Hindawi
Scientific Programming
Volume 2022, Article ID 6975490, 13 pages
https://doi.org/10.1155/2022/6975490

Research Article

Visual Emotion Analysis via Affective Semantic Concept Discovery

Yunwen Zhu,[1] **Yonghua Zhu**,[1] **Ning Ge**,[1] **Wenjing Gao**,[1] **and Wenjun Zhang**[2]

[1]*Shanghai Film Academy, Shanghai University, Shanghai 200072, China*
[2]*College of Information Technology, Shanghai Jian Qiao University, Shanghai 201306, China*

Correspondence should be addressed to Wenjun Zhang; 18096@gench.edu.cn

Received 9 November 2021; Accepted 26 February 2022; Published 14 March 2022

Academic Editor: Roberto Natella

With the development of social media, people prefer to express views and share daily life online via visual content, which has led to widespread attention in automatic emotion analysis from images. Capturing the emotions embedded in these social images has always been important yet challenging. In this paper, we propose a visual emotion prediction method that utilizes the affective semantic concepts of an image to predict its emotion. To solve the problems of narrow semantic coverage and low discriminative power of emotions in current semantic concept sets used for visual emotion analysis, we develop a concept selection model to mine emotion-related concepts from social media. Specifically, we propose several selection strategies to build an affective semantic concept set that contains various visual concepts related to emotion conveyance. And they are discovered from affective image datasets and associated tags crawled from websites. To further leverage the discovered affective semantic concepts, we train concept classifiers to predict the concept score of each concept, which are used as the intermediate features to tackle the semantic gap problem for image emotion recognition. Extensive experimental results confirm the validity of the affective semantic concepts and show the improved performance of our method.

1. Introduction

With the widespread adoption of mobile devices and social multimedia platforms such as Flickr, Twitter, and Instagram, people can easily share their daily lives and express their opinions online in the form of texts, images, and videos. Among them, the use of visual media is rising, since images and videos are more intuitive and vivid in conveying moods and sharing personal views. This creates a great demand for automatic visual semantic inference that endeavors to recognize image contents and infer their high-level semantics. In recent years, understanding emotions from visual modality in social multimedia has attracted increasing attention. Automatic emotion recognition of visual contents facilitates the provision of rich practical applications, such as retrieval [1], recommendation [2, 3], entertainment [3], and human behavior estimation [4] etc.

Unlike other computer vision tasks, visual emotion analysis is subjective and culture-dependent, which suffers from a bigger "affective gap" between low-level visual features and high-level emotional responses. Early researches on this issue extracted the low-level visual features related to emotions (e.g., colors, texture, and shapes) from input images [5–7]. However, these typical low-level features cannot effectively conserve rich visual information and fail to model the emotional content of images with various types. Recently, deep-learning methods like convolutional neural networks (CNNs) have been extensively applied to extract high-level features for visual sentiment analysis [8–10]. Nevertheless, because of the complexity and diversity of image emotion recognition, a large amount of labelled training data with small noise is required to achieve good performance. Besides, the CNNs are known as a black-box model in these works without elucidation.

The recent studies enable reliable inference of high-level visual concepts, which provide a more robust midlevel representation for capturing higher-level semantics from images [11, 12]. Leveraging semantic concepts related to visual content plays an important role in visual recognition, not only by providing effective clues for the generation of midlevel feature representation, but also requires fewer training examples. In this situation, we propose to utilize

multiple high-level visual concepts for visual emotion analysis. It is verified that, in most cases, there is a high correlation between visual concepts and emotional reactions [13, 14]. For example, images with objects such as sharks or guns evoke a feeling of fright, while images with babies or flowers convey happiness. Furthermore, in image emotion recognition, each emotional category includes much more diverse visual contents of the image, which results in a large intraclass difference. It is challenging to extract discriminative features that can effectively distinguish one class from another. The multiple high-level semantic information, such as objects, scenes, and actions, can provide more useful information to handle this issue.

In this paper, we define various visual concepts that contribute to emotion conveyance as *affective semantic concepts*: they contain objects, event, places appear in the image and actions observed in the images that are helpful for emotional transfer (see Figure 1). By introducing emotion-related visual concepts instead of low-level visual features and deep-learning features in image emotion analysis, we can reduce the semantic gap and ensure the interpretability of affective analysis to some degree. However, due to various constraints, it is difficult to define such high-level concepts and select emotion-related concepts. Several researchers proposed to extract and define midlevel visual concepts related to sentiment conveyance for visual sentiment prediction. For example, SentiBank [15] crawled a large number of images from social networks using emotion categories as keywords and excavated 1200 adjective-noun pairs (ANPs) as midlevel representations; however, it is observed that ANPs only describe a small portion of the image annotations and ignore the interactive information delivered in visual contents. Unlike SentiBank, the Sentribute [16] considers the influence brought by the scene attributes and face features, which ignores the semantic objects.

Aiming to address the problems of narrow semantic coverage and low emotional discriminability in current semantic concept sets used for visual sentiment analysis, we proposed to mine emotion-related concepts from user metadata. Nowadays, images posted to photo-sharing social platforms like Flickr and Instagram usually include tags or descriptions. Therefore, photo-sharing websites provide us with the opportunity to obtain not only a large number of images freely but metadata tags to save manual labelling. Many previous studies [17–19] have confirmed the feasibility of inferring semantic concepts from the social images and user-generated tags to help further applications. Following this thought of the line, we propose a novel affective semantic concepts discovery method by exploiting shared images and corresponding tags for image emotion classification.

To achieve this, we propose several selection strategies for affective semantic concepts to construct a set of affective concepts consistent with human affective cognition. First, we define four criteria, including semantic modelability, discriminativity, informativeness, and compactness that help to maximize the capability of affective semantic concepts subset from the entire visual concepts. Based on these selection strategies, an effective emotion-related concepts discovery scheme is developed. The semantic concepts in the images are collected with the help of rich web information resources. To further exploit these selected affective concepts, we train linear classifiers to get concept scores for each learned affective semantic concept. Then, we develop a visual emotion classification framework that exploits the affective semantic concepts scores as the intermediate representation for emotion analysis.

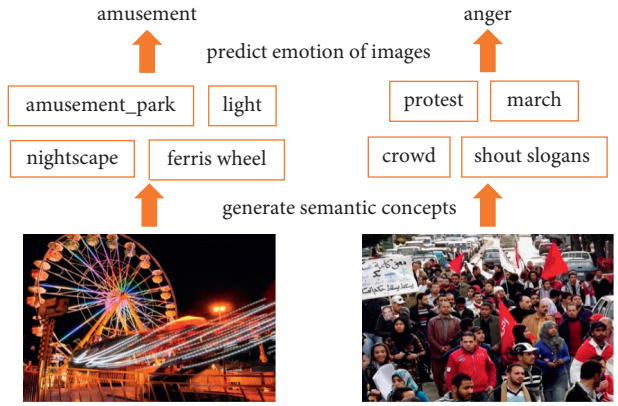

Figure 1: Affective semantic concepts as intermediate feature representations for image emotion analysis.

