J Inf Process Syst, Vol.17, No.6, pp.1138~1156, December 2021
https://doi.org/10.3745/JIPS.01.0082

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

# A Survey on Image Emotion Recognition

Guangzhe Zhao[*], Hanting Yang[*], Bing Tu[**], and Lei Zhang[*]

## Abstract

Emotional semantics are the highest level of semantics that can be extracted from an image. Constructing a system that can automatically recognize the emotional semantics from images will be significant for marketing, smart healthcare, and deep human-computer interaction. To understand the direction of image emotion recognition as well as the general research methods, we summarize the current development trends and shed light on potential future research. The primary contributions of this paper are as follows. We investigate the color, texture, shape and contour features used for emotional semantics extraction. We establish two models that map images into emotional space and introduce in detail the various processes in the image emotional semantic recognition framework. We also discuss important datasets and useful applications in the field such as garment image and image retrieval. We conclude with a brief discussion about future research trends.

## Keywords

Emotion Semantics, Image Emotion Recognition, Image Feature Extraction, Machine Learning

## 1. Introduction

With the popularity of Internet of Things (IOT) infrastructure, smartphones and wearable devices, most people generate a large amount of data every day. These data contain variable information, and 70% of that information is in the form of images, as vision is the primary way that humans access outside information. In addition to conveying information, images also affect the emotions of people viewing them. Therefore, effective analysis of emotions in images is of great significance for many applications.

One typical application is garment image emotion recognition. The emotional semantics in images affect online shoppers and determine what clothes they buy. Another common application is image retrieval. A practical image retrieval system can search for what users are looking for in a large database both quickly and precisely. However, there are obvious challenges in image emotion analysis. For example, the ambiguity of predefined features and different understandings of the image by individuals make it difficult to build a universal model. However, emotional semantics is the highest level of semantics, and they involve the human cognitive model, cultural backgrounds and aesthetic standards. As the emotion evoked by an image will determine a person’s decision-making, scholars hope to develop an appropr iate model that can describe and express the human emotional reactions caused by observing images and describe images using the semantics of subjective emotion.

Manuscript received October 11, 2019; first revision April 9, 2020; accepted May 14, 2020.

**Corresponding Author:** Bing Tu (tubing@hnist.edu.cn)

[*] College of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China (zhaoguangzhe@bucea.edu.cn, 18810903925@163.com, leizhang@bucea.edu.cn)

[**] School of Information Science and Technology, Hunan Institute of Science and Technology, Yueyang, China (tubing@hnist.edu.cn)

A traditional method called content-based image analysis relies on artificial text annotation for each image and extracts low-level features like color, texture, shape, and spatial relationships from each object in an image. Keywords are then determined based on the features and used as a basis for later retrieval or classification. These methods incur a heavy workload for the image emotion recognition system due to their inefficiency, and additionally, the system usually cannot accurately identify the emotion in an image. One reason for this poor performance is that when people judge the similarity of an image, their judgements are not necessarily based on the similarity of the image's visual characteristics.

Wang and He [1] has proposed the system framework of emotional semantics and summarized emotional semantics from three primary aspects: emotion-based image feature extraction, the establishment of emotional space, and mapping from the realization of an image feature to emotional space. Inspired by their work, we investigated these three aspects separately. This paper is organized as follows: Sections 2, 3, and 4 introduce the above three aspects, Section 5 presents our emotion semantic recognition experiment, Section 6 specifies the application of image emotion retrieval, Section 7 introduces some benchmark datasets for image emotion recognition, and Section 8 provides a conclusion.

## 2. Existing Device Discovery Schemes

Although there are many factors that affect the human emotional judgment of an image, the image features that affect people's feelings regarding images are primarily color, texture, shape and contour.

### 2.1 The Image Color Feature

The color characteristics of images have an important influence on people's emotions. In general, color hue, saturation and lightness are three primary factors that affect people's emotions. Second, different color saturations evoke different feelings. High-purity colors have a strong impact and cause strong feelings. Low-purity colors are also called neutral colors, and they cause soft and neutral feelings. Although they are dark, they are full of charm [2-5].

Lightness intuitively reflects an image's brightness, and it is defined as the arithmetic mean $\mu$ of the red, green, and blue color coordinates. People may feel uncomfortable when they see a high-brightness picture. In contrast, if the brightness is too low, the content in the picture will not be visible.

For color feature extraction, constructing a color histogram is a typical method that counts the color distribution in an image. The principal focus is the number of each type of pixel of the colors that appear. It can be applied in a variety of color spaces and has the advantages of translation, rotation and scaling invariance [6].

In recent years, with the establishment and development of deep learning theory, scholars began to use convolutional neural networks (CNNs) to extract low-level visual features from raw images as an automatic process. Yao et al. [7] trained three different CNN architectures on a dataset with sentiment tags. By learning on 15,000 scene images, he suggested that the CNN demonstrates better performance on specific prediction problems.

### 2.2 Image Shape and Contour Characteristics

The existing device discovery schemes have lengthy device discovery times because the 48-bit UID search range ($2^{48}$) tends to induce very deep tree levels along the binary search tree. Thus, we propose a new device discovery scheme called partition-based device discovery.

Different shapes cause different feelings; For example, geometric shapes impart a simple, clear mechanical and cold sense. Commonly used methods for describing shape and contour features are based on border and region-based methods. The better description methods among them are the Fourier descriptors and invariant moment method. In recent years, some new methods have emerged, such as the finite element matching method and wavelet transform description method.

Colombo et al. [8] executes Hough transform on the image to derive the line slope of a histogram, from which the horizontal and vertical lines of the ratio are determined as a shape feature. The use of linear slopes represents the different emotional theories, which can be used to determine the image emotional semantics together with other features. Jamil et al. [9] extracted structural features from outdoor photographic images in terms of shape, such as straight lines, long lines, the joint endpoint (co-terminations), L-shaped, U-shaped, parallel lines, and polygons. Then, the eigenvectors were formulated accordingly. The eigenvectors can be used to separate large-scale man-made objects in an image, such as buildings, towers, bridges, and other buildings, from natural landscape images, which demonstrates the importance of the shape feature in image content judgment. Geometric shape extraction based on emotion analysis is relatively simple. It is sufficient to extract the general shapes in the images.

In addition to analyzing the color, texture, and shape features of an image, there are methods for studying other image features. For example, Yu et al. [10] preliminarily defined the parameters of image perceptual information from the geometrical features, gray histogram, frequency band change and image comprehension change according to the perceptual characteristics of the human eye. Then, a method for describing the image perception information was proposed. In [11,12], color and texture features were used synthetically, and a method of image retrieval based on the comprehensive fuzzy histogram was proposed by using the fuzzy features of human subjective sensation. Hayashi and Hagiwara [13] combined N spectral features and color features as the image features mapped to emotional words. Wang et al. [14] used 150 large wavelet coefficients as image features to classify image emotion.

