Multimedia Tools and Applications (2022) 81:7941–7955
https://doi.org/10.1007/s11042-022-11982-5

Salient object based visual sentiment analysis by combining deep features and handcrafted features

S. Sowmyayani[1] · P. Arockia Jansi Rani[2]

Received: 16 September 2020 / Revised: 7 January 2021 / Accepted: 3 January 2022 /
Published online: 29 January 2022

Abstract
With the rapid growth of social networks, the visual sentiment analysis has quickly emerged for opinion mining. Recent study reveals that the sentiments conveyed by some images are related to salient objects in them, we propose a scheme for visual sentiment analysis that combines deep and handcrafted features. First, the salient objects are identified from the entire images. Then a pre-trained model such as VGG16 is used to extract deep features from the salient objects. In addition, hand-crafted features such as Visual texture, Colourfulness, Complexity and Fourier Sigma are extracted from all the salient objects. Deep features are combined individually with all the handcrafted features and the performance is measured. The sentiment is predicted using Convolutional Neural Network Classifier. The proposed method is tested on ArtPhoto, Emotion6, Abstract, IAPS datasets, Flickr and Flickr & Instagram datasets. The experimental results substantially proved that the proposed method achieves higher accuracy than other methods.

Keywords Salient object · Convolutional neural network · Visual sentiment · Clutter

1 Introduction

Nowadays, users share large amount of multimedia data such as images, videos on social networks. Therefore, there is a need for making machines to interpret and relate the multimedia data like humans. Although the exact interpretation of data like humans is complex, it can be

✉ S. Sowmyayani
sowmyayani@gmail.com

P. Arockia Jansi Rani
jansimsuniv@gmail.com

[1] Department of Computer Science, St. Mary's College (Autonomous), Thoothukudi, Tamilnadu, India

[2] Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India

noted that there is an emotional component that plays a role in forming the response, and this component has the main difference between how humans and machines relate to their emotion.

Images are very powerful tools for conveying moods and emotions. People can express their feelings and communicate with others using images. With the increasing development of deep learning technology, even computers are able to recognize objects, faces, and actions. It has also started to write image captions and answer questions about images. In the similar way, computers can be trained to have similar feeling as humans when looking at images.

Predicting emotion from an image is a more complex task and is still in its early stage. As the deep learning technology has its footprint in various research such as image classification, segmentation and image processing [10, 19, 20], several studies have been introduced recently that apply the deep learning for the emotion prediction [8]. Those works mostly use the Convolutional Neural Network (CNN), which has shown better performance for the emotion prediction compared to the model that uses a shallow network. Also, it is an effective network model for learning filters that capture the shapes that repeatedly appear in images.

The main issue in the emotion recognition is the affective gap. The affective gap is the lack of coincidence between the measurable signal properties, commonly referred to as features, and the expected affective state in which the user is brought by perceiving the signal [13]. To narrow this affective gap, several researchers developed emotion classification systems using psychology and art theory based high level features such as harmony, movement, rule of third, etc. [21, 22, 38]. Although the above-mentioned features help to improve the accuracy, identifying a better set of features are still in research. It is observed that the salient object appearing in the image plays an important role in determining the emotion. The idea is that salient objects can be good cues for emotions. Hence salient objects are used to extract features for sentiment classification.

The emotion may vary even if images have the same object with different background. In this paper, instead of using the entire images, salient objects are used for determining sentiment. The deep features extracted from a pre-trained deep network model are combined with hand-crafted features such as color statistics to obtain a set of features for the emotion prediction. Using this feature set, a CNN classifier is used to produce the sentiment for a given image.

The remaining of the paper is organized as follows: Section 2 briefly explains some of the related works. Section 3 and 4 describes the system architecture of the proposed method and the methodologies used in it. Section 5 analyzes the proposed method with some experiments and gives an ablation study. Section 6 concludes and gives future work.

2 Related work

Emotion of an image can be evoked by various factors. To figure out significant features for the emotion prediction problem, many researchers have considered various types of the features from color statistics to the art and the psychological features. This section briefly gives some methods that use handcrafted features for emotion recognition. The methods that bridge the affective gap between low-level features and the emotion are also discussed. Finally, the importance of local regions towards predicting visual sentiment is discussed.

Machajdik et al. [22] introduced an affective image classification system using psychology and art theory based features such as Itten's color contrast and rule of thirds. Zhao et al. [38] exploited to extract principles of art features for emotion classification. Similarly, Lu et al. [21]

computed the shape based features in natural images. As a high level concept for representing the sentiment of an image, adjective noun pairs are introduced by Borth et al. in [5]. Using the dataset from Borth et al. [5], Chen et al. [9] classified the adjective-noun pairs using the CNN and achieved better accuracy than Borth et al.'s method.

