WHEN SALIENCY MEETS SENTIMENT: UNDERSTANDING HOW IMAGE CONTENT INVOKES EMOTION AND SENTIMENT

Honglin Zheng, Tianlang Chen, Quanzeng You, Jiebo Luo

Department of Computer Science
University of Rochester, Rochester, NY 14627

ABSTRACT

Sentiment analysis is crucial for extracting social signals from social media content. Due to the prevalence of images in social media, image sentiment analysis is receiving increasing attention in recent years. However, most existing systems are black-boxes that do not provide insight on how image content invokes sentiment and emotion in the viewers. On the other hand, psychological studies have confirmed that salient objects in an image often invoke emotions. In this work, we investigate more fine-grained and more comprehensive interaction between visual saliency and visual sentiment. In particular, we partition images in several *meta-level* scene-type dimensions that are relevant to most images, including: open-closed, natural-manmade, indoor-outdoor, and face-noface. Facilitated by state of the art saliency detection algorithm and sentiment classification algorithm, we examine how the sentiment of the salient region(s) in an image relates to the overall sentiment of the image. The experiments on a representative image emotion dataset have shown interesting correlation between saliency and sentiment in different scene types and shed light on the mechanism of visual sentiment evocation.

***Index Terms*—** saliency, sentiment perception, scene attribute

1. INTRODUCTION

1.1. Visual Sentiment Analysis

The power of convolutional neural networks has been harnessed recently to discover the sentiment carried in images [1, 2]. However, most of the work are dedicated to fine-tuning pre-trained deep neural network models, such as VGGNet [3] and ResNet [4]. Most models are used as black boxes and little research has paid attention to what elements or attributes of the image are responsible for invoking emotions.

1.2. Saliency Detection

In the meantime, saliency detection [5, 6] is also shifting from common visual feature based classifiers to deep learning based models [7, 8]. Recently, researchers start to utilize CNN to capture high-level visual concepts and produce saliency models with better detection performance [7, 9].

Saliency detection focuses on attention analysis, while visual sentiment analysis focuses on emotion analysis. In psychological and neuroscientific area, there is an ongoing discrepancy between those who suggest that emotional perception is automatic, namely in the manner that it is independent of top-down factors such as attention [10], and those illustrating the dependence on attention [11, 12]. However, no computer vision research has been done to discover how emotion depends on attention, especially salient object(s), in an image. Even though it has been an active research topic in neuroscience, most of the neuroscientists fail to answer the question that whether the salient object shares similar sentiment with the entire image. If so, what category do they fall into? Do they share some common attributes like human-made objects or human faces? In this work, we address the questions above and make several contributions:

- We investigate fine-grained interaction between visual saliency and visual sentiment over several primary scene types, including open-closed, natural-manmade, indoor-outdoor, and face-noface.
- We employ the state-of-the-art saliency algorithm and visual sentiment classification model to accurately analyze region-level interactions.
- We utilize a large public image emotion dataset to discover the relationship between saliency and sentiment over different scene types to understand the evocation mechanism of visual sentiment.

2. RELATED WORK

To the best of our knowledge, there is no related work on combining saliency and sentiment analysis to understand how image invokes human emotion. The most relevant work is image sentiment localization. Sun *et al.* [13] adopt an off-the-shelf object detection algorithm [14] to generate random proposals, and then the sentiment of both the proposals and the entire image will be evaluated to discover affective sentiment regions. However, this method utilizes explicit object

recognition, which has to generate thousands of proposal windows in order to yield to a high recall rate of affective regions proposals. It also incurs a high computational overhead.

In [15], the authors use a pre-trained and fine-tuned model to predict Emotion Stimuli Map and conclude that neither saliency nor objectness can correctly predict the image regions that evoke emotion. However, they fail to justify the choice of a dataset that consists of predominantly landscape scenes, and do not give any insight on why the saliency detection method does not work for detecting affective emotion region of such images.

Inspired by these recent discoveries, we are particularly interested in analyzing the correlation between saliency and the emotion of images. Thus, we propose a framework to understand what attributes are involved in sentiment perception invoked by salient attention.

3. METHODOLOGY

3.1. Framework Overview

The overall structure of the framework is summarized in Figure 1. Details of the steps are as follows:

1. We use a saliency detection model [16] to detect salient object(s), if any, for each image. Both the detected regions and the original input images are inputs of the next step.

2. A state-of-the-art sentiment classification model [17] is used to obtain the sentiment scores for both the whole image and all of its salient regions produced in Step 1, if any.