The main contributions of this paper are summarized as follows:

(i) We propose a concept selection model excavate the affective semantic concepts from affective image datasets and their associated tags crawled from websites. We design concept selection strategies to collect and sort out affective semantic concepts that form relevant, clean and diverse affective concepts for visual emotion analysis.

(ii) Instead of mapping visual features directly to the emotion space, we propose to predict the emotion of images by exploiting the mined affective semantic concepts as the midlevel representation, which not only benefits us with fewer large-scale training examples, but also bridge the affective gap in visual emotion analysis. More important, it is able to explore relationships between the concepts and image emotions through inspecting the emotion classification results, thus improving the interpretability of emotion analysis.

(iii) We conduct both qualitative and quantitative experiments on publicly available datasets and the results demonstrate that the proposed concept discovery model is able to generate accurate semantic concepts. The affective semantic concepts have more significant improvement over the high-level concepts defined by previous approaches on affective analysis, and can be better adapted to visual emotion analysis.

2. Related Works

Visual emotion analysis is more challenging than traditional computer vision tasks due to the higher level of abstraction and subjectivity of visual emotion, which stems from the

semantic gap between low-level features and high-level semantics. Early studies on this issue explored handcrafted features inspired by artistic or psychology theories, including color, texture, SIFT-based shape descriptors, composition and symmetry [6, 20, 21]. However, the handcrafted features are unable to solve the problem of the semantic gap well, as they are most effective on small-scale datasets containing specific styles of images, like artistic images. Recently, deep learning-based features have been widely adopted in image emotion recognition extracting more discriminative features [22]. Nevertheless, deep learning models need a large amount of training data. For this reason, some researchers proposed to fine-tune the CNN models and then extract the deep features for image sentiment analysis [9, 23, 24]. These works focused on mapping visual features directly to emotions, which can be difficult for people to understand how to make decisions.

To bridge the semantic gap between low-level visual features and high-level affective semantic, learning midlevel representations is an important research direction for visual emotion analysis, which can achieve good results with smaller data. Some studies suggest that high-level concepts are crucial elements in capturing the relationships between the images and emotional responses [25]. Borth et al. [15] proposed to utilize Adjectives Noun Pairs (ANPs), which were explored based on strong co-occurrence relationships with emotion tags of web images. Then, SentiBank [15] and DeepSentiBank [26] were constructed to detect ANPs in the images as semantic feature representations to narrow the gap. The obvious drawback of these methods is that they often treat the problem as a collection of binary classification problems, indicating the presence of visual concepts while ignoring the contextual information. In addition, the set of semantic concepts constituted by ANPs covers a limited range of semantics. Ahsan et al. [27] proposed to discover event concept attributes from the website and utilize event concepts as the midlevel semantic features to predict the sentiment of event images, but they only focused on images related to events.

Other studies used a variety of general semantic concepts and pretrained models for image sentiment analysis by leveraging large-scale datasets currently constructed for image recognition, like object detection and segmentation tasks. For example, Yuan et al. [16] presented Sentribute for image sentiment analysis, exploiting 102 scene attributes and face features provided by the SUN dataset [28] as midlevel representation features. Ali et al. [29] built a nonlinear model to correlate the responses of CNN models trained on recognition tasks to emotional classes. Although they introduced concrete visual concepts to guide the generation of midlevel representations, neither covered the whole visual concept space that helped the conveyance of emotions. Some researchers attempted to mine shared images and associated metadata for a collection of semantic concepts that can be employed for computer vision tasks. Yang et al. [19] developed a method to learn visual concept automatically with the help of the webs to comprehend social event in videos. They collected a set of assistant images with corresponding text descriptions from Flickr and extracted compact semantic phrase segments as concepts. Ahsan et al. [17] utilized events as keywords to acquire event-related images and their text descriptions from Flickr, and achieved complex event recognition with few samples by visual representation calculation of segmented phrases and phrase expansion based on natural language models to obtain concepts related to events. Theses researches confirm that it is feasible to discover a set of exploitable semantic concepts from social network images and their associated metadata for further visual analysis tasks.

3. Materials and Methods

In this section, we will introduce our framework for visual emotion analysis, starting from the proposed affective semantic concepts. The overview of the proposed method is illustrated in Figure 2. Our approach consists of three main steps: (1) affective semantic concepts discovery, (2) concept classifiers training, and (3) emotion classification based on affective semantic concepts.

We first define four criteria for concept selection, including semantic modelability, discriminativity, informativeness, and compactness, through analyzing the properties of user tags and affective semantic concepts. Meanwhile, we design a quantitative calculation strategy for each property. According to these strategies, we construct a concept selection model to discover affective semantic concepts from affective image datasets with their tags. In this way, the set of affective semantic concepts with extensive semantic coverage and discriminability is collected. To employ these discovered concepts, we train concept classifiers for each concept to obtain the image feature vectors of concept scores. Once the classifiers are trained, we generate concept scores on the test images to gain the midlevel representations and finally adopt a linear SVM to classify the emotion of the test images.

3.1. Affective Semantic Concepts Discovery. The proposed approach adapts the webly supervised learning methods to discover visual concepts related to emotion conveyance. The image-sharing websites provide us with free access to a large number of images and user-generated tags without extra manual labelling. Some previous works have tried to infer semantic visual concepts from these user-shared images and their noisy tags by utilizing these freely accessible resources. Inspired by these, we propose to extend the idea of mining semantic concepts from the community-contributed resources to emotion-related concepts from affective image datasets and associated tags.

In principle, the user-defined tags include both semantically meaningful concepts that can be observed in or associated with an image and others. The former refers not only to thematic concepts such as specific objects, actions, or events that depict the theme of the image, but also to contextual information like scenes and places where the images were taken, which are obviously important factors in evoking different emotional reactions from the viewers. Additional contextual information includes information

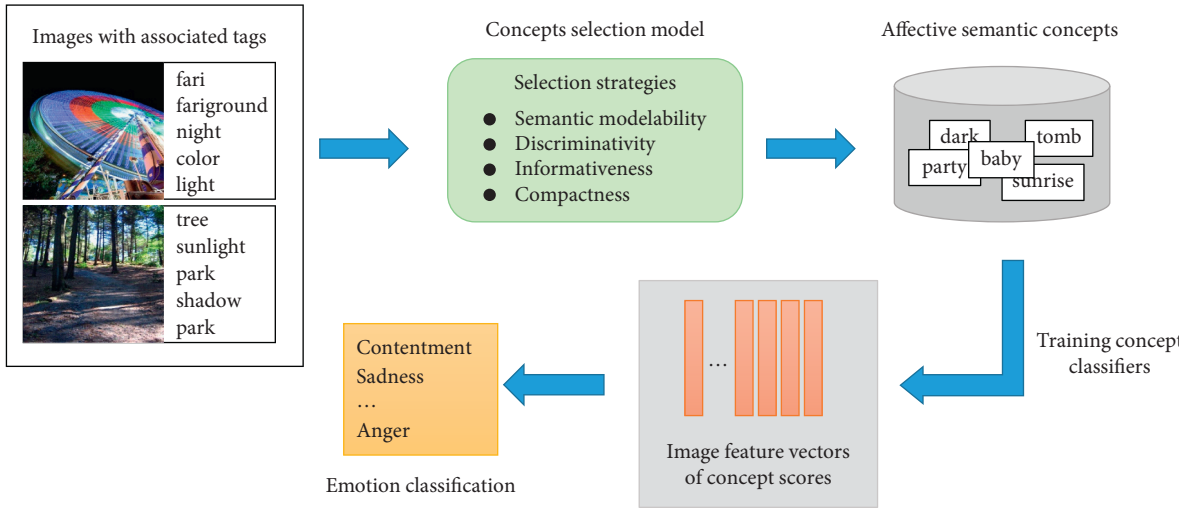

Figure 2: The overview of the proposed method for image emotion classification.

