Visual Sentiment Prediction based on Automatic Discovery of Affective Regions

Jufeng Yang, Dongyu She, Ming Sun, Ming-Ming Cheng, Paul L. Rosin and Liang Wang

***Abstract*—Automatic assessment of sentiment from visual content has gained considerable attention with the increasing tendency of expressing opinions via images and videos online. This paper investigates the problem of visual sentiment analysis, which involves a high-level abstraction in the recognition process. While most of the current methods focus on improving holistic representations, we aim to utilize the local information, which is inspired by the observation that both the whole image and local regions convey significant sentiment information. We propose a framework to leverage affective regions, where we first use an off-the-shelf objectness tool to generate the candidates, and employ a candidate selection method to remove redundant and noisy proposals. Then a convolutional neural network (CNN) is connected with each candidate to compute the sentiment scores, and the affective regions are automatically discovered, taking the objectness score as well as the sentiment score into consideration. Finally, the CNN outputs from local regions are aggregated with the whole images to produce the final predictions. Our framework only requires image-level labels, thereby significantly reducing the annotation burden otherwise required for training. This is especially important for sentiment analysis as sentiment can be abstract, and labeling affective regions is too subjective and labor-consuming. Extensive experiments show that the proposed algorithm outperforms the state-of-the-art approaches on eight popular benchmark datasets.**

***Index Terms*—Visual sentiment analysis, sentiment classification, affective region, convolutional neural networks**

I. Introduction

With the increasing popularity of social networks, more and more Internet users tend to express their opinions with different media types [1]. Algorithms to identify sentiment can be helpful to understand such user behaviors [2]. In particular, understanding the sentiment in visual media content (*i.e.*, images, videos) has attracted increasing research attention. Potential use of approaches developed for visual sentiment analysis is broad, including affective image retrieval [3], aesthetic quality categorization [4], opinion mining [5], comment assistant [6], *etc*.

Inspired by psychology and the principles of art, researchers have investigated different groups of hand-crafted features (*e.g.*, color [7], [8], texture [9], [10], shape [11]) from image level, with the goal of endowing computers with the capability of perceiving sentiment in the same manner as humans. Instead of designing visual features manually, Convolutional Neural Network (CNN) can automatically learn deep representations of images [12]. Several researchers have also applied CNN to image sentiment classification [13]–[16] and demostrated the superior performance of the deep features against hand-tuned features for sentiment classification.

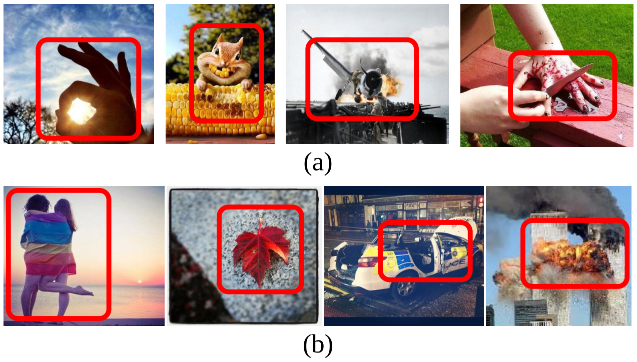


Fig. 1. Images from popular affective datasets: (a) Twitter I [14] and (b) Twitter II [17]. The bounding boxes indicate the local Affective Regions labeled by users. As can be seen, sentiments are evoked by the affective regions as well as the whole image appearance.

Visual sentiment analysis is inherently more challenging than traditional recognition tasks, since it involves a much higher level of abstraction and subjectivity in the human recognition process [18]. Recognizing sentiments evoked by images from social media is more difficult than many other visual recognition tasks, *e.g.*, object classification [19], scene recognition [20], *etc*. It is necessary to take a rich set of cues into consideration for visual sentiment prediction. Most existing methods employing CNNs try to learn sentiment representations from the global perspective of whole images, whereas the visual sentiment can also be evoked from the local regions within images [21]–[23]. Different from detecting concrete visual objects [24], there are difficulties modeling the sentiment due to the "*affective gap*" between the low-level visual features and high-level sentiment [10].

Little work has paid close attention to the use of local information for sentiment analysis. Li *et al.* [23] propose a context-aware classification model based on a bilayer sparse representation that simultaneously takes the local and global context into account. However, this approach is limited by its heavy dependence on the initial segmentation results to model appearances of different objects. In addition, they suppose that

J. Yang, D. She, M. Sun and M.-M. Cheng are at the School of Computer Science and Control Engineering, Nankai University, Tianjin 300350, China (e-mail: yangjufeng@nankai.edu.cn; sherry6656@163.com; msunming@foxmail.com; cmm@nankai.edu.cn).

P.L. Rosin is at the School of Computer Science and Informatics, Cardiff University, Wales, UK. (e-mail: Paul.Rosin@cs.cf.ac.uk).

L. Wang is at the National Laboratory of Pattern Recognition, CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wangliang@nlpr.ia.ac.cn).

all regions have the same weights for sentiment prediction, which may go against the human attention theories that the human vision system selectively processes parts of an image in detail [25]. You *et al.* [22] try to match local image regions with the descriptive visual attributes, aiming to discover the specific-attribute regions, but lack generalization ability for sentiment analysis.

To address these problems, we propose to leverage local details as well as the global information for visual sentiment analysis. We introduce a new notion named Affective Regions (ARs), which contains two distinguishing characteristics:

1) an AR is a salient region and probably contains one or more objects, which can attract people's attention, and
2) an AR conveys significant sentiments.

Fig. 1 shows some ARs in popular datasets [14], [17]. As can be seen, the visual sentiment can be induced from the ARs within images. For example, in the fourth image of (a), the sentiment is mostly evoked from the region of the bleeding hand, while in the second image of (b), the beautiful leaf rather than the gray stone conveys the positive sentiment. However, manually labeling the ARs of images for training the detector is too subjective and labor-consuming. This paper proposes a framework that only requires the image-level label to discover AR automatically, thereby significantly reducing the annotation burden.

In detail, we first use an off-the-shelf tool to generate bounding box candidates along with their objectness score for the input image, which is inspired by the strong co-occurrence relationships between objects and sentiment [26]. Then a candidate selection method is employed to remove the redundant proposals while preserving several valuable ones. The deep CNN is connected with each candidate and used to compute sentiment score. The objectness score and sentiment score are combined to calculate the AR score, based on which the top-K ARs are discovered by re-ranking the candidate regions considering both the objectness score as well as the sentiment score. Finally, the CNN outputs from the global and local views are aggregated through alternative fusion operations (*i.e.*, max pooling, sum pooling and concatenation) to produce the final predictions.

Our contributions are summarized as follows:

- We propose a deep framework for automatically discovering the affective regions of images which are likely to evoke significant sentiment information. Our framework is independent of object categories and requires no bounding box annotation, which is more general than the existing methods.
- We build a visual sentiment prediction model using a deep CNN, which utilizes the holistic and local information from both the global image and the local regions. The final representation is effective for visual sentiment classification, and outperforms the state-of-the-art approaches on the affective datasets.
- Experimental results show that our proposed framework can be generalized to the small-scale benchmarks with the help of transfer learning.

This journal paper extends our earlier work [27] in four aspects. (1) The framework is improved by adding the candidate selection module to suppress the possibly noisy proposals and reduce computational load. (2) Three alternative fusion operations are employed to combine the holistic representation with the affective regions, which aim to capture the local information in different ways. (3) More implementation details are provided and extensive experimental results on both large-scale and small-scale datasets are presented, where the hyper-parameters are determined in a systematic way. (4) The consistency of the discovered affective regions and the ground truth is evaluated on the EmotionROI benchmark [21], showing that our proposed method can automatically find high-quality ARs without human annotations.

The rest of this paper is organized as follows. Sec. II summarizes the related work on visual sentiment analysis and deep learning. Sec. III introduces the proposed method of discovering affective regions and our deep framework for sentiment prediction. In Sec. IV and V, we present and visualize the experimental results on the popular benchmark datasets. And finally, Sec. VI concludes this paper.

II. Related Work

Numerous methods for visual sentiment analysis have been developed based on still images [10], [17] and videos [28], [29]. In this section, we review the methods for affective image prediction and region-based CNNs that are closely related to this work.

A. Affective Image Prediction

Previous methods on affective image prediction can be roughly divided into dimensional approaches and categorical ones. The dimensional approaches represent sentiment in the two dimensional (2-D) valence-arousal coordinate space [30] or a three dimensional space [31]. Hanjalic [32] represents human affective response using three basic dimensions, *i.e.*, valence, arousal and control (dominance), where there is a corresponding value for every affective state. Zhao *et al.* [33], [34] propose to predict the personalized emotion perceptions of images in the valence-arousal space using shared sparse regression as a learning model. Meanwhile, the categorical approaches map sentiment into one of the representative categories. There is also some work predicting the discrete probability of different sentiment categories [35]–[39]. Since categorical approaches make it easier for a human to understand, we target categorical sentiment prediction in this work.