Some methods are developed to bridge the affective gap. In [1], an effective approach is introduced for affective classification of images by training an ensemble of kernel-based regressors that use different subset of features for different emotion classes. This method bridges the affective gap between image content and emotional response by understanding the High Level Concepts (HLC). A framework is developed to leverage Affective Regions (AR), where an objectness tool is used to generate the candidates [34], from which sentiment scores are calculated. Both the scores are used to predict the sentiment. An Affective Structural Embedding (ASE) framework is developed by utilizing mid-level semantic representation to construct an intermediate embedding space [37]. This method also bridges the affective gap between low-level visual features and high-level semantics.

Despite the rise of deep learning studies, several studies utilized the pre-trained model for the image classification and transferred the learned parameter. By changing the number of outputs to be the same as the number of labels of their dataset, the classifier can be trained. A Region-based CNN using Group Sparse Regularization (R-CNNGSR) is designed for image sentiment classification [32]. This method obtains the initial sentiment prediction model through CNN using group sparse regularization, and then detects the sentiment regions by combining the underlying features and sentimental features.

By the deep study in the psychological and neuroscientific findings, it is understood that human's visual attention is attracted to the most informative regions [18, 24, 29]. Based on this fact, some studies recently used local areas of an image instead of the whole image. From the local regions, sentiment-related features are extracted for visual sentiment analysis [5, 15, 17]. But still it has a drawback: the local areas are not truly stripped from the original images and their effects on sentiment are not considered separately, leading to a lot of noises.

In advancement to predicting sentiment using local regions, salient objects are used as it contains useful information of sentiment. The next section describes the proposed method with its architecture.

3 System architecture

The visual sentiment can be identified by the salient objects present in it. Prior to the conventional classification system, salient objects are identified from the entire image. Then deep and handcrafted features are extracted from the salient objects. These features are fused and classified using CNN classifier. The proposed method contains five main modules: Salient Object Detection, Candidate Selection, Feature Extraction, Feature Fusion and Classification. The system architecture of the proposed framework is shown in Fig. 1.

Initially, salient objects are identified from all images in the dataset. The salient objects are cropped to a required size for further processing. For each image, there may be more than one salient objects. Hence, it is necessary to select a single salient object among them. Then deep features are extracted from the salient image using pre-trained deep network model. Handcrafted features such as visual clutter, complexity, colourfulness and Fourier sigma are also extracted from the salient image. These features are combined in different ways and classified using CNN classifier.

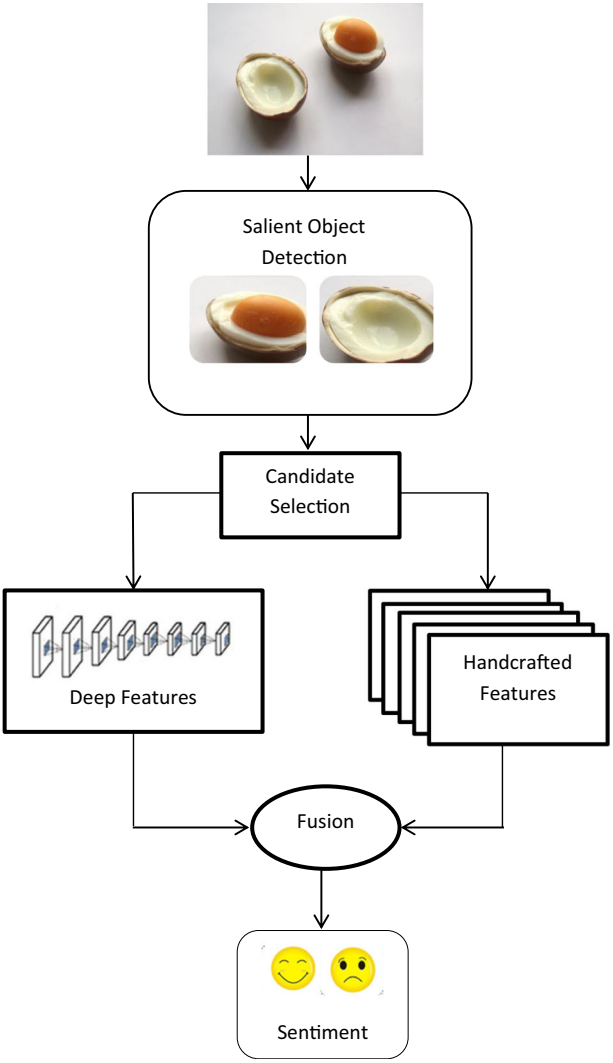

Fig. 1 Proposed System Architecture

4 Proposed framework

This section describes the methods that are used in each module.

4.1 Salient object detection

A salient object often has some unique appearance in terms of color or texture as comparing to its surroundings, implying some visual dissimilarity. Hence salient object plays a vital role in analyzing sentiment. As mentioned above, in the first module, salient objects are detected from the entire image using spectral saliency detection model [12]. This method integrates hierarchical spectral partition and saliency estimation. Each input image is divided into superpixels

and edges are identified. Then superpixels are used as vertices to construct a graph that captures intrinsic colors and edge information of an image.