## 2.3 Image Texture Features

Different materials have different textures, which cause different psychological feelings. Therefore, it has been widely used in emotion recognition [14-18]. Texture is an important characteristic of an image and is difficult to model. Even now, there is no well-defined definition. The first classic method for modeling the mathematical characteristics of texture is the co-occurrence matrix method [19]. The so-called gray level co-occurrence matrix, $M(D_x, D_y)$ mathematically represents the joint frequency distribution of two gray-scale pairings, $(D_x, D_y)$, in the image so that it can reflect the gray-scale texture space dependencies. The disadvantage of the gray level co-occurrence matrix method is that some of its texture attributes (such as entropy) do not have corresponding visual content.

In recent years, with the establishment and development of deep learning theory, scholars began to use the CNN to extract low-level visual features from raw images as an automatic process. Yao et al. [7] trained three different CNN architectures on a dataset with sentiment tags. By learning on 15,000 scene images, he suggested that the CNN demonstrates better performance on specific prediction problems.

## 3. The Establishment of Emotional Space

The result of image emotion semantic extraction is that an image is mapped to an emotional space so that each image corresponds to a point in that emotional space. Additionally, each point in the emotional space represents a semantic description of some emotion. The distance between points corresponds to the emotional distance of an image. Herein, emotional information in the form of people's feelings after seeing the image is described in the form of semantics. A corresponding quantitative comparison can then be carried out.

Wundt's emotional three-dimensional theory, Schlosberg's facial expression scale, and Plutchik's three-dimensional model of emotion are representative of emotional determinations and measurement studies [20,21]. Stallman [20] argues that emotions generally require three dimensions to be able to describe them effectively: pleasure-unpleasantness, tension-relaxation and, excitement-calm. American psychologist Schlosberg and his collaborators considered that emotional performance has three dimensions: happy-unpleasant, attention-rejection and, activation level. Plutchik [21] formulated a theoretical model of human emotional dimensions, i.e., the three-dimension emotional model (Fig. 1). It is formulated as a cone in which the vertical dimension represents the intensity of the emotion and the angle on the circle represents the degree of similarity between emotions. Eight fan-shapes represent eight basic emotional dimensions, namely: happiness, trust, fear, surprise, sadness, nausea, anger and expectation. In each basic emotional fan, the strongest emotion is at the top, and it grows weaker as it moves toward the bottom. As an example, a fan-shape may demonstrate a journey from grief and sadness to melancholy. The similarity and polarity among primitive emotions are demonstrated by the arrangement of the fan-shapes. The opposite corner on the graph contains the opposite and conflicting emotion. The emotions close to each other are complementary. The emotions in the white region of the figure are a composite of two basic emotions.

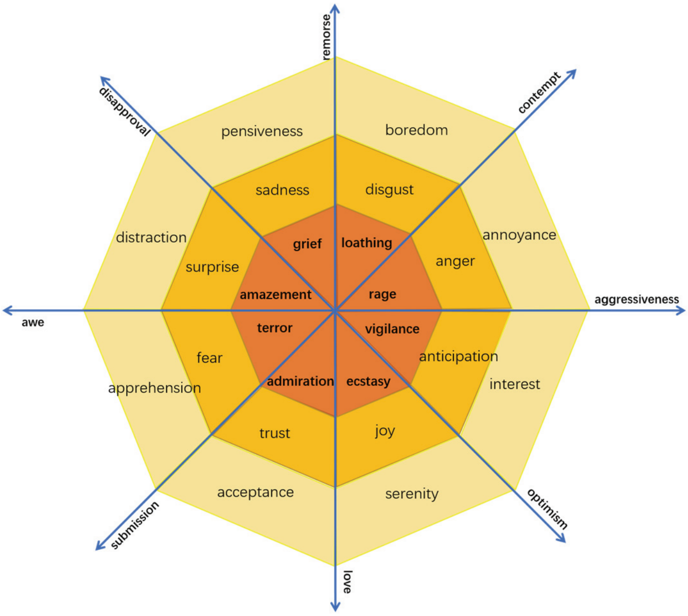


**Fig. 1.** Plutchik three-dimensional model of emotion [21].

Recently, most of the methods used to identify users' subjective emotions are based on an emotion survey regarding the image by the testers. The difference in the specific survey methods also affects the results of the emotional space. Generally speaking, there is not a unified model for the study of emotion in images because of individual subjectivity, empiricism, vagueness, and regional and cultural differences as well as gender differences [22]. We use two different emotional spatial models built by by Zhang et al. [12] and Wang et al. [14] as examples.

## 3.1 Emotion Spatial Model One

Zhang et al. [12] formulated a fabric image emotional semantic model using a large number of surveys and tests. After statistical analysis, emotional semantics can be reduced to seven-dimensional emotional languages [11], which are represented by three factors $f_1, f_2, f_3$ [8]:

$$Strong-soft = 0.412f_1 + 0.720f_2 + 0.393f_3 \tag{1}$$

$$Warm-cool = 0.468f_1 + 0.278f_2 + 0.739f_3 \tag{2}$$

$$Gorgeous-minimalist = 0.773f_1 + 0.565f_2 + 0.237f_3 \tag{3}$$

$$Elegant-plain = 0.932f_1 + 0.229f_2 - 0.164f_3 \tag{4}$$

$$Publicity-quiet = 0.626f_1 + 0.697f_2 + 0.245f_3 \tag{5}$$

$$Heavy-liftweight = 0.067f_1 + 0.193f_2 - 0.942f_3 \tag{6}$$

$$Rich-pure = 0.203f_1 + 0.892f_2 + 0.177f_3 \tag{7}$$

Parameters $f_1, f_2, and\ f_3$ can account for 90.488% of the total variance in the seven semantic variables. Then, according to the above equations, we can derive $f_1, f_2, and\ f_3$ using the following equations, where the capital character represents the seven emotional languages mentioned above:

$$f_1 = -0.133S + 0.149W + 0.371G + 0.771E + 0.130P - 0.245H - 0.421R \tag{8}$$

$$f_2 = +0.394S - 0.236W + 0.028G - 0.421E + 0.269P - 0.126H + 0.784R \tag{9}$$

$$f_3 = +0.039S + 0.485W - 0.115G - 0.107E - 0.120P + 0.773H - 0.157R \tag{10}$$

### 3.1.1 Analysis of factor 1 characteristics

From the factor load matrix, it can be observed that the elegant-plain and gorgeous-minimalist are close to factor $f_1$. According to color psychology theory, the emotional feelings of "elegant–plain," "gorgeous–minimalist" are closely related to image color saturation, temperature and contrast. According to the experiments, images with a large factor 1 have low saturation, large average color tone and low contrast. Therefore, comprehensive characteristics including the saturation, cold and warm fuzzy histogram and color contrast value of images are adopted to evaluate the first factor of the image characteristics, which is described as follows:

**Saturation description**

$$\mu(low\ saturation) = \begin{cases} 1, & C > 10 \\ \frac{27-C}{17}, & 10 \le C \le 27 \\ 0, C > 27 & \end{cases} \tag{11}$$

$$\mu(medium\ saturation) = \begin{cases} \dfrac{C-10}{17}, & 10 \le C \le 27 \\ \dfrac{51-C}{24}, & 27 \le C \le 54 \\ 0, else & \end{cases} \tag{12}$$

$$\mu(high\ 1143aturation) = \begin{cases} 0, & C < 27 \\ \dfrac{C-27}{24}, & 27 \le C \le 51 \\ 1, C > 51 & \end{cases} \tag{13}$$

**Description of the well-being of color**

The 50 hue angle is used as the warmest angle, and the cosine of the hue value and an angle with a value of 50 is used to describe a color's hue. The membership function of well-being is as follows [23]:

$$\mu(x) = \begin{cases} cos\ (x - 50^{\circ}), & x \in [0,140]\ or\ [320,360] \\ 0, & else \end{cases} \tag{14}$$

$$\mu(x) = \begin{cases} cos\ (x - 230^{\circ}), & x \in [140,320] \\ 0, & else \end{cases} \tag{15}$$

There are six combinations of saturation and color tone membership functions, namely, low saturation and warm; middle saturation, warm high saturation and warm; and low saturation and cool. They formulate a six-dimensional saturation-hue of the color fuzzy histogram.

**Color contrast description**

The equation for calculating the color contrast is:

$$ab_{contrast} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(a_i - \overline{a})^2 + \left(b_i - \overline{b}\right)^2]} \tag{16}$$

Combining color contrast with a saturation-well-being fuzzy histogram, we can obtain a seven-dimensional feature histogram that can be treated as a basis for describing factor 1 [23].

### 3.1.2 Analysis of factor 2 characteristics

From the factor load matrix, it can be seen that two factors are more closely related to the second factor: rich-pure and strong-soft. Color contrast is calculated using Eq. (16). The richness of an image's color can be directly seen from the distribution of gray of the image histogram. A histogram distribution with a few gray values indicates that the image's color richness is lower. When the grey values are evenly distributed, it indicates that the image's color is rich. Image color diversity can be represented by a 256-dimensional gray-scale histogram. The relationship between the image and factor 2 can be represented by the image color diversity plus the color contrast, for a total of 257-dimensional features.

### 3.1.3 Analysis of factor 3 characteristics

From Eq. (11), we can observe that factor 3 is closely related to heavy-lightweight and warm-cool. According to the theory of color psychology, the emotional feeling of the above meanings is related to the image's smoothness and the warmth of the colors. While smooth and cool colors provide people with a light and cool feeling, rough and warm colors provide a thick, warm feeling.

The smoothness and roughness of an image are represented by the gray level co-occurrence matrix. The gray level co-occurrence matrix can be derived by calculating the energy (ASM), contrast (CON) and entropy (ENT).

$$ASM = \sum_{i=1}^{k} \sum_{j=1}^{k} (G(i,j))^2 \tag{17}$$

$$CON = \sum_{n=0}^{k-1} n^2 \{ \sum_{|i-j|=n} G(i,j) \} \tag{18}$$

$$ENT = \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{G(i,j)}{1+(i+j)^2} \tag{19}$$

The color tone of an image is still calculated based on the average warm and cold tone based on Eqs. (16), (17), (18), and (19). Therefore, factor 3 can be explained using image hue plus the three feature values of roughness for a total of four-dimensional features.

However, there are shortcomings in this experiment. First, the total number of samples used in the above experiment is only 6, and the ability to verify the experimental results is limited.

Second, there is a gap between the experimental results and practical applications. Although the semantic categories used in the experiment are statistical results, they are different from the semantic search field (fresh, retro, etc.) commonly used in online shopping. The actual application is often not a separate analysis of a standard fabric image but a need to extract the image from the fabric image for analysis. There are still many difficulties that must be solved to realize this application.

## 3.2 Emotion Spatial Model Two

This method includes three steps for establishing emotional space [24]: one is to collect mental adjectives from different image databases; the second is to conduct cognitive psychological experiments in which respondents evaluate an image, collect data and establish a user emotional database; the third is the factor analysis method, which is adopted to analyze the data in the database and establish emotional space.

**Table 1.** Adjective group (for landscape pictures) [24]

| | |
|---|---|
| 1. like-do not like | 10. tidy-messy |
| 2. beautiful-ugly | 11. clear-blur |
| 3. coordinated-uncoordinated | 12. quiet-noisy |
| 4. romantic-not romantic | 13. impressive-plain |
| 5. comfortable-uncomfortable | 14. relaxed-depressed |
| 6. hot-cold | 15. transformative-monotonic |
| 7. warm-cool | 16. vital-desert |
| 8. light-dark | 17. broad minded-narrow minded |
| 9. soft-hard | 18. warm colors-cool colors |

In the first step, two groups of adjectives (18 pairs of landscape pictures and 15 pairs of clothing pictures) are selected (Table 1). With 1 as the perfect score, each adjective is scored using five grades (0.0, 0.25, 0.5, 0.75 and 1.0). Then, the user evaluation data is collected and the user emotional database is established. Using 206 ($M = 206$) samples from 1,300 landscape images and 200 ($M = 200$) samples from 1,486 images, 180 university students (120 men and 60 women) adopted factor analysis to process these images. Details regarding the process are described as follows:

Let user *K* evaluate the nth adjective of the image m as $z_{mnk}$ and obtain the average value $v_{mn}$ according to Eq. (20). Then, user *K* normalizes it according to Eq. (21) to achieve matrix U (*M*×*N*).

$$v_{mn} = \frac{1}{K}\sum_{K=1}^{K} z_{mnk} \tag{20}$$

$$v_{mn} = \frac{v_{mn} - v_n}{s_n} \tag{21}$$

wherein

$$\overline{v}_n = \frac{1}{M}\sum_{m=1}^{M} v_{mn}$$

$$s_m^2 = \sum_{m=1}^{M} (v_{mn} - v_n)^2$$

Factor analysis is performed on matrix X as follows:

$$U' = AF + \epsilon \tag{22}$$

The dimension of the adjective space of the original *M* dimension is reduced to *L* and the *L* dimension of the political space is formed. Then, the *m*-th row $f_m = (f_{m1}, f_{m2}, \dots, f_{mL})$ of the F matrix corresponds to the coordinate of Image m in emotion space. The *n*-th row of the *A* matrix corresponds to the coordinate of adjective *n* in the *L*-dimensional space, $a_n = (a_{n1}, a_{n2}, \dots, a_{nL})$.