1) Shallow modeling methods: Most previous methods on affective image prediction employ traditional low-level features. Machajdik *et al.* [10] define a combination of rich hand-crafted features based on art and psychology theory, including composition, color variance and image texture, *etc.* Lu *et al.* [11] investigate how shape features in natural images influence sentiments aroused in human beings, and provide evidence for the significance of roundness-angularity and simplicity-complexity for predicting sentiment content. Zhao *et al.* [8] introduce more robust and invariant visual

Dataset	GT-Box	Positive	Negative	Sum
IAPSa [43]	N	209	186	395
Abstract [10]	N	139	89	228
ArtPhoto [10]	N	378	428	806
Twitter I [14]	N	769	500	1,269
Twitter II [17]	N	463	133	596
EmotionROI [44]	Y	660	1,320	1,980
Flickr&Instagram [16]	N	16,430	6,878	23,308
Flickr [17]	N	435,798	48,424	484,222

Fig. 2. Statistics of the available affective datasets. Most datasets developed in this field contain no more than two thousand samples, mainly due to the subjective and labor intensive labeling process. Note that the Flickr dataset is weakly-labeled and none of these datasets except EmotionROI provide ground truth bounding box (GT-Box) corresponds to affective regions.

features designed according to art principles. These handcrafted visual features are proven to be effective on several small datasets, whose images are selected from a few specific domains, *e.g.*, abstract paintings and art photos [10].

To bridge the "*affective gap*" between low-level features and high-level sentiment, Borth *et al.* [17] model a mid-level concept, *i.e.*, Adjective Noun Pairs (ANPs), which are used to detect image concepts instead of expressing sentiments directly. Li *et al.* [40] further compute the weighted sum of the textual sentiment values of ANPs describing the image and take the textual sentiment into account. Yuan *et al.* [41] propose the Sentribute, an image-sentiment analysis algorithm based on 102 mid-level attributes, which are easier to interpret and ready to use for high-level understanding. Furthermore, Zhao *et al.* [42] combine features of different levels including low-level features from elements-of-art, mid-level features from principles-of-art and high-level features from a semantic concepts detector in a multi-graph learning framework. Chen *et al.* [26] build object detection models to recognize six frequent objects including car, dog, dress, face, flower and food, and propose a new classification model to handle attributive and proportional similarity between visual sentiment concepts. In contrast, our algorithm concentrates on whether a selected region contains objects or not, which is independent to object categories and more robust for real applications.

2) Deep modeling methods: In recent years, CNNs have been incorporated into a number of visual recognition systems in a wide variety of domains [45], [46]. The strength of these models lies in their ability to learn discriminative features from raw data inputs using the back propagation algorithm [47], in contrast to more traditional recognition pipelines which compute hand-engineered features on images as an initial preprocessing step [48].

Several recent methods exploit deep CNNs for image sentiment prediction. Based on their previous work [17], Chen *et al.* [49] adapt deep networks for constructing DeepSentiBank, a classification model for visual sentiment concepts, which shows significant improvements in both annotation accuracy and retrieval performance. Also, some methods incorporate the model weights learned from a large-scale general dataset [50], and further fine-tune the CNNs for the task of visual sentiment prediction [13], [15]. In [13], two types of activations from CNNs are used as image-level features for classification, namely the 4096-dimensional output from fc7 and the 1000-dimensional output from fc8. You *et al.* [14] employ a progressive strategy to train a CNN making use of half a million images that are labeled with the website meta data, and further perform benchmarking analysis on the Flickr and Instagram (FI) dataset. In [22], a method based on the attention model is developed in which local visual regions induced by sentiment related visual attributes are considered.

Due to the expensive manual annotation of sentiment labels, the existing affective datasets, including IAPSa [43], ArtPhoto [10], Abstract Paintings [10], Twitter I [14], Twitter II [17] and EmotionROI [44] typically contain less than two thousand images (see also Fig. 2). This is far from the required scale for training robust deep models. The Flickr dataset [17] is weakly-labeled with 2 categories using the meta-data provided by the up-loaders. Moreover, only the EmotionROI dataset has provided ground truth affective regions. Note that in this paper we focus on the binary sentiment (*i.e.*, positive and negative) prediction problem, for which a variety of benchmark datasets with reliable ground-truth can be employed to validate the effectiveness of the proposed algorithm.

B. Region-based CNNs

We trace the roots of our approach to region-based CNN (R-CNN) [46], an algorithm applying deep CNN to bottom-up generate region proposals in order to localize and segment objects. It has been proved that when labeled training data are scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, boosts performance significantly [51]. Girshick [52] shows that it is possible to further reduce training and testing time, while improving detection accuracy and simplifying the training process, using an approach called Fast R-CNN. Fast R-CNN reduces detection time excluding region proposal computation to 50–300ms per image, depending on network architecture. Ren *et al.* [24] introduce a fully-convolutional network version that simultaneously predicts object bounds and objectness scores at each position. Meanwhile, R-CNN has been applied to various tasks, *e.g.*, pedestrian detection [53], action detection [54], [55] and semantic segmentation [56].

Different from the traditional methods on region based CNNs for finding salient objects in an image, our work aims to automatically identify the ARs that evoke sentiment and use the local information as the supplementary sentiment representation. This requires us to analyze not only the regions containing objects but also the surrounding background [21], which may have affective influence on the selected regions. Moreover, R-CNN based methods require ground truth bounding box annotations for training, but it is time- and labor-consuming to label affective regions manually. In this paper, we employ an off-the-shelf tool to generate object proposals as candidate affective regions and propose to select the AR considering the low-level as well as the affective-level content. Compared with the methods requiring accurate segmentation [23] or concrete category information [26], it is much easier to acquire object proposals in the preprocessing stage, and will better generalize to other datasets.

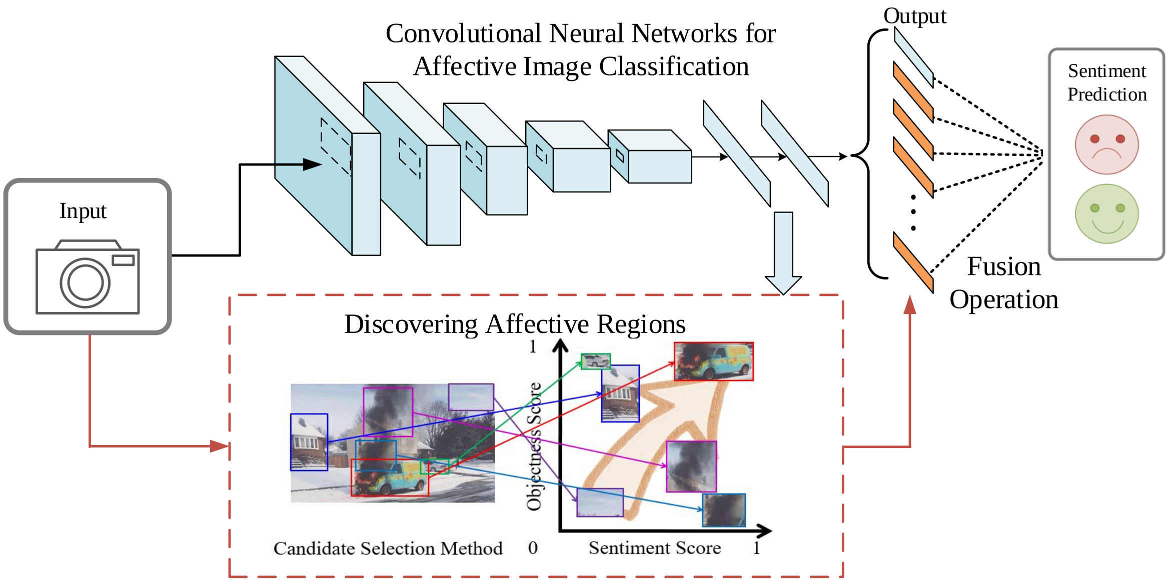

Convolutional Neural Networks for
Affective Image Classification
Output
Sentiment
Prediction
Input
Fusion
Operation
Discovering Affective Regions
1
Objectness Score
Candidate Selection Method
0
Sentiment Score
1

Fig. 3. Pipeline of the proposed approach. Given the input image, thousands of candidates along with the objectness scores are generated, and the candidate selection method is applied to remove the candidates which are overlapped and less important. The sentiment score of each proposal is roughly computed through CNN, which is then combined with the objectness score to discover affective regions. Finally, the sentiment label is predicted by fusing the local information with the holistic representation using several alternative operations.

III. Methodology

In this section, we aim to develop an algorithm to automatically discover ARs carrying significant sentiments and combine the standard holistic representation with a local representation for image sentiment analysis. Fig. 3 shows the pipeline of our proposed framework. We use an object detection technique, *i.e.*, Edgeboxes [57], to produce the candidate windows guiding the search for ARs, and then apply the candidate selection method to reduce redundant and noisy proposals. Thus, the sentiment content of each proposal is estimated at both the low-level and affective-level for the ARs detection. Finally, the deep representation of the detected ARs is combined with the holistic representation through three alternative fusion strategies, *i.e.*, max pooling, sum pooling and concatenation, to generate the final predictions.

A. Producing Candidate Proposals

1) Generating: Detecting concrete visual objects like dogs and cars has been researched extensively in computer vision [56], [58]. However, modeling abstract emotional concepts like amusement and excitement is very challenging. The difficulty comes from the “affective gap” that lies between low-level visual features and high-level sentiment. Previous methods [17], [26] have proved that associating adjectives with concrete objects can make the combined visual concepts more detectable and tractable for visual sentiment analysis. Inspired by the strong co-occurrence relationships between objects and sentiment, we suggest that object proposals can be used as the potential sentiment regions.