outside the image that informs photographer's intention and the historical origin of the image. Intuitively, emotion is a high-level concept conveyed by visual content, which is closely relevant to high-level visual concepts. Emotional concepts in this paper refer to a small portion of the visual concept set that contributes to emotion conveyance. As shown in Figure 2, this paper aims to discover the semantic concepts that consistent with human cognition from the affective image datasets and their tags crawled from websites through the proposed concept selection model. In this way, a collection of visual concepts that contributes to emotion conveyance, named affective semantic concept set, is constituted.

3.1.1. Preprocessing for User-Generated Tags. We adopt Flickr API provided by social media site Flickr to crawl user-generated tags in the affective dataset. With the aim of ensuring the correctness of concept discovery, we select meaningful semantic tags as candidate concepts of emotional semantics. Due to the existence of irregularities in user-defined tags, preprocessing are operated on the raw user-generated tags, including stop words removal, non-English words removal, and lemmatization.

3.1.2. Selection Strategy. As aforementioned, the user-defined tags can encompass a wide range of concepts that can be observed from the images or relate to the visual contents. Hence, these tags may be noisy, abstract and redundant, whence they cannot be regarded as the reliable solution for image emotion recognition. A refinement and selection process certainly helps to improve the quality of visual concepts. However, owing to various constraints, such a vast selection of emotion-related concepts is hard to accomplish. Considering the properties of affective semantic concepts and characteristics of user-generated tags, we define four criteria that assist us in maximizing the coverage of the emotional concept subset from the entire visual concept set. For this purpose, we first propose quantitative calculations of these criteria and put forward a selection process to mine concepts from community-contributed images and their tags. For convenience, some notations are defined in this section.

Given a set of training dataset $\{x_i, y_i, T_i\}$ including N emotion images. y_i is the class label of the image x_i, and $T_i = \{t_{ij}\}_{j=1,\ldots,q}$ denotes the set of the associated tags of the image. q is the amount of all the tags belonging to the image x_i. We denote the set of all the visual concepts c as C and the selected subset of affective semantic concepts as Θ, respectively. The set of emotion classes is denoted as E, and the emotion class is e. The defined four criteria of the affective semantic concept are introduced in the following.

(1) Semantic Modelability. Owing to the nature of labelling, many tags associated with images are not visually descriptive and are hard to recognize. For example, tags like "Asia" and "2008" tend not to indicate anything meaningful related to the visual content of the annotated image. This makes the utilization of these tags a difficult task. When choosing concepts for learning midlevel representations, the tags with smaller semantic gaps are implicitly favored with respect to the possibility that they would be better modelled. For instance, it is well known that modelling "Europe" is more challenging than modelling "sunset" given lacking valid visual features that can represent such a broad concept of "Europe" with limited training examples. Besides, it contributes more to the description of visual contents. This highlights a crucial demand to develop an efficient way for measuring the semantic gap. Thereby discovering those visual concepts with narrow semantic gaps that are supposed to be assigned high priority for emotional concepts selection. We call this property of emotion-related concepts semantic modelability.

An image belonging to a visual concept with a small semantic gap should have a relatively similar visual appearance. Hence, semantic modelability can be measured by concept-vision consistency. Some previous works regarded the responses of SVM detectors as the measurement of the semantic gap [30] and constructed the similarity matrix of

visual features to quantify the visibility of tags [31]. However, this is unsuitable for our task since the imbalance of sample category. In this paper, we propose to measure the visual consistency of those images for given concepts.

First, for each concept c, we compute the semantic similarity with each image based on their associated tags. The semantic similarity score(c, x_i) between the concept c and the i th image is measured as:

$$\text{score}(c, x_i) = \frac{1}{m} \sum_{t_{ij} \in T_i} d\left(c, t_{ij}\right), \tag{1}$$

where $d(\cdot)$ is the cosine distance between the input vectors. Then, according to the semantic similarity score(c, x_i), we search for each concept's top K images. In order to preserve the most representative images for each concept, we perform K-means clustering based on the similarity of image features on the top K images to get the images set IC containing k_c image clusters. To avoid selecting some images with large differences from the concept, we filter out those clusters whose sizes are smaller than the threshold.

As we stated above, semantic modelability can be measured by concept-vision consistency. Thus, the more visually consistent the image clusters of a concept are, the higher the semantic modelability of the concept is. Motivated by the literature [32], we measure the semantic modelability of a concept by calculating the intracluster dissimilarity and the intercluster dissimilarity of the representative clusters.

We first compute the intracluster similarity by computing the visual dissimilarity among each other in the k th cluster ic_k, which is formulated as

$$\text{intra}_{\text{ds}(\text{ic}_k)} = \frac{1}{\left|\text{ic}_k\right|} \sum_{x_i, x_j \in \text{ic}_k} d\left(x_i, x_j\right), \tag{2}$$

where $|\text{ic}_k|$ represents the number of images in the cluster ic_k and $d(\cdot)$ computes the cosine distance between two items. The lower the value of intra_ds(ic_k) is, the higher the visual consistency of concept c is. To calculate intercluster dissimilarity, we compute the average feature vectors in each cluster. The average feature vector $\overline{x}_{\text{ic}_k}$ in the k th cluster is computed by applying average pooling across each feature dimension. The calculation of intercluster dissimilarity inter_ds(IC) of each concept is as

$$\text{inter}_{\text{ds}(\text{IC})} == \frac{\sum_{\forall \text{ic}_i, \text{ic}_j \in \text{IC}} d\left(\overline{x}_{\text{ic}_i}, \overline{x}_{\text{ic}_j}\right)}{|\text{IC}| * (|\text{IC}| - 1)}. \tag{3}$$

We expect the visual features of each image cluster in the concept with high semantic modelability to be more coherent, which can be measured by Shannon entropy [32].

$$H(c) = -\sum_{i=1}^{K} p(\text{ic}_k) \log\, p(\text{ic}_k), \tag{4}$$

where $p(\text{ic}_k)$ is the probability of each cluster and is computed as $p(\text{ic}_k) = |\text{ic}_k|/|c|$. $|\text{ic}_k|$ is the number of images in cluster ic_k and represents the number of images for concept c. The lower the entropy is, the more coherent clusters the concept has. The concept that has coherent clusters is of higher semantic modelability. Taking the intra- and intervisual dissimilarity into consideration, we modified the entropy to obtain semantic modelability:

$$\text{SM}(c) = \left(1 + \text{inter}_{\text{ds}(\text{IC})}\right) \times \left\{ \sum_{i=1}^{K} p(\text{ic}_k) \log\, p(\text{ic}_k) * \text{intra}_{\text{ds}\left(\text{ic}_k\right)} \right\}. \tag{5}$$