3 & 4. Based on the sentiment scores obtained above, we are able to find out for each image, if any salient object shares the same sentiment with the entire image. Consequently, we can partition the entire dataset into two parts: those images where at least one of their salient objects agree with the whole image's sentiment, and those images where none of their salient objects agrees with the whole image. For simplicity, images without any salient object detected are simply excluded from further analysis. Please refer to the Visual Sentiment Analysis section for details about the definition of sentiment agreement.

5 & 6. Within the two partitions obtained from previous steps, we apply a scene attribute detector and human face detector to further partition each part based on the detected attribute of the image. We choose the four meta-level attributes that reflect general characteristics of most scenes, instead of specific scene attributes that only apply to a small percentage of images. Note that the three categories of indoor VS outdoor, open VS closed, and natural VS man-made were the subject of the early study on spatial envelope [18], which preceded the scene attribute study [19]. As for face VS noface, it is an apparent choice since a majority of images contain faces. *Open* describes vast scene, while *closed* describes images with closed up objects and narrow space. Other attributes are quite evident and are well defined in [18].

7. We evaluate the results of the classification mentioned above and gather any interesting findings.

3.2. Dataset

We use the dataset released by [20] for visual emotion prediction. Currently, this is the largest publicly available dataset labeled manually for visual emotion analysis. This dataset has 8 different categories. For each category, we randomly sample 30% of the first 8000 images for the experiment. And we end up with a collection of 13,048 images with one or more salient objects.

3.3. Salient Object Detection

We adopt a state-of-the-art saliency detection model [16] to detect whether an image has any salient object and if there is, locate those salient objects. This model leverages the high expressiveness of the VGGNet to detect whether there is any salient object in it and we save the detected salient objects for subsequent scene-attribute based partitioning.

3.4. Visual Sentiment Analysis

We use the state-of-the-art CNN model proposed by [17]. Their model is fine-tuned on an existing CNN model trained for image recognition. Several boosting techniques are added to improve the overall performance and produce one of the best results on sentiment analysis, with an accuracy of over 82%. We employ this model to extract the sentiment distribution on positive and negative categories for a given image. We denote the sentiment score of image i as $p_i = (\Pr(p)_i, \Pr(n)_i)$, which represents the probability of the emotion being positive and negative, respectively. Next, we define SAR, Sentiment Agreement Rate. For each image i,

$$\mathrm{SAR}(i) = \min\left(\parallel \Pr(p)_i - \Pr(p)_s \parallel\right), s \in S(i) \quad (1)$$

where $S(i)$ is the set of all salient objects of image i. Intuitively, SAR is designed to find the saliency region that has the closest sentiment score with the original overall image.

Given this metric, we are able to partition the dataset into two groups: $agree$ and $disagree$. We define that for an image, its salient object(s) $agrees$ with the sentiment of the whole image if $\mathrm{SAR}(i) < \theta$ where θ is the threshold. In other words, for an image, if the sentiment score of at least one of its salient object(s) is within the range of θ of the whole image's sentiment score, then it is regarded as an image whose salient object(s) agrees with the sentiment of the entire image, and is classified to the partition of $agree$. Otherwise, it will be classified to $disagree$. For experiments, we use the agreement threshold of 0.08, 0.1, 0.15, 0.18 and 0.2.

3.5. Attribute Extraction and Categorization

For each obtained partition from the previous step, we fine-tune a state-of-the-art scene recognition model, Places-

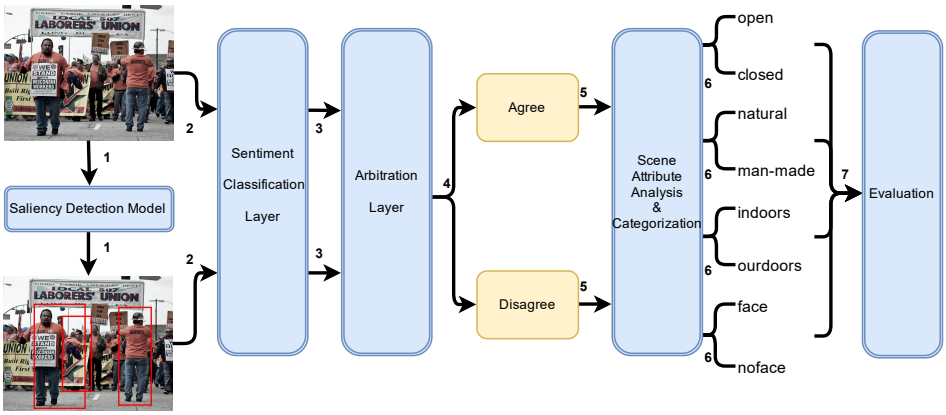

Fig. 1. Overview of our framework. Details of individual steps are described in the Methodology section.