Hierarchical spectral partition that is aimed at retaining the entire object is applied to the graph, meanwhile saliency estimation is carried out on successive hierarchies. The resultant intermediate saliency maps are integrated and then refined by graph-based semi-supervised learning, yielding final saliency map. Salient regions are obtained from the saliency map using intensity values. Salient regions are cropped, so that the extracted local regions are of required size.

4.1.1 Cropping salient regions

The salient objects obtained from the image are of irregular size. Any CNN will work in the input images of equal size. Based on the detection window for salient objects, the salient regions are cropped automatically. By analyzing the salient regions, it is found that some are so small that resizing them using a CNN can reduce the performance of the algorithm. Hence the following rules are applied while cropping salient regions:

- If the widths or heights of the detection windows are smaller than 200 pixels, the corresponding edge is extended to 200 pixels.
- If the width or the height of the detection window is greater than or equal to 200 pixels, there is no change of the size.

Based on the above rules, the salient objects will be cropped from the entire images.

4.2 Candidate selection

It is necessary to filter out the salient objects carrying little sentiment, and removing those at the initial stage of the algorithm can greatly reduce the computation time of the subsequent steps. To address this problem, the candidate selection module is used to select salient objects from the selected candidates inspired by [30].

The CNN based on the deep model VGGNet with 16 layers is used for this purpose. In order to adapt the pre-trained model on ImageNet for sentiment analysis, the CNN is first fine-tuned on the target dataset utilizing the original images to adjust the parameters of the deep model. As a supervised learning approach, the fine-tuned CNN is applied to learn a function $f : \mathscr{T} \rightarrow \mathscr{S}$, from a collection of 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. Let 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:

$$l(\mathbf{W}) = \sum_{i=1}^{N} \sum_{j \in l} \mathbf{1}(\mathbf{l_i} = \mathbf{j}) \log \mathbf{p}\left(\mathbf{l_i} = \mathbf{j} \mid \mathbf{d_i}, \mathbf{w_j}\right) \tag{1}$$

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

$$p(l_i = j \mid \mathbf{d}_i, \mathbf{w}_j) = \frac{\exp\left(\mathbf{w}_j^T \mathbf{d}_i\right)}{\sum_{j' \in l} \exp\left(\mathbf{w}_j^T \mathbf{d}_i\right)} \tag{2}$$

Since the number of classes in the dataset is not equal to that of ImageNet model, the output of fc8 layer is changed to 2 as required by the dataset.

4.2.1 Estimating sentiment score

For identifying the sentiment-level of the salient objects, the sentiment scores are computed by feeding the salient objects to the CNN. For the generated salient objects $H = \{h_i\}_{i=1}^{m}$ of the input image I, let $\left\{y_{ij}\right\}_{j=1}^{c}$ 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.

A probabilistic sampling function is used to evaluate the sentiment score of the i^{th} region in a sentiment level perspective as follows:

$$\text{Senti\_score}_i^I = \sum_{j=1}^{c} y_{ij} * \log y_{ij} + 1 \tag{3}$$

where the score ranges between 0 and 1 for binary classification. The information entropy defined in Eq. (3) represents the degree of uncertainty when predicting sentiment. The Senti_score_i^I provides a more semantic measurement compared to other methods. The salient object with maximum sentiment score are chosen for sentiment prediction.

4.3 Feature extraction

4.3.1 Hand-crafted features

Objective features like visual texture, complexity, colorfulness and Fourier Sigma influence affective observer responses to visual scenes in general. These features are described as follows:

The first four measures characterize the texture of color images. These measures are based on the Pyramid Histogram of Oriented Gradients (PHOG) image representation that was originally developed for object recognition and classification. The PHOG descriptors are global feature vectors based on a pyramidal subdivision of an image into sub-images, for which Histograms of Oriented Gradients (HOG) [11] are computed.

For calculating PHOG for color images, each image is converted to the Lab color space. HOG values are then calculated based on the maximum gradient magnitudes in the L, a and b color channels. For this, a new gradient image $\mathbf{G}_{\max}$ is generated,

$$\mathbf{G}_{\max}(x,y) = \max(\|\nabla I_L(x,y)\|, \|\nabla I_a(x,y)\|, \|\nabla I_b(x,y)\|) \tag{4}$$

The aesthetic appeal of images may depend on their degree of **complexity**. To calculate the complexity of an image, the mean norm of the gradient across all orientations over $\mathbf{G}_{\max}(x, y)$ is calculated as shown in Eq. (5).

$$M_{\mathrm{Co}}(\mathbf{G}_{\max}) = \frac{1}{N \cdot M} \sum_{(x,y)} \mathbf{G}_{\max}(x,y) \tag{5}$$

In this equation, M_{Co} corresponds to complexity, and N and M are the height and width of the new gradient image, **G**max. Since image gradients represent the changes of lightness in an image, calculating the mean gradient over the L channel will give a good prediction on image complexity. The higher the mean absolute gradient, the more complex an image is.