Since our framework takes the object proposals as inputs and obtains the final prediction by fusing the prediction of each affective region with the holistic representation, the performance of the proposed framework largely depends on the quality of the candidate regions. However, an effective candidate extraction approach is challenging since the affective region detection needs to capture not only objects but also regions of the background that may evoke sentiment. There are two criteria that should be satisfied. First, the proposed framework is based on the assumption that the candidate proposals can cover the objects in the affective images as well as parts of the background, which requires a high detection recall rate. Second, since the selected affective region proposals are then fed into the CNN, only a limited number of candidates should be produced so as to allow for efficiency whilst maintaining accuracy.

During the past decades, many object proposal methods have been proposed to tackle the object detection problem. According to [59], [60], EdgeBoxes [57] and BING [61] are faster than methods such as Selective Search [62] and Objectness [63], while EdgeBoxes achieves better quality of proposals compared to BING. Considering the balance between the speed and quality, this paper uses EdgeBoxes to generate a set of candidate windows as it provides the best trade-off. Such an off-the-shelf tool can generate thousands of candidate boxes in a fraction of a second, from which a subsequent refinement step based on object boundary estimates is applied to improve localization. For a given image I, a set of candidate bounding boxes with objectness score $B = \{b_i; Obj\_score_i^I\}_{i=1}^n$ is produced by EdgeBoxes.

2) Selecting and filtering: To achieve high recall for object detection, Zitnick *et al.* [57] employ a bottom-up strategy, generating thousands of proposals in each image. However, most of the candidate proposals are heavily overlapped and redundant for predicting sentiment. It is necessary to filter out the noisy region proposals carrying little sentiment, and removing noisy proposals at the initial stage of the algorithm can greatly reduce the computation time of the subsequent steps. To address this problem, we introduce the candidate selection module to select proposals from the affective region candidates inspired by [64].

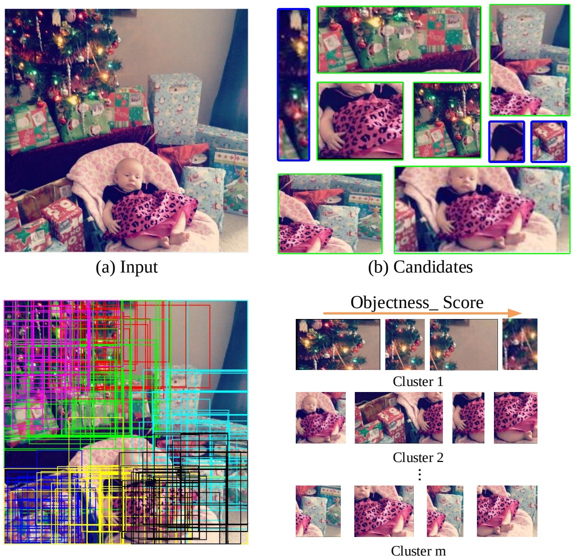
EdgeBoxes generates thousands of proposals in each image that can achieve high recall for object detection. However, since it still generates a large number of original object windows for the CNN to process, following [65], we first check the same geometric characteristics (*i.e.*, the area and height/width ratio) of candidate bounding boxes. We empirically filter out the regions with small areas ($<$ 800 pixels) or with high height/width (or width/height) aspect ratios above a threshold (> 6), since objects which are either too small or too long are unlikely to attract people's attention. Thus, a much smaller number of proposals can be fed into the candidate selection method. Following previous algorithms [64], [66], we build the affinity matrix $W \in \mathbb{R}^{n \times n}$ for each image, in which each element denotes the intersection-over-union (IoU) scores between any pair of the bounding boxes and n denotes the number of candidates:

$$W_{ij} = \frac{|b_i \cap b_j|}{|b_i \cup b_j|}, \tag{1}$$

where $|\cdot|$ is used to measure the numbers of pixels. We then apply the normalized cut algorithm [67] to group the candidate bounding boxes into m clusters. In detail, the normalized graph Laplacian matrix $L = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}$ is computed where $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n} W_{ij}$. Then the eigenvector matrix $V = [v_1, \cdots, v_m] \in \mathbb{R}^{n \times m}$ is constructed where $\{v_1, \cdots, v_m\}$ are the m smallest eigenvectors of L. Finally the *k-means* clustering algorithm is used to obtain m cluster labels where each row of V is the feature of the corresponding sample [67]. As shown in Fig. 4, the bounding boxes are first filtered out to reduce the computational load. Then with the m clusters' bounding boxes, we pick the proposal with the highest objectness score in each cluster and generate m candidate regions $H = \{h_i\}_{i=1}^{m}$ for each image. Compared to the greedy non-maximum suppression (NMS) method that is widely used for filtering [46], our candidate selection method can generate a specific number of proposals while removing the redundant and noisy bounding boxes.

B. Discovering Affective Regions

1) Initializing the framework: CNNs achieve the state-of-the-art performance in the related computer vision tasks, *e.g.*, aesthetic quality rating [4] and image style recognition [68] by fine-tuning the pre-trained ImageNet model. In this work, the CNN is based on the deep model VGGNet [69] with 16 layers. In order to adapt the pre-trained model on ImageNet for sentiment analysis, the CNN is first fine-tuned on the target affective dataset (*e.g.*, Flickr and Instagram) utilizing the original images (without any bounding boxes) to adjust the parameters of the deep model. As a supervised learning approach, the fine-tuned CNN is applied to learn a function $f : \mathcal{I} \rightarrow \mathcal{L}$, from a collection of affective training examples $\{(I_i, l_i)\}_{i=1}^{N}$, where N is the size of the training set, I_i is the input image, and l_i is the associated sentiment label. In the standard training process, the traditional classification loss is optimized to maximize the probability of the correct class [45], [69]. Let $\mathbf{d}_i$ be the output from the penultimate layer, then the fine-tuning of the last layer is done by minimizing the softmax loss function as follows:


(a) Input
(b) Candidates

(c) Candidates Selection

Fig. 4. Given the input image (a), the candidate windows are generated by EdgeBoxes and small or high aspect ratio boxes are filtered out. As shown in (b), the proposals with blue bounding boxes are dropped in this step. Different colors in (c) indicate the different clusters produced by normalized cut, from which the representative proposals are selected.

$$l(\mathbf{W}) = \sum_{i=1}^{N} \sum_{j \in l} \mathbf{1}(l_i = j) \log p(l_i = j | \mathbf{d}_i, \mathbf{w}_j), \tag{2}$$

where $\mathbf{W} = \{\mathbf{w}_j\}_{j \epsilon l}$ is the set of model parameters, and the indicator function $\mathbf{1}(s) = 1$ if s is true, otherwise $\mathbf{1}(s) = 0$. The probability of each sentiment label $p(l_i = j|\mathbf{d}_i, \mathbf{w}_j)$ can be defined by the softmax function:

$$p(l_i = j|\mathbf{d}_i, \mathbf{w}_j) = \frac{\exp(\mathbf{w}_j^T \mathbf{d}_i)}{\sum_{j' \epsilon l} \exp(\mathbf{w}_{j'}^T \mathbf{d}_i)} \tag{3}$$

Since the number of categories in the affective dataset is not equal to that of ImageNet, the fc8 classification layer is changed to 2-way required by the sentiment dataset, which can produce a probability prediction over the sentiment classes.

2) Estimating sentiment score: For the affective-level quality of the candidate proposals, we compute the sentiment scores by feeding the proposal to the CNN. For the generated affective candidates $H = \{h_i\}_{i=1}^{m}$ of the input image I, let $\{y_{ij}\}_{j=1}^{c}$ be the output vector of the last layer indicating the probability of the i-th proposal carrying the j-th class sentiment, and c is set to 2 as the number of sentiment classes. If the prediction values for each sentiment are similar then this usually indicates that it is difficult to distinguish the sentiments evoked by the proposal. Therefore, we aim to keep only those

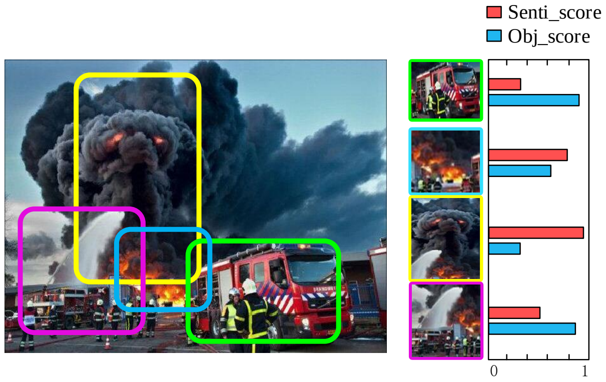


Fig. 5. Visualization of the objectness score and sentiment score in an image. For example, the region with a high objectness score indicates that the corresponding bounding box is extremely likely to be an object.

proposals which contain a dominant sentiment. We define a probabilistic sampling function to evaluate the sentiment score of the i-th region in an affective-level perspective as follows:

$$Senti\_score_i^I = \sum_{j=1}^{c} y_{ij} * \log y_{ij} + 1, \tag{4}$$

where the score ranges between 0 and 1 for binary classification. The information entropy defined in Eqn. (4) represents the degree of uncertainty when predicting sentiment, which is also consistent with the affective-level estimation of the proposal. The $Senti\_score_i^I$ can be high-level and provides a more semantic measurement compared to the traditional methods.