The method described above also has its shortcomings. The experimental group of users were primarily 20-year-old college students. In [22], the authors recognize that aesthetic cognitive models involve human cognition as well as other factors. Therefore, using a single experimental group to make aesthetic emotional judgments is not comprehensive, and the results of the entire experiment are limited to this group of college students. To provide application value, a wide range of experiments are required.

## 4. Image Emotion Semantic Recognition Framework

The process of emotion recognition establishes the mapping from an image feature space to emotional space. The linear mapping method was previously adopted. Later, to effectively evaluate the impact of perceived subjective experience on human emotions and understand and model that subjective information, a model method is established by training users to establish the relationship between image features and emotions. This model primarily includes fuzzy theory, neural networks, the genetic algorithm and support vector machines.

Dai and Yu [25] obtained good results by classifying images' texture features using the BP neural network. They also studied fuzzy neural network classifiers based on fuzzy perceptrons in detail. As a powerful tool for establishing nonlinear mapping, neural network methods have been widely used in the study of image emotion [26]. Lee and Park [27] developed a color pattern-based emotion evaluation system based on the multilayer feedback neural network, which showed higher accuracy than the previous linear system evaluation. Hayashi and Hagiwara [13] used the back propagation rule neural network to establish the relationship between image features and impression words, reaching 78.8% accuracy.

Fuzzy theory can use human-like thinking and judgment method quantification to adapt to the process of computer processing. According to fuzzy theory, people's experiences and common sense can be expressed in a form suitable for computers. It also can establish human feeling and language expression models [28]. Lee and Park [27] used an adaptive fuzzy system to establish the relationship between color legends and emotion evaluation. Compared with neural network methods and linear mapping methods, it has higher accuracy. Palani and Venkatalakshmi [29] used a fuzzy clustering method for texture image classification.

The support vector machine (SVM) method is applied to the retrieval and retrieval of image retrieval systems and constructs a classifier suitable for a user's needs [30]. Sun et al. [31] used SVM classification in a road vehicle detection system. Zhao et al. [32] proposed the concept of "principles-of-art." Different from the previous "element-of-art" feature extraction method, SVM classification and support vector regression (SVR) are used. These methods have achieved better results than other machine learning methods. The SVM is becoming a new research hotspot. It was preceded by research regarding neural networks and will promote the development of machine theory and technology.

Deep learning has gradually been applied to image emotion recognition. In the field of computer vision, the CNN has achieved great success in facial recognition and object recognition. Many researchers have begun to explore the effect of deep learning on emotion recognition. You et al. [33] built a super-large dataset of over three million emotions, which is 30 times the size of the previous dataset. Corchs et al. [34] categorized these emotions using the most popular methods (including the CNN) of the day as criteria for the following research and performed fine-tuning of different CNN layers to capture global and local feature information. They achieved the best recognition performance compared to manual features [35] and with the development of deep learning techniques, they certainly had a great impact on image emotion recognition.

The new computational model must include two processes: (1) automatic establishment of mapping between the image features and the impression words using a module with a learning kernel and (2) interactive modification of the model through an external process as well as dynamic creation of a conceptual space [35-37]. Based on the above model, Bianchi constructed a new system (K-agent). In the second part's external process, she used cognitive maps to support user and model interaction. Compared with existing image retrieval systems on the Internet (such as AltaVista), the results show that the K-agent greatly reduces the retrieval time and can meet the needs of users. Wang et al. [38] also used interactive evolutionary computation (IEC) to evaluate the similarity between the search result and the user's subjective space in image emotional retrieval, where the subjective emotion was integrated into the evolutionary process. Then, an effective result was obtained.

Wang and He [1] summarized the current research situation at home and abroad as well as the basic framework of emotional semantic extraction, which is shown in Fig. 2. It can be seen from the figure that the extraction of emotional semantics includes image feature extraction, emotional space establishment, image emotion recognition and image emotion semantic extraction.

Zhang et al. [12] used machine learning to establish the mapping relationship between low-level features and emotional semantics for fabric images. This mapping relationship was established using a SVM. The experiment used 60 sheets of fabric images as samples. The training and image semantic automatic recognition processes are shown in Figs. 3 and 4, respectively. The sample images comprised 60 images with different types of fabric. The test subject expressed seven pairs of emotional descriptions for each image [12].

The semantic recognition framework proposed by Wang et al. [14] states that each image (factor score) and each adjective (factor load) can be regarded as a vector of the L-dimensional emotional space. In addition to the samples in the database, the rest of the landscape images and clothing images can be mapped to the emotional space using emotional notes.

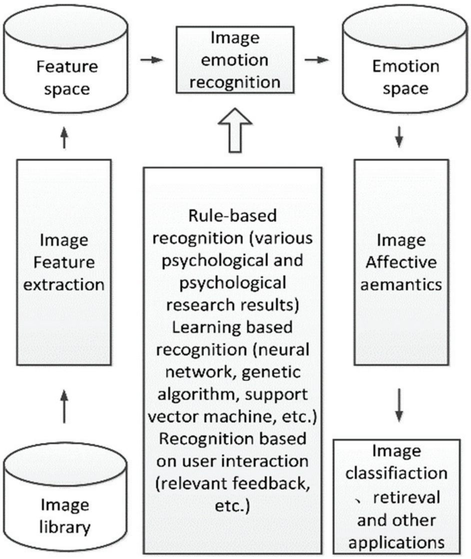

Feature space
Image emotion recognition
Emotion space
Image Feature extraction
Rule-based recognition (various psychological and psychological research results) Learning based recognition (neural network, genetic algorithm, support vector machine, etc.) Recognition based on user interaction (relevant feedback, etc.)
Image Affective aemantics
Image library
Image classifiaction 、retrieval and other applications

**Fig. 2.** Basic framework of affective semantics [1].

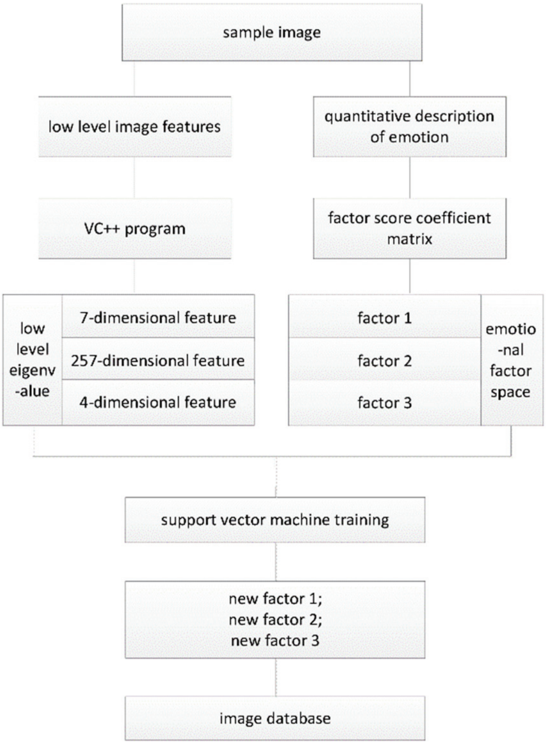


**Fig. 3.** Fabric image sample support vector machine training and storage process [12].

Each image in the annotated image library corresponds to a vector of the emotional space. Each adjective also corresponds to a vector of the emotional space through factor analysis. These vectors are used for emotional retrieval of images.