Self-similarity is computed using the Histogram Intersection Kernel (HIK) [3] to determine the similarity between HOG features at the individual levels of the PHOG. Images of natural patterns have a highly self-similarity, whereas artificial patterns have a low self-similarity.

$$\mathrm{HIK}\left(\mathbf{h}, \mathbf{h}^{'}\right) = \sum_{i=1}^{m} \min\left(h(i), h^{'}(i)\right) \tag{6}$$

In Eq. (6), h and $h^{'}$ are two different normalized histograms and m is the number of bins present in the HOG features. To calculate the self-similarity of an image, the median value of the HIK values at each level is calculated

$$M_{\mathrm{SeSf}}(I,L) = \mathrm{median}(\mathrm{HIK}(h(S), (h(\Pr(S)))) \mid \Pr(S) \in \mathrm{Sections}(I,L)) \tag{7}$$

In Eq. (7), $M_{\mathrm{SeSf}}$is the self-similarity value, I corresponds to the image, L represents the level, at which the HOG features are assessed (in this work we use L = 3), h(S) is the HOG value for a sub-image in the Sections(I,L) which corresponds to the sections in the image I in level L and Pr(S) corresponds to the parent of sub-image S.

Anisotropy describes how the gradient strength varies across the orientations in an image. Low anisotropy means that the strengths of the orientations are uniform across orientations and high anisotropy means that orientations differ in their overall prominence [27].

$$M_{\mathrm{AnI}}(L) = \sigma(H(L)) \tag{8}$$

In this equation, M_{AnI}represents the anisotropy in the image at level L, $H(L)$corresponds to a vector which is consisted of all the HOG value at level L, and σ is the variance. Consistent with the calculation of self-similarity, anisotropy is calculated at level 3.

Birkhoff-Like Measure (MBLM) [4] is described as the ratio of order and complexity in images. Birkhoff-like measure (M_{BLM}) is calculated according to Eq. (9) where M_{SeSf}represents the self-similarity in the image calculated in Eq. (7) and M_{Co}represents the complexity introduced in Eq. (5).

$$M_{\mathrm{BLM}} = \frac{M_{\mathrm{SeSf}}}{M_{\mathrm{Co}}} \tag{9}$$

The next four measures quantify the structural image complexity.

Feature Congestion(FC) is a visual clutter measure that implicitly captures the notion of spatial disorder by computing a weighted average of the local feature (color, orientation, and luminance) contrast covariance over multiple spatial scales [28]. Larger FC values correspond to higher levels of visual clutter.

Subband Entropy (SE) is a clutter measure that encodes the image information content by computing a weighted sum of the entropies of the luminance and chrominance image subbands [28]. Larger SE values correspond to higher levels of visual clutter.

The **Mean Information Gain (MIG)** is defined as the difference between the spatial heterogeneity and the non-spatial heterogeneity of an image [2, 26]. The MIG increases monotonously with spatial randomness and ranges over 0–1: MIG = 0 for uniform patterns and MIG = 1 for random patterns. The images were transformed to HSV format and the pixel value range was normalized from 0 to 10 before calculating MIG values independently for the color (Hue: MIGh), chroma (Saturation: MIGs), and intensity (Value: MIGv) components of each image.

The **Mean Gradient Strength (MGS)** or mean edge strength is a valid measure of subjective image complexity [7] and is based on the observation that the subjectively perceived level of image complexity increases with its number of edges.

The following two measures characterize the image color distribution. The **Number of Colors (NC)** represents the number of distinct colors in the RGB image. **Colorfulness** is the sensation that an image appears to be more or less chromatic. Local colorfulness has been defined as a linear combination of the mean and standard deviation of the local chrominance values in color opponent space [14]. Note that colorfulness is not strictly related to the numbers of colors: an image can be more colorful even when it contains less different colors. A global image Colorfulness (CF) metric was computed as the mean value of the local colorfulness over a set of subwindows covering the entire image support. CF varies from 0 (grayscale image) to 1 (most colorful image).

4.3.2 Deep features

As a branch of machine learning, deep learning has demonstrated impressive performance in image classification and object detection. In this work, CNN is utilized for feature extraction and classification. To verify the effectiveness of the proposed scheme, the popularly used model VGGNet was selected. Owing to the good performance of CNN on ImageNet, a fine-tuning tactic is used based on a model pre-trained on ImageNet. For VGGNet, the convolutional layers are maintained and changed only the number of outputs of the last fully connected layer from 4096 to 2. The salient objects are fed into VGG16 to extract features.

4.4 Feature fusion

The deep features and hand crafted features related to visual texture, complexity, colorfulness and Fourier Sigma which are described previously are combined to form a feature matrix. Each feature set is a $n \times 2$ vector and are normalized so that their values are all between 0 and 1.

The above features are concatenated as follows:

- The handcrafted features alone are concatenated to form a feature vector.
- The sum of the handcrafted features is used for classification.
- The deep features alone are used for classification.
- Each handcrafted features are concatenated with the deep features.
- The sum of the handcrafted features and the deep features are combined.