3) Selecting affective regions: We choose ARs according to two aspects: i) how likely the region contains an object, which is represented as $Obj\_score_i^I$, and ii) how much the region carries sentiment at the affective-level, referred to as $Senti\_score_i^I$. Fig. 5 demonstrates that an affective region should have both high $Obj\_score_i^I$ and $Senti\_score_i^I$. The reason is that the $Obj\_score_i^I$ only measures the probability of regions containing an object and is based on the texture appearance, which lacks the guidance of semantic information. The $Senti\_score_i^I$ reflects the sentiment of images at the affective level, which enables lots of noisy regions to be removed with little impact on the sentiment analysis. Such a score allows certain flexibility for the object regions, which may occur in the background as well. Considering the characteristics of each score, we introduce the AR_score to evaluate the sentiment quality of each region with the following definition:

$$AR\_score_i^I = \sqrt{(1-\alpha) * {Obj\_score_i^I}^2 + \alpha * {Senti\_score_i^I}^2}, \tag{5}$$

where α controls the trade-off between low-level and affective-level perspectives. In this paper, we select α by the cross validation of the large-scale affective dataset. The proposals with high AR_score are considered be an AR and used for sentiment prediction, while the proposals with low AR_score are removed from the candidate set.

Algorithm 1 Visual Sentiment Analysis using Affective Regions

```
Input:
    Input Image: $I$
    The number of desired affective regions: $K$
Output:
    Predicted sentiment label : $\vec{Y}$
Generate $n$ bounding boxes with their objectness scores $B = \{b_i; Obj\_score_i^I\}_{i=1}^{n}$.
Apply candidate selection method to generate $m$ candidate regions $H = \{h_i\}_{i=1}^{m}$.
Initialize the framework with pre-trained CNN.
Let $\vec{Y}_{Global}$ be the predictions of the whole image.
Pass $H$ through the CNN model from the second layer to the last layer.
Let $y \in \mathbb{R}^{m \times c}$ be the sentiment probability of $m$ proposal using the CNN model, compute the sentiment score in Eqn. (4)
Compute the AR score for the each region in Eqn. (5).
Rank proposals with AR scores and select top $K$ as affective regions.
Predict the label $\vec{Y}$ using the cross-candidates pooling operation.
return $\vec{Y}$
```

C. Sentiment Classification

Based on the initialized framework, the sentiment classification of a given image can be summarized as follows. Given a test image, we first generate the affective candidates based on EdgeBoxes. In order to reduce redundancy, we apply the candidate selection method based on their IoU scores and keep just the best candidates. Both objectness score and sentiment score are considered for selecting affective regions that are likely to attract people's attention and include emotional content. Then, for each proposal as well as the holistic image, a c-dimensional predictive result is obtained by the CNN, which is then fused into a final prediction. In particular, we consider three strategies, namely max pooling, sum pooling and concatenation. We utilize the cross-candidates pooling operation to fuse the outputs from the CNN into an integrated prediction. With max pooling, the high prediction scores from those candidates containing sentiment are preserved and the noisy ones are ignored. The sentiment probability $\vec{Y}$ of a given image can be defined as follows:

$$\vec{Y} = \max\left(\vec{Y}_{Global}, \{\vec{Y}_{AR_j}\}_{j=1}^{K}\right), \tag{6}$$

where $\vec{Y}_{Global}$ represents the prediction of the whole image and $\vec{Y}_{AR_j}$ represents the prediction of the j-th affective region, and we select the top K affective regions based on Eqn. (5). $\vec{Y}$, $\vec{Y}_{Global}$ and $\vec{Y}_{AR_j}$ share the same vector structure of (y_{pos}, y_{neg}), where y_{pos} and y_{neg} indicate the predicted probability of positive and negative sentiments, respectively.

The sum pooling fuses the prediction probability of all the proposals, where the weights of consistent proposals can be

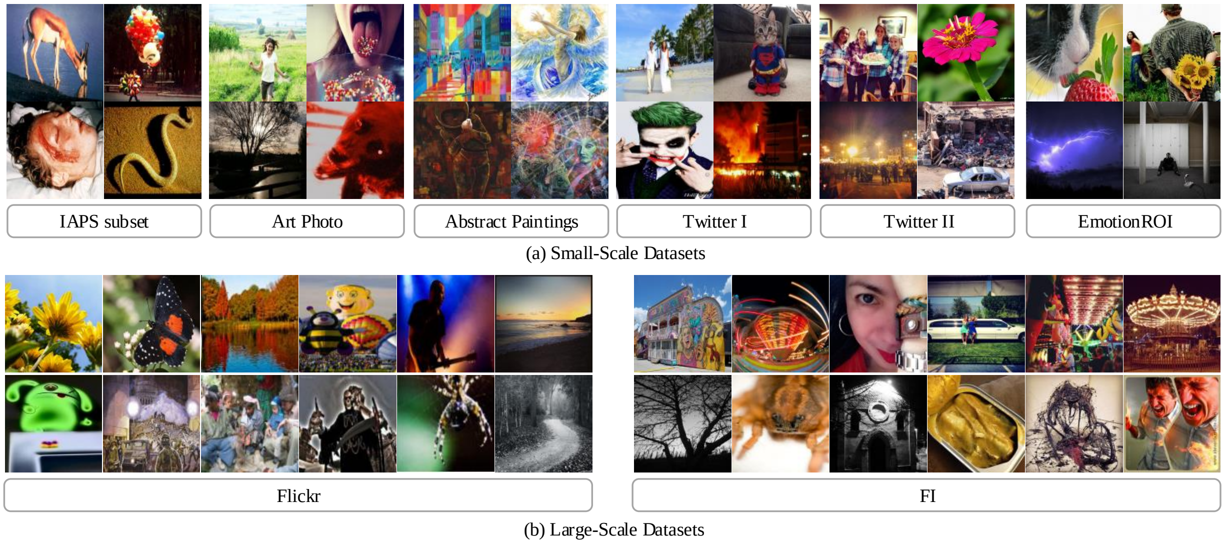


Fig. 6. Example images from (a) small-scale and (b) large-scale affective datasets. The images come from a variety of domains including art, real life, abstract and so on, in which the sentiment distributions are different.

emphasized.

$$\vec{Y} = (1-\beta) * \vec{Y}_{Global} + \beta * \frac{1}{K} * \sum_{j=1}^{K} \vec{Y}_{AR_j}, \tag{7}$$

where β is the trade-off between global and local prediction. The β is also estimated by cross-validation of the large-scale affective dataset. Both max pooling and sum pooling can generate the sentiment probability as the final prediction.

Concatenation is a simple but effective way by combing the features for a comprehensive representation:

$$\vec{Y} = \left[\vec{Y}_{Global}, \{\vec{Y}_{AR_j}\}_{j=1}^{K}\right]. \tag{8}$$

The final feature is generated by concatenating all the prediction results, and the dimension of $\vec{Y}$ is $(K+1) \times c$. In our experiments, we set the number of affective regions in all samples to be the same, making it feasible to classify the concatenated feature vector using an SVM.

IV. Experimental Results

In this section, we present our experiments and evaluate our method against the state-of-the-art deep methods to validate the effectiveness of our framework for sentiment classification and sentiment detection.

A. Dataset

We evaluate our proposed method on eight widely-used datasets, including IAPSa [43], ArtPhoto [10], Abstract Paintings [10], Twitter I [14], Twitter II [17], EmotionROI [21], Flickr [17] and Flickr and Instagram (FI) [16]. We divide the datasets into small-scale and large-scale datasets with respect to the number of images, as shown in Fig. 6.

1) Small-scale datasets: The International Affective Picture System (IAPS) [70] is a common stimulus dataset which is widely used in visual sentiment analysis research [8]–[10], [71]. **IAPSa** selects 395 pictures from IAPS and is labeled with Mikel's eight sentiment categories. **ArtPhoto** contains 806 artistic photographs from a photo sharing site and the ground truth labeling is provided by the owner of each image. **Abstract Paintings** contains 228 peer rated abstract paintings consisting of color and texture. **Twitter I** is collected from social websites and labeled with two categories (*i.e*., *positive, negative*) by Amazon Mechanical Turk (AMT) workers, and contains 1,269 images in total. We test our method on all of the three subsets of Twitter I, including "Five agree", "At least four agree" and "At least three agree", in a similar fashion to [14]. "Five agree" indicates that all the five AMT workers give the same sentiment label for a given image. **Twitter II** contains 603 images from the Twitter website, and the ground truths are obtained by AMT annotation too, resulting in 470 positive and 133 negative labels. **EmotionROI** is created as a sentiment prediction benchmark, and is collected from Flickr resulting in 1,980 images with six sentiment categories. They use AMT to collect 15 responses to the regions that evoke sentiment and represent the ground truth by assuming the influence of each pixel on evoked sentiments is proportional to the number of drawn rectangles covering that pixel.

2) Large-scale datasets: **FI** is currently the largest well-labeled dataset, which is collected by querying with eight sentiment categories as keywords from social websites. 225 AMT workers were employed to label the images which resulted in 23,308 images receiving at least three agreements. We divide FI into binary datasets the same as the IAPSa. **Flickr** contains 484,258 images in total, where each image was automatically labeled using the corresponding ANP.

