WHERE DO EMOTIONS COME FROM? PREDICTING THE EMOTION STIMULI MAP

Kuan-Chuan Peng[⋆] *Amir Sadovnik*[†] *Andrew Gallagher*[‡] *Tsuhan Chen*[⋆]

[⋆] Cornell University [†] Lafayette College [‡] Google Inc.

ABSTRACT

Which parts of an image evoke emotions in an observer? To answer this question, we introduce a novel problem in computer vision — predicting an Emotion Stimuli Map (ESM), which describes pixel-wise contribution to evoked emotions. Building a new image database, EmotionROI, as a benchmark for predicting the ESM, we find that the regions selected by saliency and objectness detection do not correctly predict the image regions which evoke emotion. Although objects represent important regions for evoking emotion, parts of the background are also important. Based on this fact, we propose using fully convolutional networks for predicting the ESM. Both qualitative and quantitative experimental results confirm that our method can predict the regions which evoke emotion better than both saliency and objectness detection.

Index Terms— Emotion stimuli map, fully convolutional networks

1. INTRODUCTION

Images, when viewed, can cause a variety of emotional responses, depending on not only the arrangement of one or more objects in the image but also the emotional state or background of the viewer. For example, an image of bungee jumping can make outdoors-loving people excited, but it can evoke fear in those afraid of heights. Even within the same image, different regions contribute to the viewer's evoked emotion differently. Imagine we crop the yellow, green, and red rectangles (Fig. 1 (c), (d), and (e)) from Fig. 1 (a) and present them individually to a viewer without showing the viewer the full image context (a). The emotional response to (e) is more similar to (a) than to either (c) or (d). We represent the varying degree of influence that regions of an image have on the emotional responses of viewers with an *Emotion Stimuli Map* (ESM), shown in Fig. 1 (b), where brighter areas represent higher influence. The ESM (b) is produced by averaging across selections from a user study, and matches the observation that (e) best captures the emotion-inducing regions of (a). In this work, we are interested in predicting the ESM.

Recently, emotion-related topics have gained increasing attention in computer vision, especially affective image classification. Machajdik and Hanbury [1] perform affective image classification on both artistic and realistic images. Solli and Lenz [2] use Internet images in their experiment, but Wang et al. [3] focus on affective image classification of artistic photos or abstract paintings. Peng et al. [4] also predict and transfer emotion distributions using Internet images. In addition, there are related works studying emotions from animated GIFs [5] and multilingual perspectives [6]. Even though different forms of multimedia have been explored, none of the previous works analyze the influence of various regions in an image on emotions. There is no benchmark for evaluating the ESM. We use the images collected in Emotion6 database [4] to build a benchmark database, EmotionROI, for predicting the ESM. The ground truth of the ESM provided in EmotionROI database is generated based on the answers marked by the users in a user study. The details of EmotionROI are explained in Sec. 2.

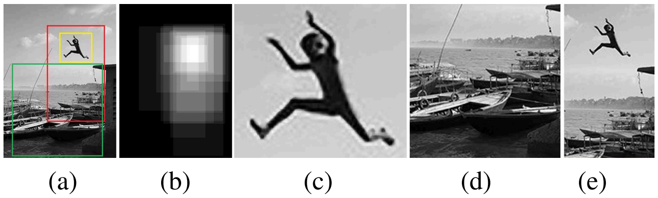

Fig. 1: An example showing that different regions in an image contribute to the viewer's evoked emotion differently. (c), (d), and (e) are cut from the yellow, green, and red rectangles of (a) respectively. (b) shows the regions in (a) which affect the evoked emotion the most marked by user study. (e) will evoke more similar emotions as those evoked by (a) compared with (c) and (d) because (e) contains not only the person jumping but other emotion-related areas, which is consistent with (b), the ground truth of the emotion stimuli map of (a).