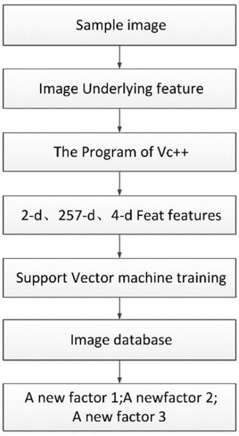

Sample image
Image Underlying feature
The Program of Vc++
2-d、257-d、4-d Feat features
Support Vector machine training
Image database
A new factor 1;A newfactor 2;
A new factor 3

**Fig. 4.** Emotional semantic recognition processes of a new fabric image [12].

## 5. Image Emotion Semantic Recognition Experiment

Zhang et al. [12] conducted their experiments according to the framework shown in Figs. 3 and 4. In the experiments, the SVM regression prediction method was used to establish the emotion recognition model for a fabric image, and then the machine learning conducted emotion prediction.

### 5.1 SVM Regression Prediction

$(X_i, y_i), \mathrm{i} = 1,2, \dots n, x_i \in R^n$ is a linear fit with the linear regression function [39] $f(x) = \omega x +$ input, where $x_i$ is the input, $y_i$ is the output, and we must choose ω and b. The penalty function is a measure of the error in the learning or training model during the learning process, and it is usually pre-determined before the model is learned. The loss function corresponding to different learning problems is also different, and the same learning model with different loss functions are also different. The standard support vector machine uses the insensitivity function.

Then, the problem is transformed into an optimization problem to minimize the objective function:

$$R(\omega, \xi, \xi^*) = \frac{1}{2}\omega \cdot \omega + C\sum_{i=1}^{n}(\xi + \xi^*) \tag{23}$$

where ξ and $\xi^*$ are relaxation factors when category error exists; ξ and $\xi^*$ are greater than 0, otherwise they are 0; $c$ is a constant; and C > 0 is a penalty coefficient that denotes the degree of punishment when the error is greater than $\varepsilon$. This problem is a convex quadratic optimization problem. Its Lagrange function is:

$$\begin{aligned} L = \frac{1}{2}\omega \cdot \omega + C\sum_{i=1}^{n}(\xi + \xi^*) - \sum_{i=1}^{n}\alpha_i[\xi_i + \varepsilon - y_i + f(x_i)] \\ - \sum_{i=1}^{n}a_i^*[\xi_i + \varepsilon - y_i + f(x_i)] - \sum_{i=1}^{n}(\xi_i\gamma_i + \xi_i^*\gamma_i^*) \end{aligned} \tag{24}$$

where $(\alpha_i, \alpha_i^*) \geq 0$ and $(\gamma_i, \gamma_i^*) \geq 0$, i=1,2,...,$n$, is the Lagrange multiplier. The function $L$ is minimized for $\omega, b, \xi_i, \xi_i^*$ and maximized for $\alpha_i, \alpha_i^*, \gamma_i, \gamma_i^*$. Using the Lagrangian dual function, the maximization function is:

$$W(\alpha,\alpha^*)=\frac{1}{2}\sum\nolimits_{i=1,j=1}^{n}(\alpha_i-a_i^*)(\alpha_j-a_j^*)(x_i\cdot x_j)+\sum\nolimits_{i=1}^{n}(\alpha_i-a_i^*)y_i \\ -\sum\nolimits_{i=1}^{n}(\alpha_i+a_i^*)\varepsilon \tag{25}$$

In fact, the above constraint is also a quadratic programming problem. For a standard support vector, the value of *b* is typically calculated for all standard support vectors and then averaged, which is expressed as:

$$b=\frac{1}{N_{nsv}}\{\sum_{0<\alpha_i<c}[y_i-\sum\nolimits_{x_j\in sv}(\alpha_j-a_j^*)\,K(x_i,x_j)-\varepsilon] \\ +\sum\nolimits_{0<a_i^*<C}[y_i-\sum\nolimits_{x_j\in SV}(\alpha_j-\alpha_k^*)\,K(x_j,x_i)-\varepsilon]\} \tag{26}$$

Therefore, the linear fitting function obtained from the sample points $(x_i,y_{i)}$ is:

$$f(x)=\omega\cdot x+b=\sum\nolimits_{i=1}^{n}(\alpha_i-\alpha_i^*)x_i\cdot x+b \tag{27}$$

The basic idea of nonlinear SVR is to map the input vector to a high-dimensional feature space using a pre-determined nonlinear mapping. Then, linear regression is performed in this high-dimensional space and the effect of non-linear regression on the original space is obtained.

First, the input vector *x* is mapped into the high-dimensional feature space *H* by mapping $R^n\rightarrow H$, and $f(x)=\omega x+b$ is fitted to $(x_i,y_i)$, i = 1,2, ..., *n*. The quadratic programming function becomes:

$$W(\alpha,\alpha^*)=\frac{1}{2}\sum\nolimits_{i=1,j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\cdot(\Phi(x_i)\cdot(\Phi(x_j))+\sum\nolimits_{i=1}^{n}(\alpha_i-\alpha_i^*)y_i \\ -\sum\nolimits_{i=1}^{n}(\alpha_i+\alpha_i^*)\varepsilon \tag{28}$$

Eq. (28) involves the dot productivity of the high dimensional feature space in which $(x_i)\cdot(x_j)$, and the function μ is unknown and high-dimensional. SVM theory only considers the dot product of the high-dimensional feature space $K(x_i,x_j)=(x_i)\cdot(x_j)$ instead of using the function $K(x_i,x_j)$ as a kernel function, and the kernel function should be calculated as the dot product of a high-dimensional feature space. There are many types of kernel functions. Common kernel functions include the linear kernel function, polynomial kernel function, radial basis function (RBF), two-layer neural network kernel function and Fourier kernel function. In contrast, this paper uses the rad basis function:

$$K(x,x')=exp\,(-\frac{|x-x'|}{2\sigma^2}) \tag{29}$$

Therefore, Eq. (25) is transformed into:

$$W(\alpha,\alpha^*)=\frac{1}{2}\sum\nolimits_{i=1,j=1}^{n}(\alpha_i-a_i^*)(\alpha_j-a_j^*)K(x_i\cdot x_j)+\sum\nolimits_{i=1}^{n}(\alpha_i-a_i^*)y_i \\ -\sum\nolimits_{i=1}^{n}(\alpha_i+a_i^*)\varepsilon \tag{30}$$