In particular, when combining handcrafted features with the deep features, sum pooling is used. The sum pooling fuses the prediction probability of all the features, where the weights of consistent features can be emphasized.

$$\overrightarrow{Y} = (1-\beta)*\overrightarrow{Y}_{DF} + \beta*\sum_{j=1}^{K}\overrightarrow{Y}_{HF} \tag{10}$$

where β is the trade-off between deep and handcrafted features prediction, $\overrightarrow{Y}_{DF}$ deep features prediction and $\overrightarrow{Y}_{HF}$ is the handcrafted features prediction. $\overrightarrow{Y}$, $\overrightarrow{Y}_{DF}$ and $\overrightarrow{Y}_{HF}$ 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. Concatenation is a simple but effective way by combining the features for a comprehensive representation:

$$\overrightarrow{Y} = \left[\overrightarrow{Y}_{DF}, \overrightarrow{Y}_{HF}\right] \tag{11}$$

The final feature is generated by concatenating all the prediction results. In this work, the number of salient regions in all samples is set to be 1, making it feasible to classify the concatenated feature vector using CNN classifier.

4.5 Summary

Based on the proposed framework, the sentiment classification of a given image can be summarized as follows. For a given image, salient objects are first generated. In order to reduce redundancy of salient objects for a single image, the candidate selection method is applied based on their sentiment scores and the best candidates are kept. Deep features are extracted using pretrained model and handcrafted features such as visual texture, complexity, colourfulness and Fourier Sigma are extracted by the described equations. All the features are classified individually and also combined with consistent weights. The feature vectors are classified using CNN classifier.

5 Experimental results

The experiment design includes image database, implementation details and the results analysis of various experiments. Experiments are carried out using two different pretrained deep network models and classifier. Next, all possible combination of deep features and handcrafted features are included to analyze the results. While combining the deep and handcrafted features, β value is changed to study the impact of deep features and handcrafted features. Finally, global features are included to check how it predicts the visual sentiment. All the above mentioned experiments are tested on four widely-used small-scale datasets, including IAPSa [22], Emotion6 [25] ArtPhoto [23], Abstract Paintings [23] and large-scale datasets including Flickr & Instagram [6] and Flickr [36]. The results of the proposed method are compared with state-of-the-art methods to show the efficacy of the proposed method.

5.1 Details of the datasets used

The International Affective Picture System (IAPS) [22] is a common stimulus dataset which is widely used in visual sentiment analysis research. IAPSa selects 395 pictures from IAPS and is labeled with Mikel's eight sentiment categories. In Emotion6 dataset [25], for each image, there are six dimensional emotion probability distribution vector (for six Ekman's emotion

(a) Emotion6 (b) IAPS

(c) Abstract (d) Art Photo

(e) FI (f) Flickr

Fig. 2 Sample Images from Tested Datasets

classes) and Valence Arousal (VA) values. These emotion probability distribution vectors were obtained through a user study and each image is no longer associated with a single emotion class.

ArtPhoto [23] contains artistic photographs from a photo sharing site and the ground truth labeling is provided by the owner of each image. Abstract Paintings [23] contains peer rated abstract paintings consisting of color and texture.

Flickr & Instagram [6] has images with eight sentiment categories (i.e. anger, amusement, awe, contentment, disgust, excitement, fear, sadness). A group of 225 Amazon Mechanical Turk (AMT) participants was asked to label the images, producing 23,308 images receiving at least three agreements. Flickr [36] dataset contains 60,745 images from Flickr with sentiment polarity (i.e. positive, negative) labels. Figure 2 shows the sample images from all the tested datasets. Table 1 shows the details of the small scale datasets.

Table 1 Types of emotion in all the tested Datasets

	Amusement	Anger	Awe	Contenment	Disgust	Excitement	Fear	Sad	Total
ArtPhoto	101	77	102	70	70	105	115	166	806
Abstract	25	3	15	64	18	36	36	32	226
IAPSa	37	8	54	63	74	55	42	62	395
Emotion6	330	330	–	–	330	330	330	330	1980