Hence, the total semantic modelability of the selected semantic concept subset SM(Θ) is computed as

$$\text{SM}(\Theta) = \sum_{c \in \Theta} \text{SM}(c). \tag{6}$$

(2) Discriminativity. As described earlier, there is a correlation between the visual concept and emotion conveyance. Nevertheless, in the huge visual concept space, each concept is related to different emotion categories separately. Some concepts such as "building" and "street" occur in multiple affective images with different emotions, which are not discriminative enough for emotion recognition. Since the affective semantic concepts are defined as those visual concepts contributing to the emotion conveyance, the selected concepts must only facilitate the prediction of a small number of emotional classes. We call this property of the selected emotion-related concepts as discriminativity, which can be measured by quantitatively analyzing the correlation between visual concepts and emotions. Inspired by using concepts in event detection [33], we aim to define a quantitative measurement based on Bayes' rule and Shannon entropy to calculate the discriminability of a visual concept to emotion recognition.

In order to get the distribution of different concepts presented in each emotion category, we first estimate the conditional distribution of each concept c when given emotion class e as

$$p(c|e) = \frac{\sum_{i:\ y_i = e} \left(t_{ij} = c\right)}{\sum_{i=1}^{n} \left(y_i = e\right)}, \tag{7}$$

where $\sum_{i=1}^{n} (y_i = e)$ is the number of images belonging to the emotion e. $\sum_{i=1}^{n} (y_i = e)$ is the amount of images for which the associated user tag has concept c and belongs to emotion e. To investigate the relationship between concepts and the whole affective dataset, the marginal distribution of each visual concept is computed as

$$p(c) = \sum_{e} p(c|e) p(e), \tag{8}$$

where the prior probability $p(e)$ is the ratio of the number of images belonging to emotion e to the total number of images. Then, we can get the conditional emotion distribution given a specific visual concept exploiting Bayes' formula:

$$p(e|c) = \frac{p(c|e) p(e)}{p(c)}, \tag{9}$$

$p(e|c)$ represents the discriminative capability of visual concept c to emotion e. To fulfill the requirement of

discriminative capability for the entire emotion classes, the responses of the selected concepts should peak at a small subset of emotion classes. Based on the property that the entropy is a natural measure to quantify the peaked nature of a probability distribution, we adopt the conditional entropy $H(E|c)$ to indicate the discriminative ability of the concept c:

$$\mathrm{ED}(c) = H(E|c) = -\sum_{e} p(e|c)\log_2\ p(e|c). \tag{10}$$

A visual concept with a higher discriminative capability should have higher value of conditional entropy $H(E|c)$. Thus, the discriminability of the selected concept subset is calculated as follows:

$$\mathrm{ED}(\Theta) = \sum_{c\in\Theta} \mathrm{ED}(c). \tag{11}$$

(3) Informativeness. Because of the casualness of user-defined tags, some users tend to describe one image with multiple tags in similar semantics to facilitate photo sharing and retrieval. Therefore, to avoid redundant information from repetitive and similar tags and to ensure the diversity of the subset, we need to consider the interrelation between pairs of classes, implying that the concepts within this space are capable of inferring other concepts effectively. In our work, we refer to this property of emotion-related concept set as informativeness modelled by a strategy like mutual information. The formula is

$$\mathrm{SC}(\Theta;\Omega) = \sum_{c_i\in C}\sum_{c_j\in\Theta} p\left(c_i, c_j\right)\log\left(\frac{p\left(c_i, c_j\right)}{p(c_i)p\left(c_j\right)} + 1\right), \tag{12}$$

where $p(c_i)$ and $p(c_j)$ are the probability of the i th and j th concept appearing in the dataset, respectively. $p(c_i, c_j)$ is the probability of the i th concept and j th concept existing in one image. Since the co-occurrence probability of two concepts may be zero, one is introduced to avoid errors.

(4) Compactness. Too many concepts may lead to a curse of dimensionality, which limits the compactness of the concept space. This work expects to construct a concept space with a finite set of objects. Therefore, the compactness can be easily measured by the size of the concept subset. The higher the number of concepts contained in the set, the lower its compactness score. The compactness of the concept set is computed as

$$L(\Theta) = \frac{1}{|\Theta|}, \tag{13}$$

where $|\Theta|$ denotes the number of the selected concepts.

3.1.3. Solution for Concept Selection. In this section, we propose a concept selection model to obtain the affective semantic concept subset based on four properties stated above: semantic modelability, discriminativity, informativeness, and compactness. The optimization objective is to maximize the emotional concept selection properties of the subset. Hence, the objective function is formulated as

$$\max_{\Theta}\left\{\alpha\mathrm{SM}(\Theta) + \beta\mathrm{ED}(\Theta) + \gamma\mathrm{SC}(\Theta, \Omega) + (1-\alpha-\beta-\gamma)L(\Theta)\right\}, \tag{14}$$

α, β, and γ are the weight parameters used to balance these four properties. As the constructed concept space $|\Theta|$ increases, the first three terms in the equation (14) will also increase accordingly, while the fourth term will decrease. To simplify the issue, we first fix the size of the constructed subset to constant m, that is, $|\Theta| = m$. Then we define U as a diagonal matrix with $u_{ii} = \mathrm{SM}(c_i)$, where c_i is a visual concept appearing in the associated tags. W is also a diagonal matrix with $w_{ii} = D(c_i)$, and V is a symmetric matrix with $v_{ij} = \mathrm{SC}(c_i, c_j)$. s is a vector with $s_i \in \{1, 0\}$ indicating whether the concept c_i is selected to Θ. With the defined variables, the optimization problem is transformed into

$$\begin{aligned} &\max_{s} \sum_{i=1}^{q} s_i^2\left(\alpha u_{ii} + \beta w_{ii}\right) + (1-\alpha-\beta)\sum_{i,j}\left\|s_i - s_j^2 v_{ij}\right\| \\ &\text{s.t.} \quad s_i \in \{1, 0\}, \sum_{i}^{q} s_i = m, \end{aligned} \tag{15}$$

where α and β are weight parameters to control the trade-off between properties that are set by cross-validation. We restrict the s_i to be a real value between 0 and 1 and the optimization problem can be further reformulated as a quadratic programming problem:

$$\begin{aligned} &\min_{s} \quad \mathbf{s}^T\mathbf{A}\mathbf{s} \\ &\text{s.t.} \quad 0 \le s \le 1,\ 1^T\mathbf{s} = m, \end{aligned} \tag{16}$$

where $\mathbf{A} = -\{\alpha U + \beta W + (1-\alpha-\beta)(D^V - V)\}$, D^V is a diagonal matrix with $d_{ii}^V = \sum_{j=1}^{n} v_{ij}$ and 1 is an all-one vector. The vector $\mathbf{s}$ is received by solving this objective function, and the corresponding m concepts with the highest scores are selected to form a subset of affective semantic concepts.