The expression of the nonlinear fitting function is:

$$f(x) = \omega \cdot \Phi(x) + b = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x, x_i) + b. \tag{31}$$

## 5.2 Image Recognition Experiment Based on Emotion

The low-level characteristics of the fabric image are extracted and calculated using Visual Studio 2008. On this basis, SVM regression analysis is performed using LIBSVM [40]. The sample is a subjective test of 60 sheets of fabric images. The low-level features of the extracted image samples are organized into an input vector for the LIBSVM system. Then, the samples are regressed using the LIBSVM regression method, where the LIBSVM type is EPSILON SVR and its kernel function is the RBF kernel, which has an accuracy of 0.0001. The SVM in the RBF kernel is primarily located in the penalty coefficient C under certain circumstances and is used to find the optimal δ. Therefore, for different penalty coefficient C, the SVM training model may be different, as shown in Fig. 5.

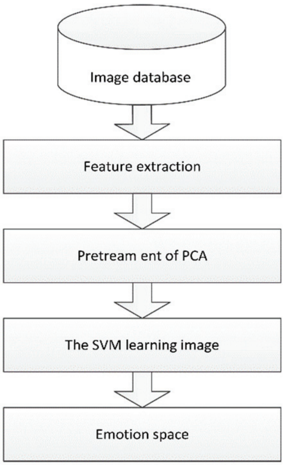
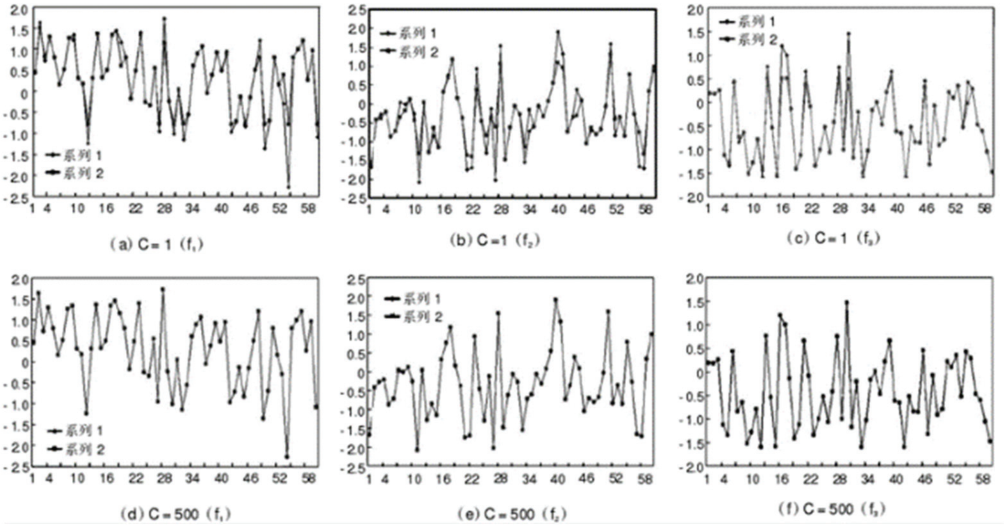


**Fig. 5.** Emotional semantic annotation framework.



(a) C= 1 (f₁) (b) C=1 (f₂) (c) C= 1 (f₃)

(d) C=500 (f₁) (e) C=500 (f₂) (f) C=500 (f₃)

**Fig. 6.** Different penalty coefficients with different training effects.

By comparison, when C is within a certain range, such as C = 500, the accuracy of the regression is already very high since the two lines are essentially coincident. This result therefore demonstrates the rationality and effectiveness of the selected lower-level features.

According to the factors $f_1, f_2,$ and $f_3$ in the training model, for any new fabric image, we calculate its three factors based on the flow chart shown in Fig. 4. Then, we calculate the seven emotional descriptions for the value and add the new image to the database. All of these steps have been realized using programming.

# 6. Image Emotion Retrieval

Image recognition can be further applied to image emotion retrieval. We establish an emotional space and collect the users' emotional information using cognitive psychological experiments and factor analysis. Using the SVM algorithm, image feature semantic annotation is formed and the mapping of the image feature space to the emotion space is established. On this basis, a method of emotion retrieval is proposed and is primarily used in garment and landscape image emotion retrieval.

Emotional semantic annotation of images uses an image's simple features to describe the image's content, establish the feature space of the image, express the emotion semantics of the image using adjectives, and establish the emotional space using cognitive psychology experiments and factor analysis. This process establishes the mapping of the image's low-level feature space to the user's high-level emotional space. It also memorizes the user's emotional characteristics through learning and it automatically comments on images that have not been evaluated.

## 6.1 Image Emotion Retrieval Process

Through emotional semantic annotation, each image is mapped into a vector of the emotional space. Hence, we realize emotional retrieval in an emotional space. The retrieval process is shown in Fig. 7 [41].

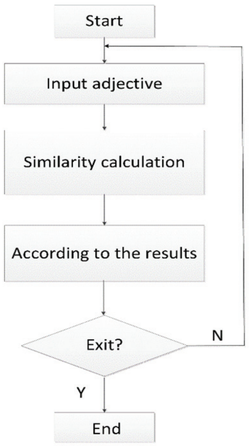


**Fig. 7.** Emotion image retrieval flow chart [41].

First, the user reports their registration information, which includes the user's gender. The user enters an adjective to begin the search. The system calculates the similarity between each image in relation to the adjective and returns the most recent images as the result. The similarity is calculated as follows:

Let $f_m = (f_{m1}, f_{m2}, \dots, f_{mL})$ and $a_n = (a_{n1}, a_{n2}, \dots, a_{nL})$ be the coordinates of image m and adjective $n$ in the emotional space, respectively. Then, the similarity between image $m$ and adjective $n$ is,

$$d_{mn} = \frac{a_n \cdot f_m}{|a_n||f_m|}. \tag{32}$$

## 6.2 Search Result

According to gender, the users participating in the experiment are classified in a simple manner. The users' evaluation data for the images are stored in the database as user emotion information and the data are analyzed using factor analysis. The cumulative contribution matrix of the load matrix is 87.78% and 87.19%, respectively. The 18-dimensional adjective for the original landscape picture is reduced to four dimensions. Moreover, the 15-dimensional adjective for the clothing picture is reduced to seven dimensions.

