Hindawi
Journal of Control Science and Engineering
Volume 2018, Article ID 3768741, 11 pages
https://doi.org/10.1155/2018/3768741

Research Article

Emotional Semantic Recognition of Visual Scene in Flash Animation

Shi Lin,[1,2] Xu Zhenguo,[1] and Meng Xiangzeng[3]

[1]*Faculty of Education, Shandong Normal University, Jinan, China*
[2]*Business School, Shandong Jianzhu University, Jinan, China*
[3]*School of Journalism and Communication, Shandong Normal University, Jinan, China*

Correspondence should be addressed to Meng Xiangzeng; mxz_sdnu@126.com

Received 5 June 2017; Revised 9 November 2017; Accepted 1 January 2018; Published 1 February 2018

Academic Editor: Ai-Guo Wu

Based on the organization structure of the Flash animation files, we first use the edge density method to segment the Flash animation to obtain the visual scenes, then extract the visual features such as color and texture as the input parameters of BP neural network, and set up the sample database. Secondly, we choose a suitable model for emotion classification, use eight kinds of emotional adjectives to describe the emotion of Flash animation, such as warm, delightful, exaggerated, funny, desolate, dreary, complex, and illusory, and mark the emotion value of the visual scene in the sample database and so use it as the output parameter of the BP neural network. Finally, we use BP neural network with appropriate transfer function and learning function for training to obtain the rules for mapping from visual features of the visual scene to semantic space and, at last, complete the automatic classification work of emotional semantic of the visual scene. We used the algorithm to carry on the emotional semantics recognition to 5012 visual scenes, and the experiment effect is good. The results of our study can be used in the classification, retrieval, and other fields of Flash animation based on emotional semantics.

1. Introduction

After nearly 20 years of development, Flash animation has become an indispensable element on the network, and it vividly describes a story in a dynamic way. Therefore, Flash animation has a wide range of applications in advertising, teaching courseware, MTV, games, and other fields. Flash animation can be so popular, mainly because it can express rich emotions. Flash animation creators tend to put their own creative emotions in the works to influence and infect the users, especially in the MTV and games. Different styles of Flash courseware and advertising can also transfer different emotions. The viewers can see through the color and texture of the picture to experience whether the creator expresses happiness or sadness, desolation, or romance. Facing the massive Flash animation resources on the network, users are interested in not only the content of the animation, but also the need to retrieve the animations containing some specific emotions in many cases. For example, a teacher wants to design a lesson and he needs to use a Flash animation that contains a dreary emotion to render the atmosphere. Through the search engine, he used the keyword "dreary" to retrieve but fails to get the desired results. We realized that the study of the emotional content of Flash animation will help to promote the development of Flash animation retrieval. Therefore, we launched the research on emotion annotation of Flash animation and established an information retrieval system based on Flash animation emotion.

Human emotion can be expressed through facial expressions, body actions, and words, and the emotion of Flash animation is expressed through color, texture, sound effect, scene changes, and so on. The emotional experience described by a Flash animation will change with the development of the plot. Such as in a MTV, the leading role may experience the emotional process from sadness to happiness. Therefore, in this study, we use the visual scene as the unit to analyze the emotional semantic. One visual scene is similar

to a shot in the video. We extract the low-level visual features such as color and texture of the visual scene and map it into the high-level emotional semantics which are represented by the visual scene.

Based on the neural network algorithm, by training a certain number of Flash animation samples, we established the mapping from low-level features to high-level semantic of the visual scene and realized the emotional semantic recognition of Flash animation visual scene. Based on the subjective needs and feelings of human beings, we classify and retrieve the Flash animation according to the content of the emotional layer. And the classification and retrieval results can be more in line with human cognitive habits. Our research can be used to establish the Flash animation emotional semantic index database, to achieve the purpose of using emotion keywords to retrieve Flash animation, and provide the basis for Flash animation management, classification, retrieval, and other research.