Saliency detection [7, 8, 9] and objectness measurement [10] are two popular topics closely related to predicting the ESM. While saliency and objectness detection tend to find salient objects in an image, the ESM captures the regions affecting the evoked emotion and those regions may contain not only the salient objects but other emotion-related areas. For example, Fig. 2 (c) and (d) are the results of saliency [7] and objectness [10] detection respectively with Fig. 2 (a) as the input. Fig. 2 (c) focuses on the dark salient areas, but Fig. 2 (d) emphasizes the withered flower. Neither Fig. 2 (c) nor (d) perfectly captures the ground truth ESM in Fig. 2 (b),

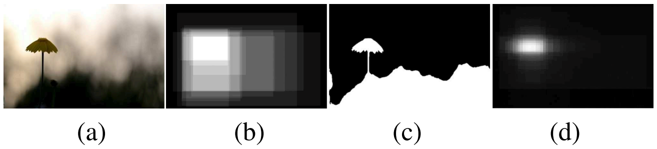

Fig. 2: An example showing the difference of saliency, objectness detection and the emotion stimuli map. (b) is the ground truth emotion stimuli map using (a) as the input image. (c) and (d) correspond to saliency [7] and objectness [10] detection, respectively. Neither (c) nor (d) perfectly captures (b), where two thirds of the subjects convey that the area affecting evoked emotions includes not only the flower but also other emotion-related areas.

where two thirds of the subjects convey that the area affecting evoked emotions the most includes not only the flower but also other emotion-related areas. In this work, we use fully convolutional networks to predict the ESM with the result closer to the ground truth versus state-of-the-art algorithms for saliency and objectness detection.

Previous work related to saliency detection [11] often considers using eye-tracking equipments to gather ground truth and perform validation. However, when building the ground truth ESM in EmotionROI, we choose not to use eye-tracking equipments because of the following reasons: 1) Saliency detection is different from predicting the ESM in terms of the task definition, and we also show their difference in Fig. 2 where (b) and (c) are not even similar. 2) Where humans look at in an image may implicitly reveal partial areas which affect the evoked emotion the most. However, we believe that directly asking the subjects to mark the emotion-related areas is a more straightforward and efficient method which can avoid potential errors caused by the inference from the eye-tracking results.

To the best of our knowledge, this is the first paper in computer vision addressing the problem of predicting the ESM. We make the following contributions: 1) We build a benchmark database, EmotionROI, for predicting the ESM by performing a user study and collecting the ground truth ESMs of the images provided in the Emotion6 database [4]. The EmotionROI database is available online [12]. 2) We propose using fully convolutional networks to predict the ESM. Our method predicts more accurate ESMs than do the state-of-the-art algorithms of saliency and objectness detection.

2. PROPOSED DATABASE AND USER STUDY

We use the images in the Emotion6 database [4] to build our proposed benchmark database, EmotionROI, for predicting the ESM. The EmotionROI database contains the ground truth ESMs collected by asking people to identify the regions in the images which most influence their evoked emotions.

Emotion6 [4] consists of 6 emotion categories with 330 images per category. For each image, the following information is provided: 1) The ground truth of evoked emotion distribution in terms of emotion keywords. 2) The emotion keyword used to search each image. Emotion6 [4] is assembled from Flickr by entering the 6 category keywords corresponding with Ekman's 6 basic emotions [13] (anger, disgust, joy, fear, sadness, and surprise) and their synonyms as the searching keywords, followed by a step of human moderation to remove erroneous images. Emotion6 contains 1980 images in total. Each image is approximately VGA resolution.

Adopting all the 1980 images in Emotion6 [4], we use Amazon Mechanical Turk (AMT) to collect responses from subjects, building the ground truth ESMs in EmotionROI. We ask the subject to draw a rectangle enclosing the part of the image that most influences the evoked emotion. The leftmost image in Fig. 3 is a snapshot of the interface. We collect the responses in a similar way as that used in Emotion6 [4]. We consider the emotion categories provided by Emotion6 [4] and create 220 different HITs (each HIT contains 10 images) for AMT that meet the following constraints: 1) Each HIT contains at least one image from each of the 6 categories. 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. We enforce the following regulations to be consistent with the previous database [14]: 1) The same subject can only respond to each image or HIT at most once, and each subject cannot respond to more than 55 different HITs to increase diversity. 2) We collect 15 responses for each image to have statistically significant results. 432 unique subjects participate in the experiment, responding to an average of 76.4 images each. We assume the influence of each pixel on evoked emotions is proportional to the number of drawn rectangles covering that pixel. The ground truth ESMs are normalized to the range between 0 to 1. Fig. 3 shows some example images in EmotionROI and the corresponding ground truth ESMs. Fig. 3 also shows the emotion keyword used to search each image (provided by Emotion6 [4]).