CNN [21], to further classify it into indoors and outdoors. For a given image, the model will produce a 205x1 vector, and each of the entry corresponds to the probability of one of 205 scenes. According to the given indoors/outdoors label reference, we use the labels of top 5 predicted places categories to vote whether it is indoor or outdoor.

Next, we use the deep features from the last fully connected layers to detect 102 SUN scene attributes [19]. We compare scores of open vs closed, natural vs man-made. Whichever attribute of a pair has a higher probability score than the other one is used to represent the image. We define that attribute as a *dominant attribute* in its attribute pair. Finally, we utilize a leading cloud-based face recognition service Face++ [22] to detect whether there is any face in a given image, based on which we can further classify the partition into images that contain face(s) and those that do not.

4. EXPERIMENT AND EVALUATION

We conduct experiments on different sentiment agreement thresholds and obtain the results shown in Table 1. To understand the table, we introduce *discrimination ratio*, DR,

$$\mathrm{DR}(P_A) = \frac{(P_A - E(P_A)) * |P_A - E(P_A)|}{E(P_A)} \qquad (2)$$

where

$$\mathrm{E}(P_A) = \frac{P * \mathrm{All}_A}{\mathrm{All}} \qquad (3)$$

where $P \in \{agree, disagree\}$ is the partition and A is the attribute, namely open, closed, natural, etc.
P_A: in partition P, the number of images that have dominant attribute A over its opposite. (open-closed, natural-manmade, indoor-outdoor, face-noface are opposite of each other). Also known as *observed frequency* in χ^2 test
$E(P_A)$: expected frequency for the number of images that have dominant attribute A in partition P
P: number of images in partition P
All_A: for the entire dataset, the number of images that have dominant attribute A
All: number of images in the entire dataset

Discrimination Ratio is a modification to χ^2 statistics, inheriting its property of measuring the dependence between two categorical data. If images are randomly split into $agree$ and $disagree$, then the DR rate for each attribute in each partition should be small, since the observed frequency will be as closed to the expected frequency as possible. Additionally, DR also preserves the information of what partition the attribute is classifying the image into. A more positive discrimination ratio of an attribute A in a partition P indicates that an image with such an attribute is easier to be classified into P. In other words, the higher the discrimination ratio is, the more likely the images in partition P are going to have attribute A, and a more negative ratio means the more likely the images in partition P will have the opposite attribute. Table 1 suggests some similar patterns for various sentiment agreement thresholds ranging from 0.08 to 0.2:

Table 1. Discrimination Ratio of each attribute in $agree$ partition with different sentiment agreement thresholds

θ	open	closed	natural	manmade
0.2	-9.68	14.82	-4.64	2.85
0.18	-10.31	15.80	-4.76	2.92
0.15	-15.78	24.21	-10.82	6.67
0.1	-16.96	25.77	-7.20	4.38
0.08	-20.38	31.47	-8.37	5.17
θ	indoor	outdoor	face	noface
0.2	17.67	-10.09	54.80	-18.85
0.18	17.57	-9.99	64.08	-22.26
0.15	26.91	-15.41	64.85	-22.76
0.1	34.56	-19.86	96.91	-33.88
0.08	40.55	-23.04	121.33	-41.93

1) Since the absolute value of all DRs, excluding natural and manmade, are greater than $\chi^2_{0.005,1} = 7.88$, we can reject the null hypothesis H_0: The partition of $agree$ and $disagree$ is independent of open-closed, indoor-outdoor, face-noface, at a significance level of 99.5%. Natural and manmade have

the weakest differentiate power, but still can discriminate the dataset at a significance level of at least 90%.

2) **For images whose salient object(s) agree with the whole images, they tend to contain face(s) or out-standing man-made object(s), or tend to be more closed or more likely to be an indoor scene than images whose salient objects disagree with the whole images.**

3) The face-noface column has higher DR value than any other column given a specific agreement threshold, indicating that **images containing faces are highly likely to express the same sentiment as the faces express**, while images without faces are more unlikely to invoke human emotion solely due to the salient objects. This observation follows the intuition that faces tend to dominate the sentiment perception by humans. For example, we can easily tell the sentiment of the left images in Figure 2 simply by looking at the human face(s) without paying attention to other objects or elements.

Fig. 2. Red boxes indicate detected salient objects. Left column shows example images whose salient objects *agree* with the whole images. Right column shows example images whose salient objects *disagree* with the whole images. Text under each image indicates its *dominant attribute*.