2. Related Researches

Influenced by subjective factors, so far there is still no unified conclusion about the emotion based classification and description model. In experimental psychology, emotion can be expressed by adjectives. Japanese scholars have studied image emotional semantic description earlier and called emotion and other subjective impressions as "Kansei" [1]. Xia [2] studied four basic emotions: "relaxation," "pleasure," "movement," and "tension." Yoshida et al. [3] discussed three kinds of image emotions, such as "monotonous," "clutter," and "comfort." Colombo et al. [4] chose several commonly used adjectives, such as "natural," "cool," and "warm," to describe images, and established an emotional space. Jing [5] made an emotional classification of clothes with adjectives such as "sexy," "elegant," "pure," "lovel," "mature," and "sports leisure." Yali et al. [6] used "happy" and "sad" to describe the natural scenery. Medical psychologists believe that the complex human emotion has four basic emotional components: "happiness," "sadness," "anger," and "fear" [7]. Osgood et al. [8] proposed that human emotion information should be expressed mainly from three dimensions, such as arousal degree (arousal/activation), pleasure degree (valence/pleasure), and control degree (power/control). Although this theory is not perfect, and it cannot fully reflect the emotion, but in calculating the distance between emotions it is relatively simple and convenient and becomes a generally accepted theory.

In the aspect of image emotional semantic recognition, in 1997, Hayashi and Hagiwara adopted the neural network method to establish the relationship between the impression words and the visual features of images, so as to achieve the extraction of emotional semantic [9, 10]. In Um et al.'s emotion evaluation system of color image based on the multilayer feedback neural network, by mapping from the color feature to the emotion characteristic, it achieved higher accuracy than the linear mapping system [11]. Weining et al. mapped the low-level features of the image into the high-level emotional semantics through a probabilistic neural network (PNN) based on Bayesian minimum risk criteria and achieved good results [12, 13]. Juanjuan proposed a method of constructing image emotional semantic ontology library based on MPEG-7 and fuzzy concept lattice, and designed a decision tree algorithm based on rough set theory to classify the emotional meaning of the image [14].

In the aspect of video emotional semantic recognition, at present, there are many research institutions engaged in research work related to the emotional content of the video, including MIT multimedia lab, Alcatel Lucent's Baer lab, and Microsoft Research Institute. MIT multimedia research lab mainly focuses on the study of emotional mechanism, emotional signal acquisition, emotional modeling and analysis, emotional expression, and other issues, especially on the study of emotional communication between human and computer. Baer laboratory mainly studies the content extraction and retrieval of video and image and proposes the low-level features of the video, such as color, sound, texture, and so on. Microsoft Research Institute, based on Microsoft's Azure cloud service, uses a set of images that mark human emotions to train and can identify the most of the characters in the image as sad or happy. Digital Media Lab of Huazhong University of Science and Technology proposed that the video "content" is divided into three different layers of abstraction, feature layer, cognition layer, and emotion layer, and puts forward the concept of video emotional computing, to establish a unified video emotional semantic space, and in this space to establish a complete set of emotion vector computing system. Professor Wang from University of Science and Technology Beijing and his team put forward modeling method based on emotion space, in which emotion is divided into several basic types, and the basic types are combined to form emotional space and through the calculation of the emotional entropy to obtain the conversion probability of emotion in different state [15]. Lin et al. of Beijing University of Posts and Telecommunications made a deep study on the analysis of video emotional semantic, proposed an analysis algorithm of "emotional syllogism" and applied it to the emotional analysis of film video, and achieved good results [16, 17].

In the analysis of the status, we found that in recent years there has been extensive research on the recognition of emotional semantics, which has made some achievements, but there is no unified emotional classification standard and universal emotional semantic recognition algorithm. Analysis of emotional semantic involves pattern recognition, artificial intelligence, machine learning, computer graphics, multimedia technology, and many other natural sciences and also includes psychology, cognitive science, and other social sciences. Emotional analysis method is still immature, there is not a unified analysis framework, and its evaluation, classification, and retrieval technology is still in the initial stage of exploration and development. The research focus is different, and the main focus of the research was on the emotion classification model, the low-level feature extraction and description, cognitive content identification, emotion mapping model, video encoding description standard, and so on.

The research of image and video emotional semantic recognition is more and more extensive, but the research of high-level emotional semantics based on Flash animation has not yet been carried out. Because of its large area coloring, interpolation to generate animation, and other creative features, Flash animation in terms of emotion recognition has incomparable advantages of image and video. Flash animation emotional semantic analysis integrated the scene segmentation, feature extraction and analysis, emotional mapping, and other aspects, and it is the most difficult and the most comprehensive research content of Flash animation retrieval system based on the content. Focusing on the visual scene of Flash animation, our study will establish a suitable model for emotion classification. Firstly we segment to obtain the visual scene and extract the low-level visual features of the visual scene, then use BP neural network to establish the mapping relationship with the high-level emotional semantics, and finally realize the emotional semantic recognition of visual scene in Flash animation.