Since we focus on the binary sentiment prediction, we

	Methods	FI	Flickr
Baseline	AlexNet [45]	60.54	55.13
	VGGNet [69]	70.64	61.28
	Fine-tuned AlexNet	72.43	61.85
	Fine-tuned VGGNet	83.05	70.12
	PCNN (VGGNet) [14]	75.34	70.48
	DeepSentiBank [49]	61.54	57.83
Ours	obj + concatenation	83.85	70.05
	senti + concatenation	84.07	70.10
	AR + concatenation	84.83	70.51
	AR + sum-pooling	84.50	70.46
	AR + max-pooling	84.21	70.49
	AR + concatenation ($K = 8$)	**86.35**	**71.13**

Fig. 7. Classification accuracy (%) on the test set of the large scale dataset, *i.e.*, FI and Flickr. We compare our proposed method with different deep methods including ImageNet models (row 1-2), fine-tuned models (row 3-4), and state-of-the-art algorithms (row 5-6). Our proposed method with different configurations are also given, *i.e.*, combining with the top-1 region (row 7-11), and leveraging more Affective Regions(row 12). Note that obj/senti indicate that only objectness score/sentiment score is used, while our "AR" method selects Affective Regions, where both objectness score and sentiment score are considered.

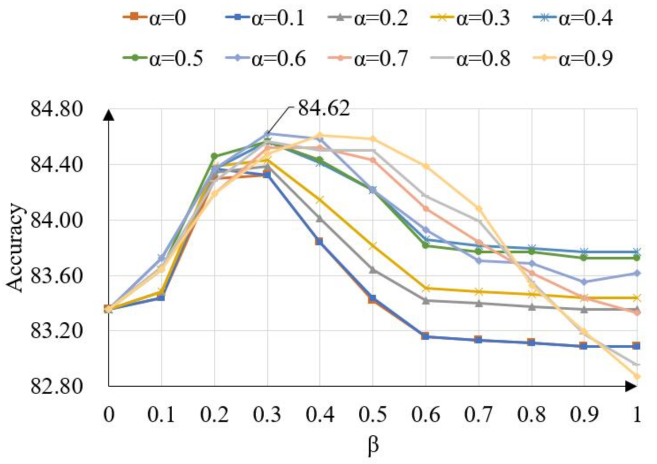


Fig. 8. Impact of different α and β on the validation sets of the FI dataset. We choose $\alpha = 0.6, \beta = 0.3$ in the remaining experiments.

convert the multi-sentiment labels into positive and negative ones according to their valance for datasets except for Twitter I, Twitter II and Flickr, which were originally labeled with binary sentiment. Specifically, for IAPSa, ArtPhoto, Abstract Paintings, and FI, we divide Mikel's eight sentiment categories into binary labels according to [43], which suggests that amusement, awe, contentment and excitement are positive sentiments and anger, disgust, fear and sadness are negative sentiments. EmotionROI is labeled with seven sentiments (*i.e.*, anger, disgust, fear, joy, sadness, surprise, neutral) along with Valance-Arousal scores, where anger, disgust, fear, sadness can similarly be considered as the negative sentiments. Since the mean valance of the set of joy and surprise images is higher than the mean valance of the set of negative images, we treat them as positive sentiment. Note that we do not include images with neutral sentiment in the experiment.

B. Implementation Details

CNNs have the capability to incorporate model weights learned from a more general dataset, which is a convenient property for tasks lacking sufficient training data. We employ the VGGNet with 16 layers [69] as our basic architecture. Following previous works [13], we initialize our model with the weights trained from ImageNet. Then the pre-trained network is fine-tuned on the large-scale datasets with the 1000-way fc8 classification layer replaced by the 2-way layer, and the data are split randomly into 80% training, 5% validation and 15% testing sets. The learning rates of the convolutional layers and the last fully-connected layer are initialized as 0.001 and 0.01 respectively. We fine-tune all layers by stochastic gradient descent through the whole net using a batch size of 64. A total of 100,000 iterations is run to update the parameters to extract more precise sentiment-related information. All our experiments are carried out on two NVIDIA GTX 1080 GPUs with 32 GB of CPU memory. For the candidate selection method, we set $m = 50$ for each image as the experimental trade-off between the performance and computational time, which provides the initial candidate proposals for discovering the affective regions.

With the help of transfer learning, we also employ our framework on small-scale datasets with limited training examples. In detail, we use the parameters of the CNN trained on FI on other datasets and fine-tune the model on the training set of other datasets. The small datasets are randomly split into 80% training and 20% testing sets except those with a specified training/testing split [17], [44] and we conduct the experiments using 5-fold cross validation and average the accuracies as the final results.

C. Baseline

In the following subsections, we evaluate the proposed method against the state-of-the-art algorithms for image sentiment prediction, including those based on hand-crafted features and deep methods. In addition, we also show the results with different configurations of the proposed method on the validation set, especially with different components and fusion strategies.

1) Hand-crafted features: We extract several low-level features from the small-scale datasets, including local descriptors like SIFT, HOG, GIST, *etc*. The global color histograms (**GCH**) features consists of 64-bin RGB histogram, while the local color histogram features (**LCH**) first divide the image into 16 blocks and use a 64-bin RGB histogram for each block [72]. We use the **ColorName** to count the pixels of each of the 11 basic colors presented on the image using the algorithm in [10]. We also use **SentiBank** [17], a concept detector library based on the constructed ontology, to exploits the 1,200 dimensional features as mid-level representation. Zhao *et al.* [8] propose the principle of art features (**PAEF**) for sentiment analysis. We use a simplified version provided by the author to extract 27 dimension features.

2) Deep methods: **PCNN** proposed by You *et al.* [14] is a novel progressive CNN architecture. They suggest that leveraging larger amounts of weakly supervised data can

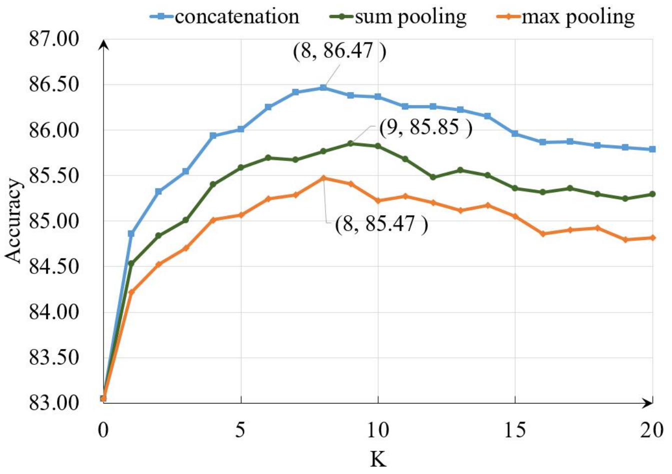


Fig. 9. Impact of different K on the validation set of the FI dataset. We set $K = 8$ in the remaining experiments.

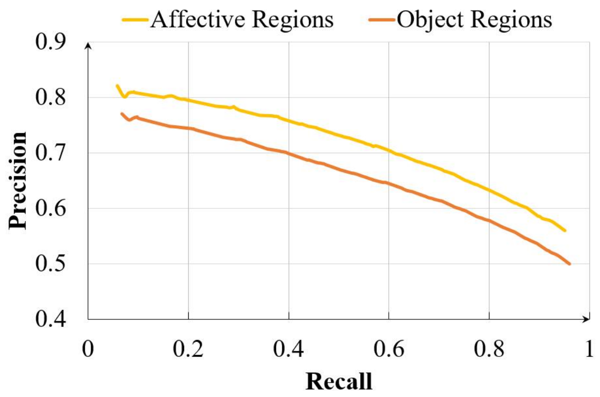


Fig. 10. Precision-recall curve for discovering affective regions. Our method is more consistent with human annotation than objectness (*i.e.*, using the proposals with the highest objectness scores generated by EdgeBoxes).

improve the generalizability of the model. We fine-tune the PCNN with the noisy Flickr dataset based on VGGNet and extract the deep visual features. **DeepSentiBank** [49] is a visual sentiment concept classification based on CNNs for discovering ANPs. We apply the pre-trained DeepSentiBank to extract 2,089 ANPs as mid-level representations for sentiment. We also show the performance of deep visual features of **CNN** models pre-trained on ImageNet and fine-tuned on the affective datasets, including different architectures, *i.e.*, AlexNet and VGGNet. To compare with the ImageNet CNN, we show the results of using LIBSVM [73] trained on features extracted from the second to the last layer of the model and reduce the dimensionality employing PCA. In practice, we find that different cost values (parameter C in LIBSVM) produce similar accuracy, so we just use the default value and use the *one v.s. all* strategy following the same evaluation routine described in [10].

D. Results on Large-Scale Datasets

We first fine-tune the CNN on the large scale datasets (*i.e.*, FI and Flickr), and compare the performance of our framework with the deep methods. Fig. 7 reports the performance of the baselines on the test set of the FI and Flickr datasets. As can be seen, the pre-trained model on the ImageNet is inferior to the fine-tuned model due to the differences between the distributions in the ImageNet and sentiment datasets, while VGGNet with a deeper architecture performs better than AlexNet. The fine-tuned VGG achieves 83.05% on the FI dataset, which outperforms DeepSentiBank (61.54%) and PCNN (75.34%). Compared to the weakly-labeled Flickr, the fine-tuned CNN on FI shows a greater improvement in performance due to the reliable annotation.

