propose a novel image database, Emotion6, and provide ground truth of valence, arousal, and probability distributions in evoked emotions.

References

[1] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994. 3, 4

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. 4

[3] Y.-Y. Chen, T. Chen, W. H. Hsu, H.-Y. M. Liao, and S.-F. Chang. Predicting viewer affective comments based on image content in social media. In *ICMR*, page 233, 2014. 2

[4] E. S. Dan-Glauser and K. R. Scherer. The geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior Research Methods*, 43(2):468–477, 2011. 3, 4

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 4

[6] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia*, 19(3):34–41, 2012. 3

[7] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis. The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. *Lecture Notes in Computer Science*, 4738:488–501, 2007. 3, 4

[8] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In *the 28th Annual Conference on Computer Graphics and Interactive Techniques*, pages 341–346, 2001. 2

[9] P. Ekman, W. V. Friesen, and P. Ellsworth. What emotion categories or dimensions can observers judge from facial behavior? *Emotion in the Human Face*, pages 39–55, 1982. 2, 3, 4, 6

[10] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth. The world of emotions is not two-dimensional. *Psychological Science*, 18(2):1050–1057, 2007. 2, 3, 4

[11] M. Gendron and L. F. Barrett. Reconstructing the past: a century of ideas about emotion in psychology. *Emotion Review*, 1(4):316–339, 2009. 2

[12] M. Grundland and N. A. Dodgson. Color histogram specification by histogram warping. In *SPIE*, volume 5667, pages 610–621, 2005. 2

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 4, 5

[14] D. Joshi, R. Datta, Q.-T. Luong, E. Fedorovskaya, J. Z. Wang, J. Li, and J. Luo. Aesthetics and emotions in images: a computational perspective. *IEEE Signal Processing Magazine*, 28(5):94–115, 2011. 2, 3

[15] B. Jou, S. Bhattacharya, and S.-F. Chang. Predicting viewer perceived emotions in animated gifs. In *ACM Multimedia*, Grand Challenge, 2014. 2

[16] J. Jun, L.-C. Ou, B. Oicherman, S.-T. Wei, M. R. Luo, H. Nachilieli, and C. Staelin. Psychophysical and psychophysiological measurement of image emotion. In *the 18th Color and Imaging Conference: Color Science and Engineering Systems, Technologies, and Applications*, pages 121–127, 2010. 2

[17] S. G. Koolagudi and K. S. Rao. Emotion recognition from speech: a review. *International Journal of Speech Technology*, 15(2):99–117, 2012. 3

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012. 4

[19] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (IAPS): affective ratings of pictures and instruction manual. technical report a-8. 2008. 3

[20] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the International Conference on Multimedia*, pages 83–92, 2010. 2, 4, 5, 6

[21] K. Matsumoto, K. Kita, and F. Ren. Emotion estimation of wakamono kotoba based on distance of word emotional vector. In *NLPKE*, pages 214–220, 2011. 5

[22] M. Mignotte. An energy-based model for the image edge-histogram specification problem. *IEEE Transactions on Image Processing*, 21(1):379–386, 2012. 2

[23] A. Ortony and T. J. Turner. What's basic about basic emotions? *Psychological Review*, 97(3):315–331, 1990. 2, 3

[24] K.-C. Peng, K. Karlsson, T. Chen, D.-Q. Zhang, and H. Yu. A framework of changing image emotion using emotion prediction. In *IEEE International Conference on Image Processing*, pages 4637–4641, 2014. 2

[25] T. Pouli and E. Reinhard. Progressive histogram reshaping for creative color transfer and tone reproduction. In *the 8th International Symposium on Non-Photorealistic Animation and Rendering*, pages 81–90, 2010. 2, 6

[26] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980. 2

[27] E. M. Schmidt and Y. E. Kim. Modeling musical emotion dynamics with conditional random fields. In *ISMIR*, pages 777–782, 2011. 5

[28] M. Solli and R. Lenz. Emotion related structures in large image databases. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 398–405, 2010. 2, 3, 4, 6

[29] X. Wang, J. Jia, and L. Cai. Affective image adjustment with a single word. *The Visual Computer*, 2012. 2, 8

[30] X. Wang, J. Jia, J. Yin, and L. Cai. Interpretable aesthetic features for affective image classification. In *IEEE International Conference on Image Processing*, pages 3230–3234, 2013. 2, 3, 4, 5, 6

[31] Y.-H. Yang and H. H. Chen. Machine recognition of music emotion: a review. *ACM Transactions on Intelligent Systems and Technology*, 3(3):1–30, 2012. 3