The first 130 samples from the 206 landscape images are used as the learning samples and the last 76 as the test samples. The output cosine of the SVM and the standard output are used for the similarity measurement. Table 3 presents the results of the experiment. The similarity rate is defined as follows:

For each image, the similarity between the original data, the output data from the SVM and each adjective is calculated and sorted according to the similarity value. The similarity ratio for the first n adjectives in the two rankings is calculated and is called the coincidence rate. It indicates the degree of agreement between the results of the image annotation and the original evaluation.

From Table 2, the average coincidence rate is approximately 70%. Therefore, an SVM is a good method for annotating images with emotion semantics.

**Table 2.** Average coincidence rate of the system

| | **Study samples** | **Experimental samples** |
|---|---|---|
| Average similarity rate | 0.8703 | 0.7086 |
| Coincidence rate (1) | 76.00% | 61.29% |
| Coincidence rate (1–2) | 68.00% | 58.06% |
| Coincidence rate (1–3) | 70.67% | 62.37% |
| Coincidence rate (1–4) | 78.00% | 68.55% |
| Coincidence rate (1–5) | 82.40% | 72.26% |

Twenty students, aged approximately 20 years old, used the system for emotional image retrieval. The results show that the image retrieval accuracy is 40%, the success rate is 78%, the image retrieval accuracy is 50%, the success rate is 80%, and the rate of success is 80%.

# 7. Image Emotional Data Set

Datasets are the basis of image emotion recognition and are an important part of the image emotion recognition field.

## 7.1 Artistic Photographs

The abstract photographs contain only colors and textures, and there are no identifiable objects. Using approximately 230 people to mark 280 pictures, the abstract photographs are divided into eight categories including entertainment, anger, satisfaction, nausea, excitement, fear and, sadness. Each picture is marked approximately 14 times and the category with the largest number of votes is the final marked category. Disputed pictures are discarded and there are 228 photographs remaining.

### 7.2 Abstract Photographs

The abstract photographs contain only colors and textures, and there are no identifiable objects. Using approximately 230 people to mark 280 pictures, the abstract photographs are divided into eight categories including entertainment, anger, satisfaction, nausea, excitement, fear and, sadness. Each picture is marked approximately 14 times and the category with the largest number of votes is the final marked category. Disputed pictures are discarded and there are 228 photographs remaining.

### 7.3 You's large dataset

You et al. [32] established a large-scale dataset to label more than 3 million pictures to accommodate the need for in-depth study. These images are classified into categories such as entertaining, angry, satisfaction, nauseous, excitement, fear and, sadness. It is currently the largest public image emotional dataset.

## 8. Conclusion

Using the emotion recognition of clothing images as an example, this paper discusses the research status of image emotion recognition from three aspects: emotion feature extraction, emotion space establishment and emotion semantic recognition framework. To address the aspect of emotional feature extraction, we review the methods based on color feature, shape feature and texture feature. To address the aspect of emotional space, we establish two emotional spaces and discuss the composition of their coordinates. To address the aspect of emotional semantic recognition, we verify different recognition methods and prove the effectiveness of these methods through the application of image retrieval. Finally, we introduce the benchmark datasets for image emotion recognition.

Image emotion recognition involves many fields and multiple disciplines. We believe that further developments in psychology, cognitive science and computer science will promote it and that it will play a more meaningful role in the future.

## Acknowledgement

This work was supported by the Pyramid Talent Training Project of BUCEA and funded by open research fund program of Beijing Key Laboratory of Robot Bionics and Function Research.

## References

[1] W. Wang and Q. He, “Emotion-based image semantic query through color description,” *Journal of South China University of Technology (Natural Science Edition)*, vol. 36, no. 1, pp. 60-66, 2008.

[2] T. C. Merlo, I. Soletti, E. Saldana, B. S. Menegali, M. M. Martins, A. C. B. Teixeira, S. D. S. Harada-Padermo, M. D. B. Dargelio, and C. J. Contreras-Castillo, "Measuring dynamics of emotions evoked by the packaging colour of hamburgers using Temporal Dominance of Emotions (TDE)," *Food Research International*, vol. 124, pp. 147-155, 2019.

[3] J. Yi, A. Chen, Z. Cai, Y. Sima, M. Zhou, and X. Wu, "Facial expression recognition of intercepted video sequences based on feature point movement trend and feature block texture variation," *Applied Soft Computing*, vol. 82, article no. 105540, 2019. https://doi.org/10.1016/j.asoc.2019.105540

[4] A. Barman and P. Dutta, "Facial expression recognition using distance and texture signature relevant features," *Applied Soft Computing*, vol. 77, pp. 88-105, 2019.

[5] S. M. Alarcao, "Reminiscence therapy improvement using emotional information," in *Proceedings of 2017 7th International Conference on Affective Computing and Intelligent Interaction (ACII)*, San Antonio, TX, 2017, pp. 561-565.

[6] S. Wang, K. Han, and J. Jin, "Review of image low-level feature extraction methods for content-based image retrieval," *Sensor Review*, vol. 39, no. 6, pp. 783-809, 2019.

[7] J. Yao, Y. Yu, and X. Xue, "Sentiment prediction in scene images via convolutional neural networks," in *Proceedings of 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, 2 Wuhan, China, 016, pp. 196-200.

[8] C. Colombo, A. Del Bimbo, and P. Pala, "Semantics in visual information retrieval," *IEEE Multimedia*, vol. 6, no. 3, pp. 38-53, 1999.

[9] N. Jamil, S. Lqbal, and N. Iqbal, "Face recognition using neural networks," in *Proceedings of IEEE International Multi Topic Conference: Technology for the 21st Century (INMIC)*, Lahore, Pakistan, 2001, pp. 277-281.

[10] Y. Yu, Z. Tian, and Z. Cai, "Research of perceptive information of image," *Acta Electronica Sinica*, vol. 29, no. 10, pp. 1373-1375, 2001.

[11] Z. Li, Z. Tan, L. Cao, H. Chen, L. Jiao, and Y. Zhong, "Directive local color transfer based on dynamic look-up table," *Signal Processing: Image Communication*, vol. 79, pp. 1-12, 2019.

[12] H. Zhang, T. Huang, and L. Liu, "Study on the emotion factor space of fabric," in *Proceedings of the International Conference on Kansei Engineering and Emotion Research*, Penghu, Taiwan, 2012, pp. 1174-1178.

[13] T. Hayashi and M. Hagiwara, "Image query by impression words-the IQI system," *IEEE Transactions on Consumer Electronics*, vol. 44, no. 2, pp. 347-352, 1998.

[14] S. Wang, E. Chen, J. Li, and X. Wang, "Kansei-based image evaluation and retrieval," *Pattern Recognition and Artificial Intelligence*, vol. 14, no. 3, pp. 297-301, 2001.