3.2. Training Concept Classifiers. To leverage these relevant emotional concepts selected by the concept discovery approach proposed above, this section introduces our method for training concept classifiers. Given a set of discovered affective semantic concepts $C = \{c_1, c_2, \ldots, c_m\}$, we use each concept as a keyword in Microsoft Bing to search the top 100 images. The retrieved images are applied to train concept classifiers. For all its retrieved images of each concept, we adopt the pretrained AlexNet [34] model to extract the image features. We extract the CNN features on each image and feed them into the linear classifiers to generate the concept scores. Assuming the feature vector of each image I is denoted as $f_I = \{f_i\}_{i=1}^{m}$, where m is the overall number of concepts, f_i is the score produced for the concept c_i classifier and the feature vector f_I is a series of all concept classifier scores produced on the image I.

3.3. Emotion Classification Based on Affective Semantic Concepts. Based on the trained concept classifiers, we can concatenate all concept classifier scores generated from the

image. These obtained concept score responses can thus be used as concept representation features to predict the emotion of an image. Given an affective image, we first generate the affective semantic concept scores based on the concept classifiers. Finally, we use these concept scores as the mid-level representations to accomplish image emotion classification through a linear SVM classifier.

4. Experimental Results and Analysis

In this section, we perform extensive experiments to investigate the performance of our method for image emotion classification. We carry out qualitative experiments to demonstrate the proposed concept discovery strategy can effectively mine affective semantic concepts. Furthermore, we also compare our method with state-of-the-art methods for image emotion classification.

4.1. Experiment Setups

4.1.1. Datasets. We perform our experiments on five widely used affective datasets for emotion classification and sentiment prediction, including Flickr and Instagram (FI) [9], EmotionROI [35], Flickr [36], Instagram [36], and Twitter [15]. The FI dataset is collected from Flickr and Instagram websites with 23308 weakly labelled web images. Each of them is labelled with Mikels' eight emotion classes. The EmotionROI dataset is also collected from Flickr, containing 1980 images classified into six emotion categories (anger, disgust, fear, happy, sadness, and surprising). In addition, we also assess the performance of our method on several positive and negative classes' datasets for binary classification, including Flickr, Instagram, and Twitter datasets. They contain 60745, 42856, and 603 web images from Flickr, Instagram, and Twitter social networks.

For purpose of implementing the affective semantic concept mining approach based on user metadata information, it is necessary to gain the images with reliable emotion labels and user tags. Based on this, we choose the FI dataset for emotion-related concepts discovery, in which all images are crawled from Flickr and Instagram websites. Most images are rich in user metadata, which is in line with the resources required for the concept discovery proposed in this paper. In our experiment, we only adopt the part of the dataset corresponding to the Flickr resource. Specifically, we apply the Flickr API provided by the Flickr website to crawl user-generated tags in the affective dataset. Then, preprocessing is operated on the raw user-generated tags, including stop words removal, non-English words removal, and lemmatization. Moreover, to ensure the availability of tags, we remove the user-generated tags containing less than 30 associated images. Table 1 shows the statistics of the image dataset with its tags after preprocessing. The filtered 894 user-generated tags constitute the initial concept space. Examples for each emotion category with its associated tags are shown in Figure 3.

4.1.2. Implementation Details. Our experiments begin with mining the emotion-related concepts by the proposed concepts selection strategies. The parameter settings involved in the proposed concept selection model are as follows. For the parameters of the semantic modelability, we select the top 150 images for each concept to generate a set of representative images, and the number of image clusters is predefined as $k_c = 10$. The threshold of the number of images contained in the image cluster is set to 5. According to the importance of different emotion properties, the weight parameters involved in the optimization objective function of the concept selection model are set as $\alpha = 0.45$, $\beta = 0.50$ to control the proportion of semantic modelability and discriminativity in the objective function. They are set by grid-search for cross-validation following [40].

We train the concept classifiers by extracting the activations of CNN layer 7 as visual features for all the training images exploiting the Caffe [37] deep learning framework, since employing the features from the pretrained CNN on ImageNet for various visual recognition tasks has shown state-of-art performance [41]. The CNN model is built on AlexNet [34] architecture pretrained on ImageNet [38] and Places [39] datasets. Then, we use the trained concept classifiers to generate the semantic concept scores and concatenate the concept scores to form the feature vector for each image, which can be regarded as the mid-level representations. Finally, we apply the publicly available LibSVM for image emotion classification. The FI dataset is split randomly into 80% for training, 5% for validation, and 15% for testing. For the Flickr dataset and Instagram dataset, we randomly sample the same number of images for each class following the same configuration in [42], which are split randomly into around 90% for training and 10% for testing. The remaining datasets are all randomly divided into 80% training set and 20% testing set.

4.1.3. Baselines. We compare our method against several baselines, including methods using low-level features, midlevel semantic features as well as high-level concept features. For the methods based on low-level features, we compare with the principle-of-art features (PAEF) designed by Zhao et al. [21]. We adopt the simplified version to extract 27-dimensional features and utilize the LibSVM classifier for image emotion classification. For the methods based on midlevel representation features, we compare with SentiBank [15] and the pretrained DeepSentibank [26]. SentiBank utilizes 1200-dimensional binary features detected by a concept detector library. While the DeepSentiBank is based on the CNN model to extract 2089-dimensional features and then perform image sentiment classification by a fully connected layer. For the methods based on high-level concepts, Ali et al. [29] proposed to use the pretrained AlexNet model on the ImageNet to extract 1000 dimensional objects features and on the Places dataset to obtain 365-dimensional scenes features. Inspired by this literature, we design three variants to compare the performance with our method, including a method based on object concept (HLCs-object), a method based on scene concepts

Table 1: The statistics of the dataset used for affective semantic concept discovery.

Original images	Original tagged images	Tagged images	Tag types	The average number of tags
23308	20557	13534	894	6.2

Figure 3: Examples of images with different emotion categories and their user-generated tags.

(HLCs-scene) and a method fusing object and scene concepts (HLCs).

4.2. Qualitative Experiment. To evaluate our selected affective semantic concepts, we first show the qualitative results in terms of the mined concepts by the proposed selection model.

Table 2 lists the top 30 concepts with the highest semantic modelability scores to better understand and verify the validity of defined semantic modelability. We calculated the average of five times scores as the final score to reduce errors. As shown in Table 2, most of the top 30 concepts belong to visual cognitive semantic, such as objects like "baby" and "cat," scenes with more visual consistency like "park" and "fair," and simple actions like "mourning" and "cry." Moreover, we also observe that proper nouns like "califonia", which are not relevant to the description of visual content, and abstract concepts like "happy" have lower semantic modelability scores. According to general knowledge, it can be concluded that these concepts with high scores describe specific contents and have more visual consistency. These results are consistent with the definition of semantic modelability stated above, confirming the feasibility of the proposed quantitative calculation for semantic modelability.

Figure 4 displays the top five concepts with the highest scores of discriminativity excluding abstract nouns in eight emotion categories. It can be seen that these concepts almost conform to the important elements of human emotion perception. For example, concepts with high scores in the sadness emotion category include actions that express sadness, such as "cry," "loneliness," and "mourning." Concepts with higher discriminability scores are easier to distinguish from other emotion categories. As shown in Figure 4, the score of the concept "roller coaster" is 1.0 in the amusement category, which indicates it only appears in affective images with amusement emotion. Consequently, it coincides with the property of discriminativity defined in this paper.