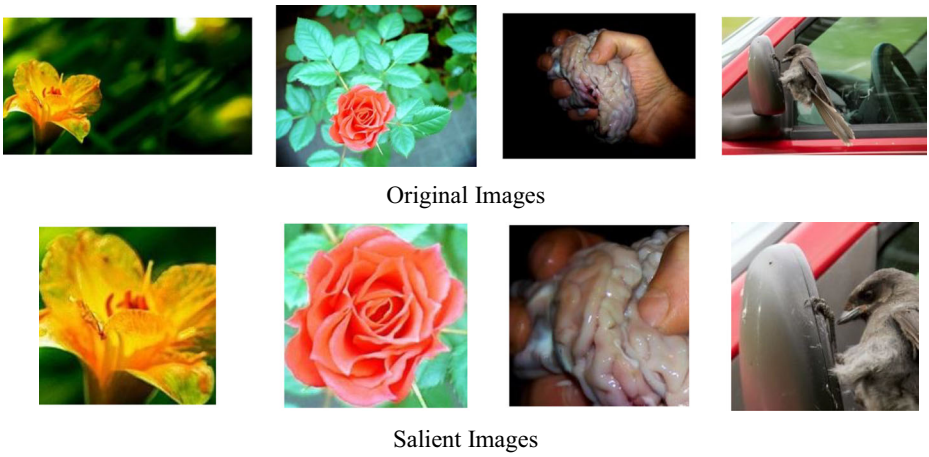

Fig. 3 Original and its Cropped Salient Images

5.2 Implementation details

The VGGNet with 16 layers is employed as the basic architecture for features extraction. It is initialized with the weights trained from ImageNet. Then the pre-trained network is fine-tuned on the 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.

All layers are fine-tuned by stochastic gradient descent through the whole net using a batch size of 64. A total of 100,000 iterations run to update the parameters to extract more precise sentiment-related information. All the experiments are carried out on NVIDIA GTX 1080 GPUs with 32 GB of CPU memory.

Figure 3 shows the some sample images and its salient objects. In this work, the 5-fold cross-validation is used and the average emotion recognition accuracy is shown as final results.

5.3 How different models influenced the performance

In Deep feature extraction, features are extracted using pre-trained model. Two pre-trained models such as AlexNet and VGGNet are used for this study. The SVM and CNN classifier is used to predict the sentiment. Table 2 shows the accuracy obtained by each model and each classifier. It is clear that, in all the datasets, VGG16 achieves higher accuracy than Alexnet model.

Table 2 Accuracy obtained by different models

Classifier	Method	IAPS	Abstract	ArtPhoto	Emotion6	Flickr	FI
SVM	AlexNet	89.15	73.25	74.12	77.68	80.25	85.12
	VGG16	90.11	73.77	74.35	77.90	82.1	89
CNN	AlexNet	89.2	73.5	74.9	80.97	81.5	85.22
	VGG16	92.2	75.91	75.9	81.86	82.92	89.72

Table 3 Accuracy obtained by different Feature Fusion

Method	IAPS	Abstract	ArtPhoto	Emotion6	Flickr	FI
Deep Features+Visual Texture	90.34	75.6	75.8	79.4	81.24	85.8
Deep Features+Complexity	91.9	75.9	78	79.12	81.38	86.23
Deep Features+Colourfulness	91.66	75.9	75.8	78.1	81.3	87.6
Deep Features+Fourier Sigma	91.11	75.89	75.89	79.2	81.44	87.65
Deep Features+Handcrafted Features	92.2	75.91	75.9	81.86	82.92	89.72

Table 4 Analysis of global features and local features

Dataset	Global Feature	Global Feature+Local Deep Features	Global Feature+Local Handcrafted Features	Global Feature+Local Deep Features +Local Handcrafted Features
IAPS	90.33	91.21	91.2	91.9
Abstract	74.67	74.6	74.5	75.11
ArtPhoto	74.6	75.34	74.69	75.54
Emotion6	80.77	81.8	80.79	81.45
Flickr	80.2	81.37	81.19	81.9
FI	85.6	85.4	84.45	87.89

From Table 2, it is also proved that CNN classifier predicts visual sentiment better than SVM classifier. Hence, CNN classifier and VGG16 model are used for further analysis.

5.4 How different features influenced the performance

In this work, deep and hand crafted features are extracted for sentiment analysis. The performance of the proposed method vary depends on the combination of the features. The possible combination of features is tested and the accuracy is analyzed. From various literatures, it is studied that handcrafted features alone could not attain a high accuracy. Hence, those features are combined with deep features. Each handcrafted feature such as visual texture, complexity, colorfulness and Fourier Sigma are combined with deep features and the results are shown in Table 3.

From Table 4, it is analyzed that there is no great improvement in accuracy by including global features. Also, there are only slight differences in the accuracy obtained by any combinations. The accuracy is little higher for Emotion6 dataset than the results obtained in Table 3 of the combination of deep and handcrafted features.

5.5 How β influenced the performance

In Eq. (10), β is used to adjust the weights of the outputs of deep features and handcrafted features. It ranges from 0 to 1. We tested different sets of values of β on validation sets of the Emotion6 dataset to determine the best performance. In Fig. 4, the accuracy first increased and then decreased.