When selecting and combining affective regions in the deep model, we have several choices: we can use the objectness score or sentiment score only, or use the AR score proposed in this work. We roughly consider the objectness score as a low-level cue and sentiment score as a high-level cue. The experimental results show that the sentiment score is more effective than the objectness score, which is mainly because the objectness score just indicates how likely a region contains an object. When both scores are combined into a deep model, our method using the most confident affective regions achieves 84.83%, which performs favorably against the state-of-the-art methods as well as combing the proposals selected by only one score, demonstrating the benefit of using local details for classification. Analyzing the objectness and sentiment score of different regions, we observe that the sentiment score often gives different values even when the area of overlap of two different proposals is more than half. For two different regions proposals both containing an affective region, the sentiment scores are usually similar, and thus it only needs to evaluate whether the proposal contains an affective region and ignore the area of proposal.

1) The effect of the hyper-parameters: We report the classification performance of using “AR + sum pooling” methods on the validation set of the FI dataset, and different α and β are employed for comparison. As shown in Fig. 8, setting $\alpha = 0.6$ achieves the best overall accuracy for discovering affective regions in the validation set. Using only the objectness score ($\alpha = 0$) gives limited performance, which indicates it is necessary to use the sentiment score for selecting affective regions. On the other hand, combining local regions can boost the classification performance compared with using a single global representation. Setting $\beta = 0.3$ achieves a balance in most cases. Therefore, we use $\alpha = 0.6, \beta = 0.3$ in the remaining experiments.

2) The effect of the fusion operations: When fusing the outputs of the affective region and the entire image, we consider three fusion operations for combing the most confident affective regions. Fig. 7 (bottom) shows that all three combinations are useful for capturing information in the holistic and regional view, while concatenation is the most effective way since it retains all the information.

3) The effect of the hyper-parameter K: Given an input image, we not only predict the sentiment of the whole image but also find the affective regions. Although the dataset does not provide annotations of affective regions, the number of affective regions is usually small. Here we show an experiment to determine how many affective regions should be evoked in

Algorithm	IAPS-Subset	Abstract	ArtPhoto	Twitter I			Twitter II	EmotionROI
				Twitter I_5	Twitter I_4	Twitter I_3		
GCH	71.76	71.50	67.00	67.91	67.20	65.41	77.68	66.53
LCH	52.91	73.26	64.01	70.18	68.54	65.93	75.98	64.29
ColorName + BoW	57.72	73.28	66.26	64.51	64.79	60.83	70.10	60.13
Gist	65.05	60.97	63.40	65.87	61.47	60.68	77.68	60.38
LBP	56.73	59.85	55.06	55.78	53.94	57.29	65.15	55.26
Gabor	79.21	50.43	58.43	55.37	54.03	53.90	63.72	58.73
SIFT + BoW	86.06	53.54	59.05	63.15	63.71	60.36	70.32	65.30
SIFT + VLAD	83.02	60.53	64.75	70.29	68.91	67.14	77.34	72.15
SIFT + FisherVector	83.28	60.10	62.40	71.09	67.29	65.56	76.34	70.92
DenseSIFT + BoW	56.22	54.38	56.58	64.29	59.94	58.94	60.07	59.85
DenseSIFT + VLAD	58.25	55.74	64.38	67.12	66.49	65.01	77.17	62.13
DenseSIFT + FisherVector	62.55	59.21	64.01	71.76	68.01	65.96	78.01	62.97
HOG + BoW	79.99	60.95	62.40	68.48	61.92	60.99	61.23	61.05
HOG + VLAD	82.52	57.49	68.97	71.99	67.74	66.43	61.92	63.38
HOG + FisherVector	83.76	61.41	68.11	76.07	70.34	68.32	68.12	65.33
PAEF [8]	62.81	70.05	67.85	72.90	69.61	67.92	77.51	75.24
SentiBank [17]	81.79	64.95	67.74	71.32	68.28	66.63	65.93	66.18
DeepSentiBank [49]	85.63	71.19	68.73	76.35	70.15	71.25	70.23	70.11
PCNN (VGGNet) [14]	88.84	70.84	70.96	82.54	76.52	76.36	77.68	73.58
VGGNet	88.51	68.86	67.61	83.44	78.67	75.49	71.79	72.25
Fine-tuned VGGNet	89.37	72.48	70.09	84.35	82.26	76.75	76.99	77.02
obj + concatenation	88.47	73.38	71.34	84.24	81.81	76.68	75.97	77.83
senti + concatenation	88.74	74.23	72.86	84.35	82.44	76.57	78.18	77.95
AR + concatenation	89.39	74.71	73.76	86.10	83.25	77.97	78.89	78.52
AR + sum-pooling	90.32	73.72	73.63	86.39	83.41	77.57	78.32	78.43
AR + max-pooling	89.04	73.92	73.32	86.19	83.11	77.67	78.52	78.32
AR + concatenation ($K = 8$)	**92.39**	**76.03**	**74.80**	**88.65**	**85.10**	**81.06**	**80.48**	**81.26**

Fig. 11. Classification results of different methods on the small-scale datasets. GCH represents the features of global color histogram and LCH corresponds to local color histogram. Note that "obj" means that we only regard the proposals with high objectness score as affective regions, "senti" refers to the proposals having high sentiment score are used. Note that our method is based on the fine-tuned VGGNet.

our proposed framework. It is hard to evaluate the quality of the discovered affective regions directly due to lack of annotations. Therefore, our aim is to discover how many affective regions can boost sentiment prediction accuracy. We show the classification performance when combing different numbers of affective regions for sentiment analysis. As shown in Fig. 9, as the number of affective regions is increased, the accuracy increases as more information becomes available. However, a further increase in the number of regions leads to a slight decrease in performance due to the introduction of noisy regions. Therefore, as a good balance, we choose to combine 8 affective regions for sentiment analysis in the remaining experiments, which outperforms the fine-tuned VGGNet by 3.3% on FI (86.35%) and 1% on Flickr (71.13%). We also report the true positive rate of different sentiments on the large-scale datasets. In detail, the positive and negative sentiments achieve 92.10% and 72.65% on FI, respectively; and on Flickr achieve 73.56% and 47.92%, respectively. For both datasets, the positive class receives a higher accuracy than the negative class, which is consistent with the number of training images. More training images can lead to a higher probability that the corresponding sentiment receives a higher true positive rate.

E. Results on Small-Scale Datasets

We transfer the parameters learned on the FI dataset to small-scale datasets, then show our experimental results in Fig. 11 and provide comparisons to several state-of-the-art works. Note that our method is based on the fine-tuned VGGNet. "obj" means that we only regard the proposals with relatively high objectness score as affective regions and "senti" refers to the proposals having high sentiment score. Our "AR" method selects affective regions, where both objectness score and sentiment score are considered.

For the color features, ColorName is usually not enough to describe the distribution of image color compared to GCH and LCH except for the Abstract dataset. For the texture features, the HOG descriptor is able to achieve the best prediction accuracy in most datasets compared with other texture representations like SIFT, Gist, LBP and Gabor. Texture has better discriminative power than color on these small datasets. The reason is that sentiments are usually conveyed through complicated texture regions, *e.g.*, faces, dogs, buildings *etc*. In addition, we also compare the different encoding algorithms in Fig. 11. As can be seen, it achieves better performance while using the Fisher Vector to encode these descriptors on most datasets.

Compared with the traditional representations based mainly on color and texture information, the deep methods achieve better results, as expected. Our proposed method employs affective regions and outperforms both hand-crafted features-based methods and deep approaches, and achieves the best accuracy in all the small datasets. In detail, compared to the SentiBank and DeepSentiBank which do not use affective