3. PREDICTING EMOTION STIMULI MAPS

We propose Fully Convolutional Networks with Euclidean Loss (FCNEL) to predict the ESM. Fully Convolutional Networks (FCN) have been shown to achieve the state-of-the-art performance in semantic segmentation since Long et al. [15] popularized this approach. We leverage FCN because FCN provides an end-to-end training framework which generates pixel-wise dense prediction of the same resolution as the input image. Specifically, we adopt the FCN in Long's work [15] with single stream, 32-pixel-prediction-stride version based on the AlexNet [16] architecture. We choose this standard and relatively simple architecture versus other deeper or more complicated networks because the size of our database is relatively small. Therefore, we want to keep the number of parameters which need to be trained manageable.

In Long's work [15], a softmax loss layer is used as the objective function in the FCN for semantic segmentation

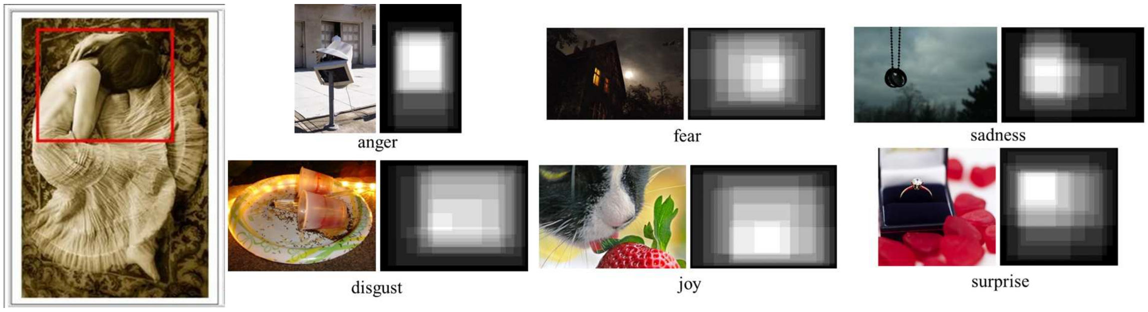

Fig. 3: The leftmost image is a screenshot of the interface of our user study on Amazon Mechanical Turk. We ask the subject to draw a rectangle enclosing the part of the image that most influences the evoked emotion. The other images are some examples from EmotionROI with the corresponding ground truth emotion stimuli maps. The emotion keyword used to search each image (provided by Emotion6 [4]) is displayed under the image.

where any two different semantic labels are mutually exclusive. However, in predicting the ESM, we want to predict the influence on evoked emotions at each pixel location, not one out of many mutually exclusive class labels. Therefore, we change the topmost fully connected layer of FCN such that only one output representing the influence on evoked emotions is predicted at each pixel location. We also change the softmax loss layer to a Euclidean loss layer such that the modified FCN can be trained to predict the ESM close to the corresponding ground truth in terms of L2-norm. To distinguish FCN using Euclidean loss from the common FCN used in semantic segmentation, we use FCNEL to refer to the former method.

We train the FCNEL for predicting the ESM by using the Caffe [17] framework. We pre-train our FCNEL with the reference model, FCN-AlexNet, which is trained for PASCAL VOC segmentation task [18] and provided by Long et al. [15]. After pre-training, we fine-tune all the parameters of the FCNEL with the EmotionROI training data. To efficiently train FCNEL but also avoid a convergence issue of the learned parameters, we empirically set the base learning rate to 10^{-8}. The number of training iterations is set such that each training example is visited at least 20 times. For other training details, we adopt the same setting provided by Long et al. [15] unless otherwise specified.

4. EXPERIMENTAL SETTING

We experiment on our proposed EmotionROI database, and we use the same training/testing split as that used in Peng's work [4] unless otherwise specified. Therefore, there are 1386/594 training/testing images out of all 1980 images in EmotionROI database.

Evaluation metrics: We use 8 evaluation metrics — mean absolute error (MAE), $precision$, $recall$, 4 commonly used F-measures ($F_{0.5}$, $F_{\sqrt{0.3}}$, F_1, and F_2 scores), and the Precision-Recall (PR) curve. All the predicted ESMs are normalized to 0 to 1 before evaluation. MAE corresponds to the mean absolute error between the value of the predicted map and the ground truth at all pixel locations. $precision$ is defined as the ratio of emotionally involved pixels correctly assigned to all the pixels identified in the predicted map, while $recall$ represents the percentage of detected emotionally involved pixels out of all the pixels marked in the ground truth. Before computing $precision$ and $recall$, we binarize each predicted map adaptively according to its Otsu threshold [19]. F-measure is defined in terms of $precision$ and $recall$ as follows:

$$F_\beta = \left(1+\beta^2\right) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}, \qquad (1)$$

where β controls the weighting between $precision$ and $recall$. In addition to the 3 common F-measures ($F_{0.5}$, F_1, and F_2), we also include $F_{\sqrt{0.3}}$ because it is a standard metric for saliency detection [20]. For the PR curve, we binarize the predicted map using each threshold between $[0, 255]/255$, which is similar as the method used in [20].