Table 2: The top 30 concepts with the highest scores of semantic modelability.

Road	Rain	Baby	Protest	amusement_rides
Sunshine	Park	Snow	Cat	Mourning
Insect	Head	Party	Fair	Animal
Grass	Reflection	Painting	Dog	House
Sunrise	Sea	Eye	Cry	Portrait
Broad	Walk	Nightlife	Grave	Downtown

Additionally, we also show the selected affective semantic concepts when the size of concept space m is set to 300. Table 3 reports the top 30 concepts with the highest scores calculated by the proposed concept selection model. We divide them into object, scene and action semantics based on the image semantic hierarchy description. The results imply that almost all the selected concepts with high scores belong to the cognitive semantics, which proves the property of semantic modelability. Meanwhile, these concepts possess a certain degree of discriminability. For example, "tear" and "cry" are consistent with the concepts of human emotion cognition that conveys sadness.

From the qualitative experimental results, we can conclude that the affective semantic concepts not only accord with the visual concepts that contribute to the emotion conveyance in human cognition, but also satisfy the visual discriminability. Besides, the affective concept set contains fewer concepts with high semantic similarity, which ensures the diversity of the concept set and avoids the repetition of redundant information. These results demonstrate the availability of the proposed concepts selection model for affective semantic concepts discovery.

4.3. Quantitative Experiment. To evaluate the performance of our method for image emotion classification, we conduct comparison experiments to compare our method with the above-mentioned baselines on five public affective datasets. Table 4 reports the performances of the baselines along with our approach measured by the accuracy metric. The accuracy is the ratio of the number of correctly classified test

Figure 4: The top 5 concepts with the highest scores of discriminativity in different emotion categories.

Table 3: Top 30 concepts with highest scores in the affective semantic concepts set.

Semantic hierarchy	Concepts
Objects	thrill_ride, ferris_wheel, roller_coaster Grave, baby, cat, insect, boardwalk, mask
Scenes	Fair, nightlife, marathon, protest, birthday, journey Picnic, valley, amusement_park, cemetery Downtown, park, dawn, rain, snow, party, dark
Actions	Demonstration, racing, mourning, swimming

Table 4: Emotion classification accuracy (%) on five datasets for different methods.

Methods	FI	EmotionROI	Flickr	Instagram	Twitter
PAEF	46.13	34.84	66.61	64.17	67.51
SentiBank	49.23	35.24	69.26	66.53	65.93
DeepSentiBank	51.54	42.53	70.16	67.13	70.23
HLCs-object	49.10	36.36	69.35	67.03	70.33
HLCs-scene	53.48	45.45	73.52	72.67	75.96
HLCs	54.80	42.59	74.55	73.98	77.81
Ours	**61.55**	**49.95**	**75.13**	**75.02**	**78.87**

samples to the total number of test samples. It shows that the handcrafted features perform worse than other approaches. This result can be explained that there is still a large semantic gap between the low-level features, inclusive of color and texture features designed based on the principle of art and the affective semantic. In contrast, other models bridge the semantic gap to some extent by using the generated semantic features as an intermediate representation. Among them, SentiBank and DeepSentiBank utilize sentiment-specific concepts ANPs as the mid-level features. They cover a limited range of semantic concept space and the sentiment-specific concepts may be relevant for general images shared on websites but ignore the contextual information like scenes and events, which leads them to fail to achieve satisfying results. As for the HLCs method, they apply the large number of objects provided by ImageNet and scenes provided by Places datasets to cover larger semantic concept space that results in outperforming other comparison methods on each dataset. Particularly, the HLCs method that fused objects and scenes information achieves better performance. However, the HLCs methods show a disadvantage that the semantic concepts defined by the visual recognition task may lead to insufficient discriminative and the semantic redundancy of the generated semantic concepts, which affects the accuracy of image emotion classification.

From the results shown in Table 4, we can observe that our method achieves the best accuracy in all datasets. The prediction accuracy of our method reaches 61.55% on FI