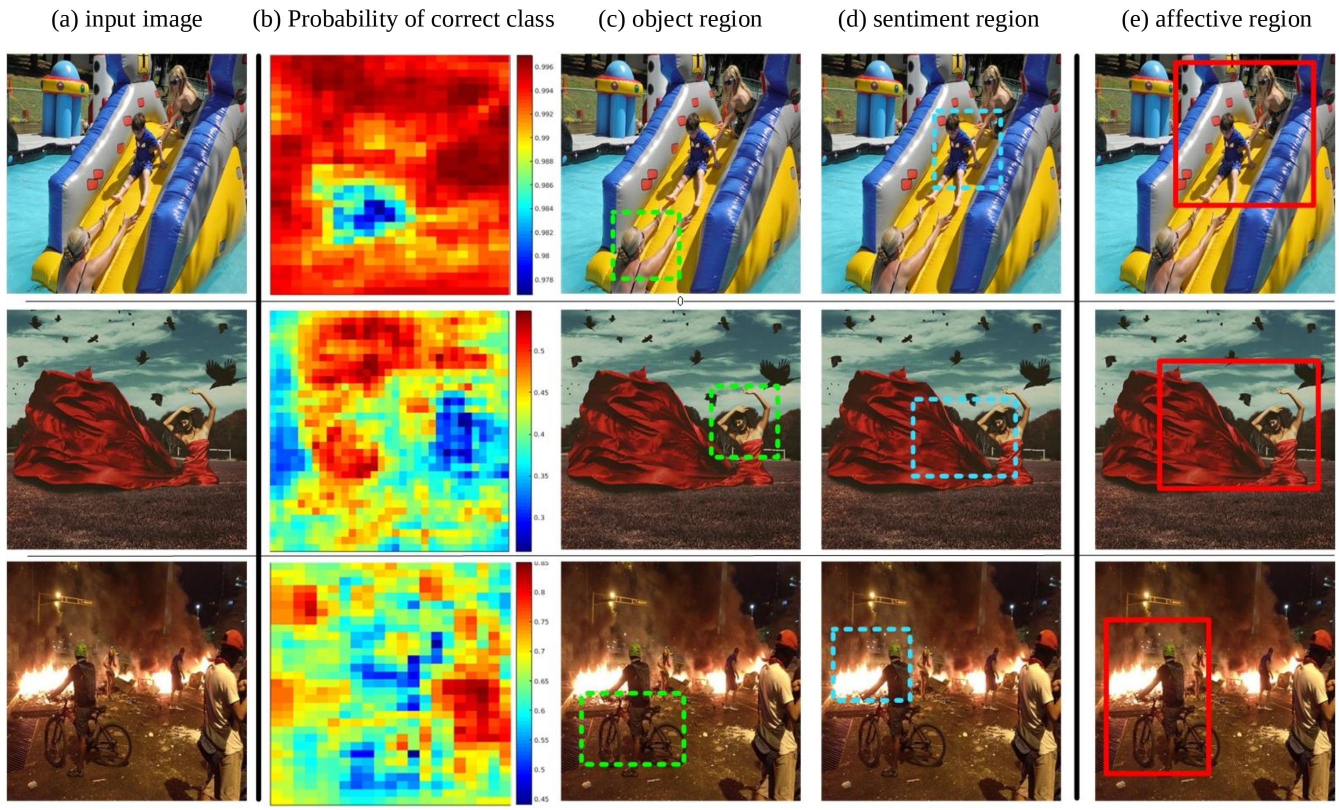


Fig. 12. Visualization of images from the FI dataset. Given the input image (a), we systematically cover up different portions of the image with a gray square and see how the classifier output (b) changes. Column (b) denotes a map of the probabilities estimated by the CNN for the ground-truth class, indicating the relative importance of locations in the affective image for the CNN. We also show the top-1 regions ranked by different scores (*i.e.*, Obj_score, Senti_score, AR) as object region (c), sentiment region (d), and affective region (e).

regions and represent images at the mid-level, our method outperforms them by a large margin. Furthermore, our method also shows an advantage over PCNN on all the affective datasets, and the three fusion operations are all useful with concatenation being the most effective method. According to our experimental findings on the large-scale datasets, when we increase the number of affective regions many regions have little impact on image sentiment and can even decrease the prediction accuracy. Therefore, we combine the same number of regions for the final sentiment prediction on the small-scale datasets and achieve the best performance. This shows another advantage of our method, which is that we do not need many local regions to be included in the deep model, ensuring an acceptable increase in computation overhead.

F. Affective Regions Evaluation

We evaluate the affective regions detected by our framework on the EmotionROI dataset, taking the same training/testing split as the previous works [44], [74]. Since the dataset only provides as ground truth the normalized Emotion Stimuli Map, which is based on 15 bounding boxes, we first binarize the Emotion Stimuli Map with threshold values $\gamma \in [0..255]/255$, and compare the ground truth region with the most confident discovered affective regions. Precision and recall are employed, which represents the percentages of detected emotionally involved pixels out of all the pixels identified in the predicted region or the ground truth. Following [44], all the predicted affective regions and ground truth are normalized to 0 to 1 for evaluation. Fig. 10 shows the precision-recall curve of the objectness score and our proposed AR score. The average precision and recall of our method are 0.69 and 0.59, while the objectness measure achieves 0.63 and 0.53, indicating that the selected affective regions are more consistent with the human annotation.

V. Visualization Of Affective Regions

For the image classification approaches, a natural question is whether the proposed model can identify the target part in the image. In this section, we attempt to answer this question by visualizing the crucial location for classifying sentiment. Following the previous works [75], we use sliding windows to occlude different portions of the input image with a gray square, and then generate the heat-map by plotting the estimated probability of the ground truth class at that location. Compared to other visualization methods, *e.g.*, embedding the features with t-SNE or visualizing the filters of the network, this method tends to directly show the regions that the CNN focuses on. As shown in Fig. 12, the first column is the input image and the second column is the prediction probability of the correct class using the fine-tuned VGGNet when occluding the corresponding portions of the image. If the occluded portion is essential for the sentiment prediction, the corresponding probability in the heat-map will obviously decrease (blue pixels). As can be seen in the three examples, the fine-tuned deep model has the ability to discover the parts in the images that can evoke the sentiment. For example, occluding the salient objects (*e.g.*, people, fire) that can evoke the sentiment leads to decreasing prediction probabilities. However, due to the affective gap, the CNN is not discriminative enough to capture the most significant sentiment information in the images.

We also visualize the top-1 regions by re-ranking the candidate proposals according to different scores (*i.e.*, Obj_score, Senti_score, AR) in Fig. 12 (c) (d) (e), respectively. Column (c) and (d) refer to the regions that are selected using only objectness or sentiment scores. The objectness score selects the regions which contain rich information at the low level, while the sentiment score usually evaluates the regions' sentiment at the affective level. Considering information from both of these two aspects, our proposed method is able to discover more accurate affective regions, see column (e). The detected affective regions can be not only complementary for the salient objects in the image (first example), but also extend the regions of interest to the additional contextual background (last two examples). Thus, combing the global and local information can be discriminative for the visual sentiment analysis.

VI. Conclusion

In this paper, we address the problem of automatically recognizing sentiments in images. Inspired by the observation that both global appearance and local regions produce significant sentiment responses, we propose a framework to discover affective regions and combine both information using CNN. We estimate the level of sentiment content in a region considering the objectness score and sentiment score. The objectness score usually finds regions containing rich texture information while the sentiment score evaluates the regions' sentiment at the affective level. We also consider three alternative fusion operations and implement the proposed model on VGGNet. The experimental results show that our method outperforms the state-of-the-art methods on the popular affective datasets.

Acknowledgments

This research was supported by NSFC (NO. 61620106008, 61572264, 61633021, 61525306, 61301238, 61201424), the Open Project Program of NLPR, China (NO. 201800012), Huawei Innovation Research Program, CAST YESS Program, and IBM Global SUR award.

References

[1] S. Zhao, Y. Gao, G. Ding, and T. S. Chua, "Real-time multimedia social event detection in microblog," *IEEE Trans. Cyber.*, vol. PP, no. 99, pp. 1–14, 2017.

[2] B. Pang, L. Lee *et al.*, "Opinion mining and sentiment analysis," *Found. Trends Inf. Ret.*, vol. 2, no. 1–2, pp. 1–135, 2008.

[3] L. Pang, S. Zhu, and C. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2008–2020, 2015.

[4] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "RAPID: Rating pictorial aesthetics using deep learning," in *ACM Int. Conf. Multimedia*, 2014.

[5] Q. You, J. Luo, H. Jin, and J. Yang, "Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia," in *ACM Int. Conf. Web Search and Data Mining*, 2016.

[6] Y.-Y. Chen, T. Chen, T. Liu, H.-Y. M. Liao, and S.-F. Chang, "Assistive image comment robota novel mid-level concept-based representation," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 298–311, 2015.

[7] A. Sartori, D. Culibrk, Y. Yan, and N. Sebe, "Who's afraid of Itten: Using the art theory of color combination to analyze emotions in abstract paintings," in *ACM Int. Conf. Multimedia*, 2015.

[8] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in *ACM Int. Conf. Multimedia*, 2014.

[9] V. Yanulevskaya, J. Van Gemert, K. Roth, A.-K. Herbold, N. Sebe, and J.-M. Geusebroek, "Emotional valence categorization using holistic image features," in *IEEE Int. Conf. Image Process.*, 2008.

[10] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *ACM Int. Conf. Multimedia*, 2010.

[11] X. Lu, P. Suryanarayan, R. B. Adams Jr, J. Li, M. G. Newman, and J. Z. Wang, "On shape and the computability of emotions," in *ACM Int. Conf. Multimedia*, 2012.

[12] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: a decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.

[13] V. Campos, A. Salvador, X. Giró i Nieto, and B. Jou, "Diving deep into sentiment: Understanding fine-tuned CNNs for visual sentiment prediction," in *International Workshop on Affect & Sentiment in Multimedia*, 2015.

[14] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *AAAI Conf. Artif. Intell.*, 2015.

[15] V. Campos, B. Jou, and X. Gir-I-Nieto, "From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction," *Image Vision Comput.*, vol. 65, pp. 15–22, 2017.

[16] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *AAAI Conf. Artif. Intell.*, 2016.

[17] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *ACM Int. Conf. Multimedia*, 2013.

[18] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Proc. Mag.*, vol. 28, no. 5, pp. 94–115, 2011.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[20] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Adv. Neural Inform. Process. Syst.*, 2014.

[21] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen, "Where do emotions come from? predicting the emotion stimuli map," in *IEEE Int. Conf. Image Process.*, 2016.

[22] Q. You, H. Jin, and J. Luo, "Visual sentiment analysis by attending on local image regions." in *AAAI Conf. Artif. Intell.*, 2017.

[23] B. Li, W. Xiong, W. Hu, and X. Ding, "Context-aware affective images classification based on bilayer sparse representation," in *ACM Int. Conf. Multimedia*, 2012.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Adv. Neural Inform. Process. Syst.*, 2015.