Baselines — saliency and objectness detection: As the first work predicting the ESMs, this paper compares the ESM with saliency and objectness to emphasize the difference between these tasks. We apply our proposed method to all the EmotionROI testing images to predict the ESMs, and compare the results with those of the state-of-the-art method of saliency [7] and objectness [10] detection. We also compute the MAE of context-aware saliency detection [8], and the results are similar to those of Cheng's method [7]. Therefore, we only report the results using Cheng's method [7] for saliency detection.

5. EXPERIMENTAL RESULTS

We evaluate the predicted ESMs of the 594 testing images using the 8 evaluation metrics mentioned in Sec. 4. We show the PR curve in Fig. 4, and report the average $precision$ and $recall$, and 4 F-measures in Fig. 5. FCNEL outperforms saliency [7] and objectness [10] detection under most evaluation metrics. The only exception is the results of $precision$, where objectness [10] and FCNEL show comparable performance. Since the most salient object usually has a relatively

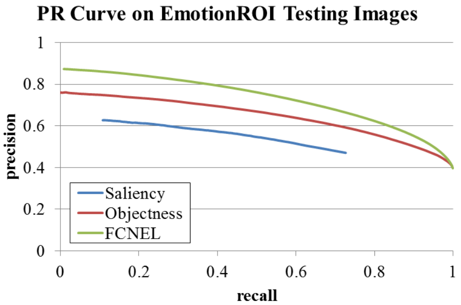

Fig. 4: The performance of predicting the ESMs in PR curve, where FCNEL outperforms saliency [7] and objectness [10] detection.

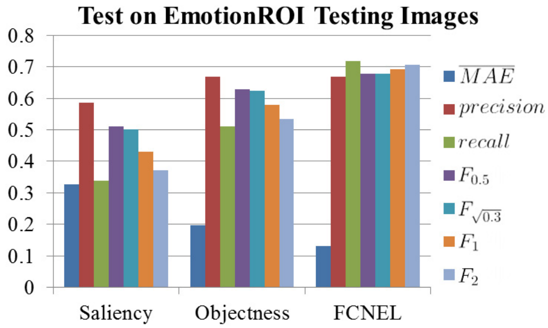

Fig. 5: The performance of predicting the ESMs in 7 evaluation metrics. $\overline{MAE}$ represents the mean of MAE (lower is better). For all the other metrics, higher is better. FCNEL outperforms saliency [7] and objectness [10] detection in all these metrics.

high value on the ESM, it makes sense that both saliency [7] and objectless [10] achieve reasonable $precision$. However, what affects evoked emotions is not only salient objects but also other emotionally involved areas, as shown in the ground truth of EmotionROI, Fig. 1 and 2. FCNEL shows better ability in identifying those emotionally involved areas compared with both saliency [7] and objectness [10] detection, which is reflected in metrics involving $recall$.

Fig. 6 shows the qualitative and quantitative results of predicting the ESMs with some EmotionROI testing images as input. Column (a) to (e) are the input image, the ground truth ESM, the result of saliency detection [7], the result of objectness detection [10], and the result of FCNEL respectively. For column (a), the emotion keyword under each image is the keyword used to search that image according to the information provided in Emotion6 [4]. For column (b) to (e), the corresponding MAE is shown under each image. Compared with saliency [7] and objectness [10] detection, the ESMs predicted by FCNEL show that the features learned from EmotionROI training images improve the results of predicting the ESM.

6. CONCLUSION

We identify a novel problem, predicting the emotion stimuli map (ESM), in computer vision. Building a new image database, EmotionROI, as a benchmark for predicting the ESM, we address the major difference between the ESM, saliency and objectness detection — the regions affecting evoked emotions contain both the main objects and additional contextual background necessary for the viewer to fully experience the emotion of the image.