[15] S. Zhao, X. Zhao, G. Ding, and K. Keutzer, "EmotionGAN: unsupervised domain adaptation for learning discrete probability distributions of image emotions," in *Proceedings of the 26th ACM International Conference on Multimedia*, Seoul, Korea, 2018, pp. 1319-1327.

[16] H. Sadeghi and A. A. Raie, "Human vision inspired feature extraction for facial expression recognition," *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 30335-30353, 2019.

[17] S. Zhao, H. Yao, Y. Gao, R. Ji, and G. Ding, "Continuous probability distribution prediction of image emotions via multitask shared sparse regression," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 632-645, 2016.

[18] L. Cai, H. Xu, Y. Yang, and J. Yu, "Robust facial expression recognition using RGB-D images and multichannel features," *Multimedia Tools and Applications*, vol. 78, no. 20, pp. 28591-28607, 2019.

[19] F. Liu, B. Wang, and Q. Zhang, "Deep learning of pre-classification for fast image retrieval," in *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, Sanya, China, 2018, pp. 1-5.

[20] K. Stallman, *Emotional Psychology*. Liaoning, China: Liaoning People Press, 1986.

[21] R. Plutchik, “The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice,” *American Scientist*, vol. 89, no. 4, pp. 344-350, 2001.

[22] M. Y. Tsalamlal, M. A. Amorim, J. C. Martin, and M. Ammi, “Combining facial expression and touch for perceiving emotional valence,” *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 437-449, 2016.

[23] Y. Y. Gao, X. P. Wang, and Y. X. Yin, “Research on affective annotation for natural scene images,” *Journal of Chinese Computer Systems*, vol. 32, no. 4, pp. 767-771, 2011.

[24] Y. Ye, X. Zhang, Y. Lin, and H. Wang, “Facial expression recognition via region-based convolutional fusion network,” *Journal of Visual Communication and Image Representation*, vol. 62, pp. 1-11, 2019.

[25] Q. Dai and Y. Yu, “A kind of image retrieval based on texture features and BP neural network,” *Computer Science*, vol. 27, no. 6, pp. 55-57, 2000.

[26] A. R. Kurup, M. Ajith, and M. M. Ramon, “Semi-supervised facial expression recognition using reduced spatial features and deep belief networks,” *Neurocomputing*, vol. 367, pp. 188-197, 2019.

[27] J. Lee and E. Park, “Fuzzy similarity-based emotional classification of color images,” *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 1031-1039, 2011.

[28] R. V. Priya, “Emotion recognition from geometric fuzzy membership functions,” *Multimedia Tools and Applications*, vol. 78, no. 13, pp. 17847-17878, 2019.

[29] D. Palani and K. Venkatalakshmi, “An IoT based predictive modelling for predicting lung cancer using fuzzy cluster based segmentation and classification,” *Journal of Medical Systems*, vol. 43, article no. 21, 2019. https://doi.org/10.1007/s10916-018-1139-7

[30] C. Singh, E. Walia, and K. P. Kaur, “Enhancing color image retrieval performance with feature fusion and non-linear support vector machine classifier,” *Optik*, vol. 158, pp. 127-141, 2018.

[31] Z. Sun, G. Bebis, and R. Miller, “Quantized wavelet features and support vector machines for on-road vehicle detection,” in *Proceedings of 7th International Conference on Control, Automation, Robotics and Vision*, Singapore, 2002, pp. 1641-1646.

[32] 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*, Orlando, FL, 2014, pp. 47-56.

[33] Q. You, J. Luo, H. Jin, and J. Yang, “Building a large scale dataset for image emotion recognition: the fine print and the benchmark,” in *Proceedings of the 30th AAAI Conference on Artificial Intelligence,* Phoenix, AZ, 2016, pp. 308-314.

[34] S. Corchs, E. Fersini, and F. Gasparini, “Ensemble learning on visual and textual data for social image emotion classification,” *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 8, pp. 2057-2070, 2019.

[35] L. Poretski, J. Lanir, and O. Arazy, “Feel the image: the role of emotions in the image-seeking process,” *Human–Computer Interaction*, vol. 34, no. 3, pp. 240-277, 2019.

[36] A. Hernandez-Garcia, “Perceived emotion from images through deep neural networks,” in *Proceedings of 2017 7th International Conference on Affective Computing and Intelligent Interaction (ACII)*, San Antonio, TX, 2017, pp. 566-570.

[37] D. Kong, X. Shen, L. Cao, and G. Jin, “Phase retrieval for attacking fractional Fourier transform encryption,” *Applied Optics*, vol. 56, no. 12, pp. 3449-3456, 2017.

[38] S. Wang, E. Chen, J. Li, and X. Wang, “Kansei-based image evaluation and retrieval,” *Pattern Recognition and Artificial Intelligence*, vol. 14, no. 3, pp. 297-301, 2001

[39] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, UK: Cambridge University Press, 2004.

[40] C. C. Chang and C. J. Lin, “LIBSVM: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article no. 27, 2011. https://doi.org/10.1145/1961189.1961199

[41] S. Wang, E. Chen, Z. Wang, and Z. Wang, "Research of emotion semantic image annotation and retrieval algorithm using support vector machine," *Pattern Recognition and Artificial Intelligence*, vol. 17, no. 1, pp. 27-33, 2004.

**Guangzhe Zhao** **https://orcid.org/0000-0002-6850-9335**

He received the Ph.D. degree in computer science from Nagoya University, Japan in 2012. He is currently an associate professor with Beijing University of Civil Engineering and Architecture. His research interests include image processing and pattern recognition.

**Hanting Yang** **https://orcid.org/0000-0002-5777-5777**

He received his B.S. degree in building electricity and intelligence from Beijing University of Civil Engineering and Architecture, China in 2013. His current research interests include emotion recognition, fatigue detection and deep learning.

**Bing Tu** **https://orcid.org/0000-0001-5802-9496**

He received the M.S. degree detection technology and automatic equipment from Guilin University of Technology, Guilin, China, in 2009. And Ph.D. degree in mechanical engineering from the Beijing University of Technology, Beijing, China, in 2013. From 2015 to 2016, he was a visiting researcher with the Department of Computer Science and Engineering, University of Nevada, Reno, NV, USA, supported by the China Scholarship Council. Since 2018, he has been an associate professor with the School of Information Science and Engineering, Hunan Institute of Science and Technology. His research interests include sparse representation, pattern recognition, and analysis in remote sensing.

**Lei Zhang** **https://orcid.org/0000-0002-2311-9421**

He received his Ph.D. degree in Beijing Institute of Technology, Beijing, China in 2007. He is currently professor of School of Electrical and Information Engineering, BUCEA. And he is currently deputy director of Beijing Key Laboratory of Robot Bionics and Function Research. His main research interests are humanoid robot, human-machine interaction and machine vision.