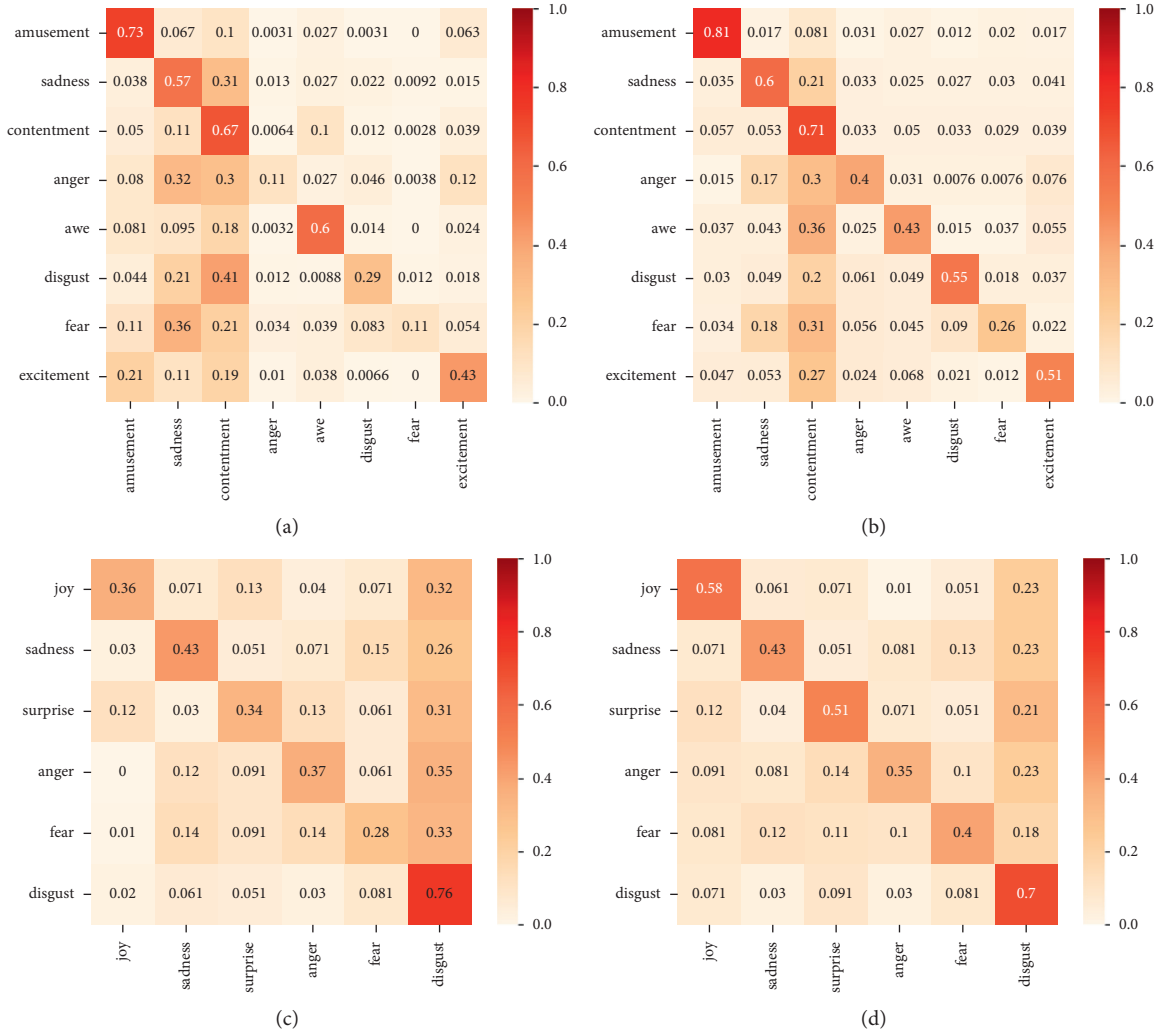

Figure 5: Confusion matrices for the HLCs and our method on FI and EmotionROI datasets. (a) HLCs on FI dataset. (b) Our method on FI dataset. (c) HLCs on EmotionROI dataset. (d) Our method on EmotionROI dataset.

dataset and 49.95% on EmotionROI dataset, which surpasses the state-of-art HLCs method by over 6.75% on FI dataset and 7.36% on EmotionROI dataset, respectively. Meanwhile, our method gains improvement of 0.58%~1.06% on other binary sentiment datasets. This implies that our proposed method performs better in dealing with multiclass emotion analysis, the reason is that our method introduces affective semantic concepts as intermediate representations. These excavated semantic concepts are more emotional discriminative and have wider coverage, thus, they benefit more for multi-class emotion datasets. For binary sentiment datasets, the prediction accuracy of our method is also improved, which proves that our proposed method is able to more effectively tackle the problems that existed in visual sentiment analysis methods based on midlevel features. Our method has superior performance on small-scale datasets compared to other baseline methods, which further demonstrates the advantages of our method that introducing affective semantic concepts can decrease the requirements of training data. In summary, applying the affective semantic concepts that conform to emotion properties as intermediate representations of images, our method shows significant advantages over the semantic-based image understanding approaches.

To further evaluate the ability of the learned concepts to distinguish between different emotional categories on multiclass image affective datasets, we visualize the confusion matrices of HLCs and our method on FI and EmotionROI datasets. In the confusion matrices, the value on the diagonal indicates the ratio of images classified to the correct emotional category. The predicted result is visualized based on the color shades, where the darker the color indicates the more samples classified into that category. As illustrated in Figure 5, the color of the squares on the diagonal of our method is darker than that of HLCs method, which means our method outperforms the HLCs and shows more ability to distinguish between various emotion categories. In the FI dataset, the HLCs and our method have less incorrectness on *amusement*, *sadness*, *contentment,* and *awe* emotions, while the HLCs tends to incorrectly classify the *anger* and *fear*

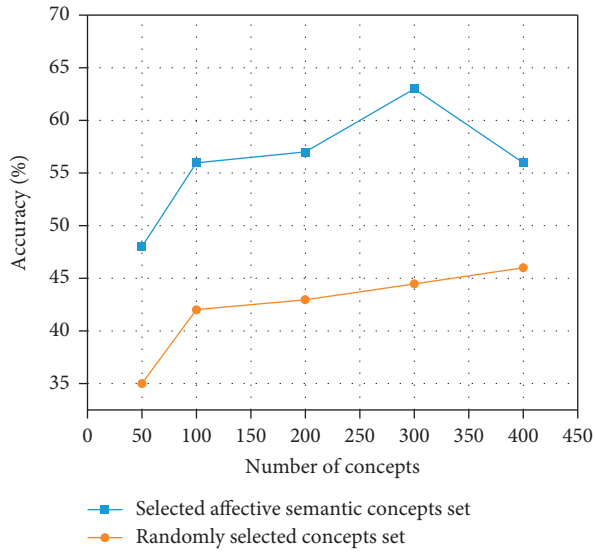

Figure 6: Accuracy results with the different number of concepts.

emotions. This confusion is because the high-level visual concepts used as intermediate representations in HLCs method have narrow semantic coverage and low discriminative power of emotions. In contrast, our method utilizes the mined affective semantic concepts that contribute to emotion conveyance, which can better enhance emotional discrimination.

4.4. Parameter Analysis. As stated above, our method mines the emotion-related concepts as the midlevel semantic representations by constructing an affective concept set. The size of the constructed concept set may influence the quality of the learned midlevel representation. To explore the optimal value for the size of the concept set, we test the different values of m to ensure the best performance for emotion classification. Meanwhile, we compare with a random concept set that randomly selects a given number of concepts from the set of user-generated tags. The results illustrated in Figure 6 shows that the midlevel semantic features generated by the constructed concept set for image emotion classification perform better than the randomly selected concept set, which further demonstrates the contribution of the mined emotion-related concepts to improving the prediction accuracy of image emotion classification. Additionally, as the value of m increases, the accuracy of emotion classification first increases and then decreases. The peak value reaches for accuracy when $m = 300$. The reason lies in two aspects. On the one hand, the increase in the size of the concept space will cause difficulties for recognition. On the other hand, the concepts within this space can reasonably infer the semantic information of other concepts by calculating the mutual information. As the best performance is obtained with $m = 300$, we choose it in all our experiments.

5. Conclusion

In this paper, we propose a novel image emotion prediction method based on affective semantic concept discovery. We mine emotion-specific concepts from the affective image datasets and their user-generated tags crawled from social websites to predict the emotions of images. To sort out and acquire a clean, relevant and diverse affective concepts set, we define the selection strategies and propose a concept selection model by combing the properties of concepts and user tags. Then, we train concept classifiers to learn the concept scores as the intermediate representations for emotion recognition, which shows the strength of our method in narrowing the semantic gap. Moreover, leveraging a concept-based intermediate representation can benefit us by requiring fewer labelled training data and enhancing the interpretability of visual emotion analysis. The qualitative experiments demonstrate the availability of the proposed concept discovery method. And the quantitative experiments conducted on five public affective datasets show that our method achieves the best accuracy in all datasets, which proved the superiority of the proposed visual emotion classification method. However, the method proposed in this paper exists some disadvantages, such as it detects each concept independently by the concept classifiers. We are aware that our work is just one of many steps in visual emotion recognition task. In the future, how to predict the concept labels of images based on the discovered affective semantic concepts and combine them with deep learning networks will be a possible extension. Additionally, this concept-based visual emotion can be introduced to aesthetics analysis as well.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Key Research and Development Plan of China (no. 2017YFD0400101).

References

[1] X. Yao, D. She, S. Zhao, J. Liang, Y.-K. Lai, and J. Yang, "Attention-aware polarity sensitive embedding for affective image retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1140–1150, Seoul, Republic of Korea, October 2019.

[2] Y. Zhang, G. Lai, M. Zhang, Yi Zhang, Y. Liu, and S. Ma, "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis," in *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 83–92, New York, NY, USA, July 2014.

[3] Y.-Y. Chen, T. Chen, T. Liu, M. Liao, and S.-F. Chang, "Assistive image comment Robot-A novel mid-level concept-based representation," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 298–311, 2015.