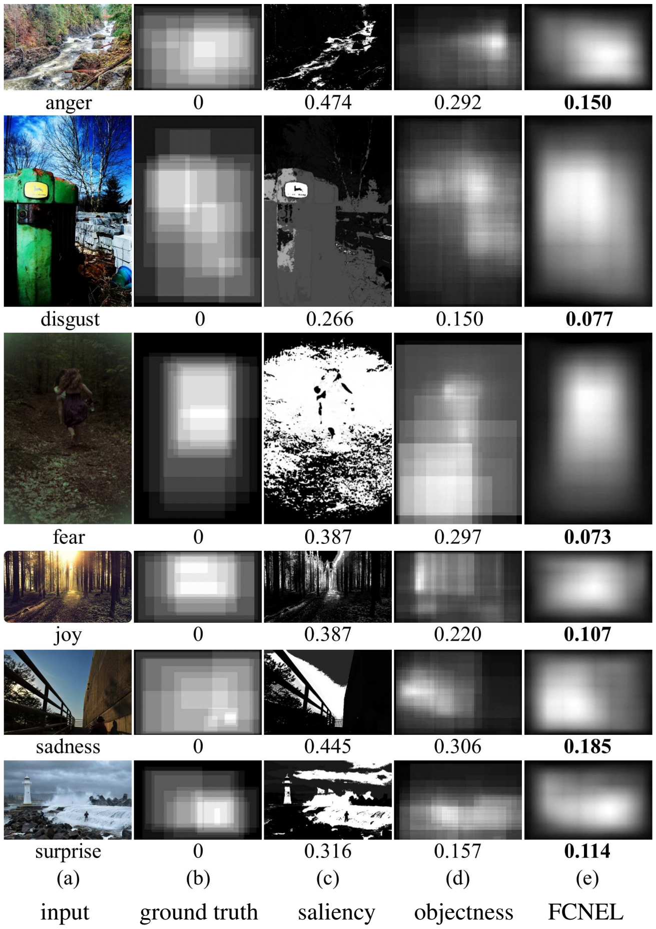

Fig. 6: The qualitative and quantitative results of predicting emotion stimuli maps with some EmotionROI testing images as input. The representation of each column is as follows: (a) input image, (b) the ground truth emotion stimuli map, (c) the result of saliency detection [7], (d) the result of objectness detection [10], (e) the result of FCNEL. The emotion keyword used to search each input image is shown under each image in column (a) according to the information provided in Emotion6 [4]. For column (b) to (e), the corresponding MAE is shown under each image. FCNEL (column (e)) predicts more accurate emotion stimuli maps than other baselines (column (c) and (d)) do for these examples.

Based on the above finding, we propose FCNEL for predicting the ESM. FCNEL leverages fully convolutional networks to directly learn from the EmotionROI training images. Our qualitative and quantitative results show that FCNEL predicts more accurate ESMs compared with saliency and objectness detection.

7. REFERENCES

[1] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *International Conference on Multimedia*. ACM, 2010, pp. 83–92.

[2] M. Solli and R. Lenz, "Emotion related structures in large image databases," in *International Conference on Image and Video Retrieval*. ACM, 2010, pp. 398–405.

[3] X. Wang, J. Jia, J. Yin, and L. Cai, "Interpretable aesthetic features for affective image classification," in *International Conference on Image Processing*. IEEE, 2013, pp. 3230–3234.

[4] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen, "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in *Computer Vision and Pattern Recognition*, 2015.

[5] B. Jou, S. Bhattacharya, and S.-F. Chang, "Predicting viewer perceived emotions in animated GIFs," in *International Conference on Multimedia*. ACM, 2014, pp. 213–216.

[6] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S.-F. Chang, "Visual affect around the world: A large-scale multilingual visual sentiment ontology," in *International Conference on Multimedia*. ACM, 2015.

[7] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *International Conference on Computer Vision*. IEEE, 2013, pp. 1529–1536.

[8] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.

[9] J. Luo, A. Singhal, S. P. Etz, and R. T. Gray, "A computational approach to determination of main subject regions in photographic images," *Image and Vision Computing*, vol. 22, pp. 227–241, 2004.

[10] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189–2202, November 2012.

[11] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *International Conference on Computer Vision*, 2009, pp. 2106–2113.

[12] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen, "The Cornell EmotionROI Image Database," http://chenlab.ece.cornell.edu/downloads.html.

[13] P. Ekman, W. V. Friesen, and P. Ellsworth, "What emotion categories or dimensions can observers judge from facial behavior?," *Emotion in the Human Face*, pp. 39–55, 1982.

[14] 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*, vol. 43, no. 2, pp. 468–477, 2011.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Computer Vision and Pattern Recognition*, 2015.

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

[17] 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.

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results," http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html.

[19] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[20] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Computer Vision and Pattern Recognition*, 2013, pp. 3166–3173.