[25] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annu. Rev. Neurosci.*, vol. 18, no. 1, pp. 193–222, 1995.

[26] T. Chen, F. X. Yu, J. Chen, Y. Cui, Y.-Y. Chen, and S.-F. Chang, "Object-based visual sentiment concept analysis and application," in *ACM Int. Conf. Multimedia*, 2014.

[27] M. Sun, J. Yang, K. Wang, and H. Shen, "Discovering affective regions in deep convolutional neural networks for visual sentiment prediction," in *Int. Conf. Multimedia and Expo*, 2016.

[28] S. Benini, L. Canini, and R. Leonardi, "A connotative space for supporting movie affective recommendation," *IEEE Trans. Multimedia*, vol. 13, no. 6, pp. 1356–1370, 2011.

[29] B. Jou, S. Bhattacharya, and S.-F. Chang, "Predicting viewer perceived emotions in animated GIFs," in *ACM Int. Conf. Multimedia*, 2014.

[30] M. A. Nicolaou, H. Gunes, and M. Pantic, "A multi-layer hybrid framework for dimensional emotion classification," in *ACM Int. Conf. Multimedia*, 2011.

[31] M. Solli and R. Lenz, “Color based bags-of-emotions,” in *Int. Conf. Comput. Anal. Images Patterns*, 2009.

[32] A. Hanjalic, “Extracting moods from pictures and sounds: Towards truly personalized tv,” *IEEE Signal Proc. Mag.*, vol. 23, no. 2, pp. 90–100, 2006.

[33] S. Zhao, H. Yao, Y. Gao, R. Ji, W. Xie, X. Jiang, and T. Chua, “Predicting personalized emotion perceptions of social images,” in *ACM Int. Conf. Multimedia*, 2016.

[34] S. Zhao, H. Yao, Y. Gao, G. Ding, and T.-S. Chua, “Predicting personalized image emotion perceptions in social networks,” *IEEE Trans. Affect. Comput.*, 2018.

[35] S. Zhao, H. Yao, and X. Jiang, “Predicting continuous probability distribution of image emotions in valence-arousal space,” in *ACM Int. Conf. Multimedia*, 2015.

[36] J. Yang, D. She, and M. Sun, “Joint image emotion classification and distribution learning via deep convolutional neural network,” in *Int. J. Conf. Artif. Intell.*, 2017.

[37] J. Yang, M. Sun, and X. Sun, “Learning visual sentiment distributions via augmented conditional probability neural network,” in *AAAI Conf. Artif. Intell.*, 2017.

[38] S. Zhao, G. Ding, Y. Gao, and J. Han, “Approximating discrete probability distribution of image emotions by multi-modal features fusion,” in *Int. J. Conf. Artif. Intell.*, 2017.

[39] S. Zhao, H. Yao, Y. Gao, R. Ji, and G. Ding, “Continuous probability distribution prediction of image emotions via multitask shared sparse regression,” *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 632–645, 2017.

[40] Z. Li, Y. Fan, W. Liu, and F. Wang, “Image sentiment prediction based on textual descriptions with adjective noun pairs,” *Multimed. Tools Appl.*, pp. 1–18, 2017.

[41] J. Yuan, S. Mcdonough, Q. You, and J. Luo, “Sentribute: image sentiment analysis from a mid-level perspective,” in *ACM International Workshop on Issues of Sentiment Discovery and Opinion Mining*, 2013.

[42] S. Zhao, H. Yao, Y. Yang, and Y. Zhang, “Affective image retrieval via multi-graph learning,” in *ACM Int. Conf. Multimedia*, 2014.

[43] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, “Emotional category data on images from the international affective picture system,” *Behavior Research Methods*, vol. 37, no. 4, pp. 626–630, 2005.

[44] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen, “Where do emotions come from? predicting the emotion stimuli map,” in *IEEE Int. Conf. Image Process.*, 2016.

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Adv. Neural Inform. Process. Syst.*, 2012.

[46] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014.

[47] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural. Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[48] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, “Part-based R-CNNs for fine-grained category detection,” in *Eur. Conf. Comput. Vis.*, 2014.

[49] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, “DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks,” *ArXiv e-prints*, 2014.

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009.

[51] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Region-based convolutional networks for accurate object detection and segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, 2016.

[52] R. Girshick, “Fast R-CNN,” in *Int. Conf. Comput. Vis.*, 2015.

[53] L. Zhang, L. Lin, X. Liang, and K. He, “Is faster R-CNN doing well for pedestrian detection?” in *Eur. Conf. Comput. Vis.*, 2016.

[54] X. Peng and C. Schmid, “Multi-region two-stream R-CNN for action detection,” in *Eur. Conf. Comput. Vis.*, 2016.

[55] G. Gkioxari, R. Girshick, and J. Malik, “Contextual action recognition with R*CNN,” in *Int. Conf. Comput. Vis.*, 2015.

[56] S. Gidaris and N. Komodakis, “Object detection via a multi-region and semantic segmentation-aware cnn model,” in *Int. Conf. Comput. Vis.*, 2015.

[57] C. L. Zitnick and P. Dollár, “Edge boxes: Locating object proposals from edges.” in *Eur. Conf. Comput. Vis.*, 2014.

[58] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[59] J. H. Hosang, R. Benenson, and B. Schiele, “How good are detection proposals, really?” in *Brit. Mach. Vis. Conf.*, 2014.

[60] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, “What makes for effective detection proposals?” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 814–830, 2016.

[61] M. Cheng, Z. Zhang, W. Lin, and P. Torr, “BING: Binarized normed gradients for objectness estimation at 300fps,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014.

[62] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders, “Segmentation as selective search for object recognition,” in *Int. Conf. Comput. Vis.*, 2011.

[63] B. Alexe, T. Deselaers, and V. Ferrari, “Measuring the objectness of image windows,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, 2012.

[64] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, “HCP: A flexible CNN framework for multi-label image classification,” vol. 38, no. 9, pp. 1901–1907, 2016.

[65] H. Wu, H. Zhang, J. Zhang, and F. Xu, “Typical target detection in satellite images based on convolutional neural networks,” in *Int. Conf. Syst. Man. Cy.*, 2015.

[66] E. Elhamifar and R. Vidal, “Sparse subspace clustering,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009.

[67] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[68] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller, “Recognizing image style,” in *Brit. Mach. Vis. Conf.*, 2013.

[69] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Int. Conf. Learn. Represent.*, 2015.

[70] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, “International affective picture system (IAPS): Affective ratings of pictures and instruction manual,” *Technical report*, 2008.

[71] W. Weining, Y. Yinglin, and J. Shengming, “Image retrieval by emotional semantics: A study of emotional space and feature extraction,” in *Int. Conf. Syst. Man. Cy.*, 2006.

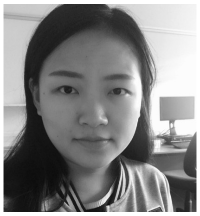
[72] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, “Analyzing and predicting sentiment of images on the social web,” in *ACM Int. Conf. Multimedia*, 2010.

[73] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Trans. Intel. Syst. Tec*, vol. 2, no. 3, pp. 1–27, 2011.

[74] K.-C. Peng, T. Chen, A. Sadovnik, and A. Gallagher, “A mixed bag of emotions: Model, predict, and transfer emotion distributions,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015.

[75] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Eur. Conf. Comput. Vis.*, 2014.

Jufeng Yang is an associate professor in the College of Computer and Control Engineering, Nankai University. He received the PhD degree from Nankai University in 2009. From 2015 to 2016, he was working at the Vision and Learning Lab, University of California, Merced. His research falls in the field of computer vision, machine learning and multimedia.



Dongyu She is currently a Master student with the College of Computer and Control Engineering, Nankai University. Her current research interests include computer vision, machine learning, pattern recognition.

Ming Sun received his Master degree from Nankai University in 2017. He is now a research associate in Baidu Institute of Deep Learning (IDL). His research interests include computer vision, deep learning and machine learning. He was named as Excellent Graduate, and received Outstanding Dissertations Award from Nankai University.

Ming-Ming Cheng received his PhD degree from Tsinghua University in 2012. Then he did 2 years research fellow, with Prof. Philip Torr in Oxford. He is now a professor at Nankai University, leading the Media Computing Lab. His research interests includes computer graphics, computer vision, and image processing. He received research awards including ACM China Rising Star Award, IBM Global SUR Award, CCF-Intel Young Faculty Researcher Program, *etc*.

Paul L. Rosin is a professor at the School of Computer Science & Informatics, Cardiff University. His research interests include the representation, segmentation, and grouping of curves, knowledge-based vision systems, early image representations, low level image processing, machine vision approaches to remote sensing, methods for evaluation of approximation algorithms, medical and biological image analysis, mesh processing, non-photorealistic rendering and the analysis of shape in art and architecture.

Liang Wang received the Ph.D. degree in pattern recognition and intelligent system from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (CAS), China, in 2004. He was a Research Assistant with the Imperial College London, U.K., and Monash University, Australia, and a Research Fellow with the University of Melbourne, Australia. He was a Lecturer with the Department of Computer Science, University of Bath, U.K. Currently, he is a Professor of the Hundred Talents Program of CAS with the Institute of Automation, CAS. His major research interests include machine learning, pattern recognition, computer vision, and data mining.