We conduct another experiment to further evaluate the partitioning power of attribute combinations of closed, manmade, indoor and face, which individually has a great tendency of partitioning the images to agree. In this experiment, we partition the *agree* and *disagree* classes according to all the possible combinations of these four pairs of attributes. For example, for each class, we compute the total number of closed, manmade, indoor and face images, and also the total number of its entire opposite, namely open, natural, outdoor and noface images. For each combination in each partition, we will also calculate its DR metric.

The observations from Table 2 lead to the following interesting findings:

1) The combination with the most positive DR is 0, 0, 1 and 1, while the one with the most negative DR is exactly its opposite. That is to say, an image with an attribute combination of closed, manmade, indoor and face is the most likely to be classified to the partition of agree, verifying our aforementioned hypothesis.

2) Combinations with the face attribute are all with high positive DR, which further confirms that images with faces will be more likely to be classified into the partition of agree.

Table 2. Discrimination Ratio of all pairwise combinations of four attributes in the partition of agree

open /closed	natural /manmade	indoor /outdoor	face /noface	DR
0	0	0	0	-0.48
1	0	0	0	-16.91
0	1	0	0	+0.01
0	0	1	0	+2.32
0	0	0	1	+4.22
1	**1**	**0**	**0**	**-34.49**
1	0	1	0	+1.12
1	0	0	1	+8.87
0	1	1	0	+0.27
0	1	0	1	+2.69
0	**0**	**1**	**1**	**+38.49**
1	1	1	0	-0.01
1	1	0	1	+11.67
1	0	1	1	+1.21
0	1	1	1	+0.91
1	1	1	1	+0.16

[a] Threshold $\theta = 0.15$.

[b] For each cell, 1 stands for the first attribute of the pair and 0 stands for the other one.

5. CONCLUSION AND FUTURE WORK

In this work, we obtain salient object proposals and analyze their sentiment agreement with the whole images. We extract scene attributes from the images and examine the fine-grained interaction between visual saliency and visual sentiment. Our results suggest that images that contain outstanding man-made objects or human faces, or are indoors and closed, tend to express sentiment through their salient objects. On the other hand, images in which natural objects are more outstanding than man-made objects or do not contain human faces, or are outdoors and open, usually do not convey their sentiment information solely through their salient objects.

We will conduct further experiments on the scene attribute distribution for each emotion class to study the evocation mechanism for each specific emotion. Furthermore, we may also explicitly take salient maps into account in image sentiment classification. We hope ultimately this line of research will provide guidance for both in-depth image sentiment analysis and psycho-visual understanding.

6. REFERENCES

[1] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[2] S. Jindal and S. Singh, "Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning," in *2015 International Conference on Information Processing (ICIP)*, Dec 2015, pp. 447–451.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[5] Chenlei Guo, Qi Ma, and Liming Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *Computer vision and pattern recognition, 2008. cvpr 2008. ieee conference on*. IEEE, 2008, pp. 1–8.

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

[7] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang, "Deep networks for saliency detection via local estimation and global search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3183–3192.

[8] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang, "Saliency detection by multi-context deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1265–1274.

[9] Dingwen Zhang, Junwei Han, Chao Li, and Jingdong Wang, "Co-saliency detection via looking deep and wide," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2994–3002.

[10] Patrik Vuilleumier, Jorge L Armony, Jon Driver, and Raymond J Dolan, "Effects of attention and emotion on face processing in the human brain: an event-related fmri study," *Neuron*, vol. 30, no. 3, pp. 829–841, 2001.

[11] Stephen Grossberg, "How does a brain build a cognitive code?," in *Studies of mind and brain*, pp. 1–52. Springer, 1982.

[12] Elaine Fox, Riccardo Russo, Robert Bowles, and Kevin Dutton, "Do threatening stimuli draw or hold visual attention in subclinical anxiety?," *Journal of Experimental Psychology: General*, vol. 130, no. 4, pp. 681, 2001.

[13] Ming Sun, Jufeng Yang, Kai Wang, and Hui Shen, "Discovering affective regions in deep convolutional neural networks for visual sentiment prediction," in *Multimedia and Expo (ICME), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–6.

[14] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari, "What is an object?," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 73–80.

[15] Kuan-Chuan Peng, Amir Sadovnik, Andrew Gallagher, and Tsuhan Chen, "Where do emotions come from? predicting the emotion stimuli map," pp. 614–618, 2016.

[16] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomır Mech, "Unconstrained salient object detection via proposal subset optimization," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] Victor Campos, Brendan Jou, and Xavier Giro-i Nieto, "From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction," *arXiv preprint arXiv:1604.03489*, 2016.

[18] Aude Oliva and Antonio Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[19] Genevieve Patterson and James Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2751–2758.

[20] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 308–314.

[21] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495.

[22] Megvii Inc, "Face++ research toolkit. www.faceplusplus.com," 2013.