If $\beta = 1$, only deep features was used, but accuracy was the worst; and if $\beta = 0$, accuracy was also poor. This shows that it is effective to combine deep features and handcrafted features. Accuracy reached its peak value of 82.12% when $\beta = 0.8$. This shows that deep

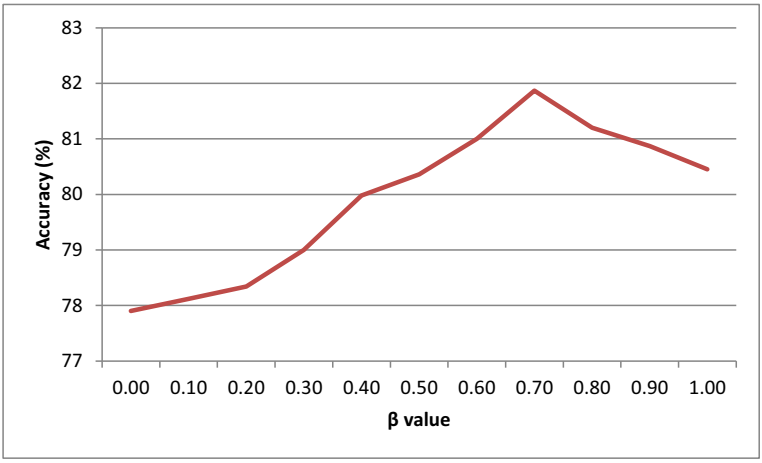

Fig. 4 Performance of proposed method for different values of β

features is more important than handcrafted features. However, the significance of local handcrafted features cannot be ignored. Based on the above results, $\beta = 0.8$ was selected in subsequent experiments.

5.6 Comparison with the state-of-the-art methods

The efficacy of the proposed method can be described only when it is compared with recent methods. Hence, the accuracy of the proposed method is compared with AR Discovery [34], ASE [37], R-CNNGSR [32] methods, Label Distribution Learning (LDL) [33], Global Model & Local Region Model (GM & LRM) [31], Multi-Attentive Pyramidal Model (MAPM) [16] and Weakly Supervised Coupled Network (WSC – Net) [35]. Some of these methods are discussed in Section 2. Table 5 shows the comparison of these methods.

From Table 5, it is understood that the proposed method achieves higher accuracy than other methods for large-scale datasets. For IAPS dataset, the accuracy of the proposed method is lesser than AR Discovery [34] method. In all other datasets, there is a little increase in accuracy. Hence it is studied that the proposed method suits well for large-scale datasets.

Table 5 Accuracy Comparison of the Proposed Method with Recent Methods

Method	IAPS	Abstract	ArtPhoto	Emotion6	Flickr	FI
AR Discovery [12]	92.39	76.03	74.80	–	–	–
ASE [13]	57.82	64.71	53.22	59.94	–	68.87
R-CNNGSR [14]	92.14	75.89	75.02	81.36	–	–
LDL [35]	–	–	–	52.4	–	67.48
Wu et al. [36]	–	–	–	–	72.39	88.84
He et al. [37]	–	–	–	–	80.96	68.13
Yang et al. [38]	–	–	–	–	81.36	70.07
Proposed Framework	92.2	75.91	75.9	81.86	82.92	89.72

6 Conclusion

Visual Sentiment analysis has its own growth in the research field. Deep Learning has a rapid development in any computer vision methods. In this work, Deep learning and hand-crafted methods are combined to classify the sentiment. In order to increase the accuracy, salient objects are identified in each image for analyzing the sentiment. The proposed method is tested on publicly available datasets such as IAPS, Emotion6, Artphoto, Abstract Paintings, Flickr and Flickr & Instagram. Several feature fusion is done to know the performance of the proposed method. The one with all the handcrafted features and the deep features achieves a higher accuracy than the state-of-the-art methods for all the datasets. It is also concluded that the proposed method works good for large-scale datasets. In further research, recent pretrained models such as ResNet can be used for feature extraction. Also, sentiments can be classified hierarchically.

References

1. Ali AR, Shahid U, Ali M, Ho J (2017)High-level concepts for affective understanding of images. In *2017 IEEE winter conference on applications of computer vision (WACV)* (pp. 679-687). IEEE.
2. Andrienko YA, Brilliantov NV, Kurths J (2000) Complexity of two dimensional patterns. Eur Phys J B 15: 539–546. https://doi.org/10.1007/s100510051157
3. Barla, A., Franceschi, E., Odone, F., and Verri, A. (2002). "Image Kernels," in Pattern Recognition with Support Vector Machines: First International Workshop, SVM 2002 Niagara Falls, Canada, August 10, 2002 Proceedings, eds S.-W. Lee and A. Verri (Berlin: Springer), 83–96. https://doi.org/10.1007/3-540-45665-1_7
4. Birkhoff G (1933) Aesthetic measure. Harvard University Press, Cambridge
5. D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in Proceedings of the 21st ACM international conference on Multimedia. ACM, 2013, pp. 223–232.
6. Borth D, Ji R, Chen T, Breuel T, Chang SF (2013)Large-scale visual sentiment ontology and detectors using adjective noun pairs. In proceedings of the 21st ACM international conference on multimedia (pp. 223-232).
7. Braun J, Amirshahi SA, Denzler J, Redies C (2013) Statistical image properties of print advertisements, visual artworks and images of architecture. Front Psychol 4:808. https://doi.org/10.3389/fpsyg.2013.00808
8. V. Campos, A. Salvador, X. Giro-i Nieto, and B. Jou, "Diving deep into sentiment: Understanding fine-tuned cnns for visual sentiment prediction," in Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia. ACM, 2015, pp. 57–62.
9. T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," arXiv preprint arXiv:1410.8586, 2014.
10. Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 415–423.
11. Dalal, N., and Triggs, B. (2005). "Histograms of oriented gradients for human detection," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (Piscataway, NJ: IEEE), 886–893. https://doi.org/10.1109/CVPR.2005.177
12. Fu K, Gu IYH, Yang J (2018) Spectral salient object detection. Neurocomputing 275:788–803
13. Hanjalic A (2006) Extracting moods from pictures and sounds: towards truly personalized tv. IEEE Signal Process Mag 23(2):90–100
14. Hasler, D., and Süsstrunk, S. E. (2003). "Measuring colorfulness in natural images," in Human Vision and Electronic Imaging VIII, eds B. E. Rogowitz and N. P. Thrasyvoulos (Santa Clara, CA: The International Society for Optical Engineering), 87–95. https://doi.org/10.1117/12.477378
15. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in IEEE Conf. Comput. Vis. Pattern Recog., 2016.
16. He X, Zhang H, Li N, Feng L, Zheng F (2019, July) A multi-attentive pyramidal model for visual sentiment analysis. In 2019 international joint conference on neural networks (IJCNN) (pp. 1-8). IEEE.