[4] H. Lin, J. Jia, L. Nie, G. Shen, and T.-S. Chua, "What Does social media Say about Your stress?" *IJCAI*, pp. 3775–3781, 2016.

[5] V. Yanulevskaya, J. C. van Gemert, K. Roth, A. K. Herbold, N. Sebe, and J. M. Geusebroek, "Emotional valence categorization using holistic image features," in *Proceedings of the 2008 15th IEEE International Conference on Image Processing*, pp. 101–104, San Diego, CA, USA, October 2008.

[6] 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*, pp. 83–92, New York, NY, USA, Octomber 2010.

[7] B. Li, S. Feng, W. Xiong, and W. Hu, "Scaring or pleasing: exploit emotional impact of an image," in *Proceedings of the 20th ACM International Conference on Multimedia*, pp. 1365-1366, New York, NY, USA, Octomber 2012.

[8] J. Xu, Z. Li, F. Huang, C. Li, and S. Y. Philip, "Visual sentiment analysis with social Relations-Guided Multiattention networks," *IEEE Transactions on Cybernetics*, 2020.

[9] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: the fine print and the benchmark," *Proceedings of the AAAI Conference on artificial intelligence*, vol. 30, no. 1, 2016.

[10] W. Zhang, X. He, and W. Lu, "Exploring discriminative representations for image emotion recognition with CNNs," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 515–523, 2019.

[11] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1785, Miami, FL, USA, June 2009.

[12] P. Blandfort, T. Karayil, J. Hees, and A. Dengel, "The Focus-Aspect-Value model for predicting subjective visual attributes," *International Journal of Multimedia Information Retrieval*, vol. 9, no. 1, pp. 47–60, 2020.

[13] J. Vogel and B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *International Journal of Computer Vision*, vol. 72, no. 2, pp. 133–157, 2007.

[14] C. Tzelepis, Z. Ma, V. Mezaris et al., "Event-based media processing and analysis: a survey of the literature," *Image and Vision Computing*, vol. 53, pp. 3–19, 2016.

[15] D. Borth, R. Ji, T. Chen et al., "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of the 21st ACM International Conference on Multimedia*, pp. 223–232, New York, NY, USA, October 2013.

[16] J. Yuan, S. Mcdonough, Q. You et al., "Sentribute: image sentiment analysis from a mid-level perspective," in *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, pp. 1–8, New York, NY, USA, August 2013.

[17] U. Ahsan, C. Sun, J. Hays, and I. Essa, "Complex event recognition from images with few training examples," in *Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 669–678, Santa Rosa, CA, USA, March 2017.

[18] J. Tang, S. Yan, R. Hong, and G.-J. Qi, "Tat-Seng Chua"Inferring semantic concepts from community-contributed images and noisy tags," in *Proceedings of the 17th ACM International Conference on Multimedia*, pp. 223–232, New York, NY, USA, October 2009.

[19] X. Yang, T. Zhang, C. Xu, and M. S. Hossain, "Automatic visual concept learning for social event understanding," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 346–358, 2015.

[20] H. Zhang, E. Augilius, T. Honkela et al., "Analyzing emotional semantics of abstract art using low-level image features, Advances in Intelligent Data Analysis X," in *Proceedings of the International Symposium on Intelligent Data Analysis*, pp. 413–423, Porto, Portugal, October 2011.

[21] 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*, pp. 47–56, New York, NY, USA, November 2014.

[22] T. Rao, X. Li, and M. Xu, "Learning multi-level deep representations for image emotion classification," *Neural Processing Letters*, vol. 51, no. 3, pp. 2043–2061, 2020.

[23] S. Zhao, Y. Ma, Y. Gu et al., "An end-to-end visual-audio attention network for emotion recognition in user-generated videos," pp. 303–311, 2020.

[24] V. Campos, B. Jou, and X. Giró-i-Nieto, "From pixels to sentiment: fine-tuning CNNs for visual sentiment prediction," *Image and Vision Computing*, vol. 65, pp. 15–22, 2017.

[25] S. Nasim, M. Rehan, and N. Sabahat, "Emotional understanding of an image by applying high-level concepts on image parts," in *Proceedings of the 2020 IEEE 23rd International Multitopic Conference (INMIC)*, pp. 1–5, Bahawalpur, Pakistan, November 2020.

[26] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "Deepsentibank: visual sentiment concept classification with deep convolutional neural networks," 2014, https://arxiv.org/abs/1410.8586.

[27] U. Ahsan, M. De Choudhury, and I. Essa, "Towards using visual attributes to infer image sentiment of social events," in *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 1372–1379, Anchorage, AK, USA, May 2017.

[28] G. Patterson and J. Hays, "Sun attribute database: discovering, annotating, and recognizing scene attributes," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2751–2758, Providence, RI, USA, June 2012.

[29] A. R. Ali, U. Shahid, M. Ali, and J. Ho, "High-level concepts for affective understanding of images," in *Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 679–687, Santa Rosa, CA, USA, March 2017.

[30] C. Lang, J. Feng, and Y. Zheng, "Towards a universal detector by mining concepts with small semantic gaps," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11312–11320, 2012.

[31] M. A. Kastner, I. Ide, F. Nack et al., "Estimating the imageability of words by mining visual characteristics from crawled image data," *Multimedia Tools and Applications*, vol. 79, no. 25, pp. 18167–18199, 2020.

[32] J. W. Jeong, X. J. Wang, and D. H. Lee, "Towards measuring the visualness of a concept," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pp. 2415–2418, New York, NY, USA, Octomber 2012.

[33] L. Wang, Z. Wang, Y. Qiao, and L. V. Gool, "Transferring deep object and scene representations for event recognition in still images," *International Journal of Computer Vision*, vol. 126, no. 2, pp. 390–409, 2018.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[35] K. C. Peng, A. Sadovnik, A. Gallagher, and T. Chen, "Where do emotions come from? predicting the emotion stimuli map," in *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*, pp. 614–618, Phoenix, AZ, USA, September 2016.

[36] M. Katsurai and S. I. Satoh, "Image sentiment analysis using latent correlations among visual, textual, and sentiment views," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2837–2841, Shanghai, China, March 2016.

[37] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Technical report, Department of Computer Science, National Taiwan University, Taipei, Taiwan, 2003.

[38] Y. Jia, E. Shelhamer, J. Donahue et al., "Caffe: convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 675–678, New York, NY, USA, November 2014.

[39] A. S. Razavian, H. Azizpour, and J. Sullivan, "Stefan Carlsson. "CNN features off-the-shelf: an Astounding baseline for recognition," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 806–813, Columbus, OH, USA, June 2014.

[40] O. Russakovsky, J. Deng, H. Su et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[41] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proceedings of the 27th Advances in Neural Information Processing Systems (NIPS)*, pp. 487–495, Cambridge, MA, USA, December 2014.

[42] D. She, J. Yang, M. M. Cheng, Y.-K. Lai, P. L. Rosin, and L. Wang, "WSCNet: weakly supervised coupled networks for visual sentiment classification and detection," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1358–1371, 2019.