17. Joshi D, Datta R, Fedorovskaya E, Luong Q-T, Wang JZ, Li J, Luo J (2011) Aesthetics and emotions in images. IEEE Signal Proc Mag 28(5):94–115
18. B. Jou, S. Bhattacharya, and S.-F. Chang, "Predicting viewer perceived emotions in animated GIFs," in ACM Int. Conf. Multimedia, 2014.
19. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.
20. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
21. 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 Proceedings of the 20th ACM international conference on Multimedia. ACM, 2012, pp. 229–238.
22. J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010, pp. 83–92.
23. Mikels JA, Fredrickson BL, Larkin GR, Lindberg CM, Maglio SJ, Reuter-Lorenz PA (2005) Emotional category data on images from the international affective picture system. Behav Res Methods 37(4):626–630
24. M. A. Nicolaou, H. Gunes, and M. Pantic, "A multi-layer hybrid framework for dimensional emotion classification," in ACM Int. Conf. Multimedia, 2011.
25. K.-C. Peng, T. Chen, A. Sadovnik, and A. Gallagher (2005) A Mixed Bag of Emotions: Model, Predict, and Transfer Emotion Distributions. In 2015 IEEE conference on computer vision and pattern recognition (CVPR), pages 860–868. IEEE.
26. Proulx R, Parrott L (2008) Measures of structural complexity in digital images for monitoring the ecological signature of an old-growth forest ecosystem. Ecol Indic 8:270–284. https://doi.org/10.1016/j.ecolind.2007.02.005
27. Redies, C., Amirshahi, S. A., Koch, M., and Denzler, J. (2012). "PHOG-Derived aesthetic measures applied to color photographs of artworks, natural scenes and objects," in Computer Vision – ECCV 2012. Workshops and Demonstrations: Florence, Italy, October 7–13, 2012, Proceedings, Part I, eds A. Fusiello, V. Murino, and R. Cucchiara (Berlin: Springer), 522–531. https://doi.org/10.1007/978-3-642-33863-2_54
28. Rosenholtz R, Li Y, Nakano T (2007) Measuring visual clutter. J Vis 7:1–22. https://doi.org/10.1167/7.2.17
29. M. Solli and R. Lenz, "Color based bags-of-emotions," in Int. Conf. Comput. Anal. Images Patterns, 2009.
30. 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.
31. Wu L, Qi M, Jian M, Zhang H (2019) Visual sentiment analysis by combining global and local information. Neural Processing Letters, pp:1–13
32. Xiong H, Liu Q, Song S, Cai Y (2019)Region-based convolutional neural network using group sparse regularization for image sentiment classification. EURASIP J Image Video Processing 2019(1):1–9
33. Yang J, She D, Sun M (2017) Joint image emotion classification and distribution learning via deep convolutional neural network. In *IJCAI* (pp. 3266-3272).
34. Yang J, She D, Sun M, Cheng MM, Rosin PL, Wang L (2018) Visual sentiment prediction based on automatic discovery of affective regions. IEEE Trans Multimedia 20(9):2513–2525
35. Yang J, She D, Lai YK, Rosin PL, Yang MH (2018) Weakly supervised coupled networks for visual sentiment analysis. In proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7584-7592).
36. You, Q., Luo, J., Jin, H., & Yang, J. (2016). Building a large scale dataset for image emotion recognition: the fine print and the benchmark. arXiv preprint arXiv:1605.02677.
37. Zhan, C., She, D., Zhao, S., Cheng, M.M. and Yang, J., 2019. Zero-shot emotion recognition via affective structural embedding. In *proceedings of the IEEE international conference on computer vision* (pp. 1151-1160).
38. S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014, pp. 47–56.