3. Visual Scene Segmentation

3.1. Visual Scene. Visual scene generally refers to the picture fragment of a Flash animation, which is composed of continuous frames with similar visual features. One visual scene can express the same picture environment, a continuous action, and the same event. Thus, a visual scene can often fully describe an emotional experience. A Flash animation is made up of a number of visual scenes to express more complex story and more complex emotional changes. Therefore, the visual scene can be used as the basic unit of emotion recognition of Flash animation, and the visual scene segmentation of Flash animation has become a key link of the emotional semantic recognition.

The pictures in the same visual scene have a similar visual effect and contain similar media objects or background. A visual scene fragment contains a number of frames, and each frame contains text, graphics, image, video, audio, and other media objects. The basic structure of the visual scene is shown in Figure 1.

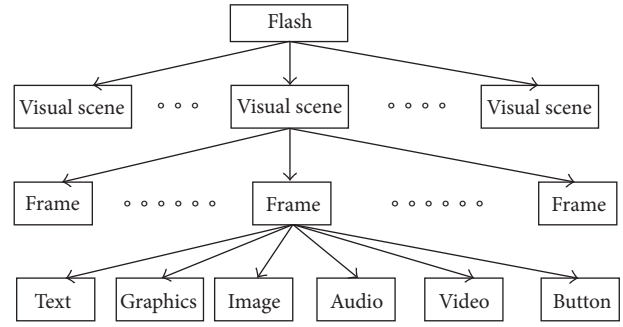

Figure 1: Basic structure of the visual scene.

3.2. Segmentation Process of Visual Scene. The segmentation of visual scene is essentially the clustering analysis based on the similarity of key frames, that is, to determine whether the scene has changed or not through the variety of the visual features of two frame images. This paper mainly studies the segmentation of visual scenes using edge density feature. The work mainly includes three steps: the key frame extraction, edge density segmentation, and the visual scene representative frame extraction.

3.2.1. The Key Frame Extraction. Frames in a Flash animation are divided into key frames and ordinary frames. The key frame refers to the frame which is defined as the object properties changes or object action, and the ordinary frame refers to the middle frame generated by the interpolation operation. Flash animation creators generally use a large number of gradients to present dynamic content. So if we use frame by frame comparison, most of the comparisons are unnecessary, and the algorithm efficiency is low. In this paper, we first obtain the key frames in a SWF file and then complete the visual scene segmentation on the key frame sequence, which can greatly reduce the invalid comparisons and improve the analysis efficiency. In the key frames, it created the dynamic interaction and a variety of dynamic effects of the objects. There are five kinds of actions such as placing objects, removing objects, defining action, changing shape, and changing properties. These five kinds of actions are implemented through the corresponding tags, such as DoAction, DoInitAction, PlaceObject, RemoveObject, RemoveObject2, DefineMorph, PlaceObject2, and PlaceObject3. So we determine whether the current frame is a key frame by judging whether it contains the above tags.

3.2.2. Edge Density Segmentation. The algorithm mainly judge whether the scene change is generated by calculating the change degree of the edges of the object in the frame. In this paper, we use the Canny operator which can better suppress the noise to get the number of edge pixels of the frame and then calculate the edge density of the frame.

The specific steps of using Canny operator to get the edge points are as follows.

Firstly, the Gauss filter is used to smooth the image; secondly, the finite difference method is used to calculate the gradient magnitude and direction; then the gradient magnitude is suppressed by nonmaxima; finally, we use double threshold algorithm to detect and connect edges. On the basis of this, the edge density is defined as the proportion of the number of edge pixels in the total number of pixels in a frame.

Compare the difference of the edge density between two frames, and if it is greater than the preset threshold value, then we can determine that the scene change appears. Otherwise, the two frames are judged to belong to the same visual scene.

3.2.3. Visual Scene Representative Frame Extraction. A representative frame is a static picture frame which can reflect the meaning of a visual scene in Flash animation. It usually can represent the theme of a visual scene. Representative frame extraction target is to use a still frame image to represent the contents of a visual scene. We represent the characteristics

of a visual scene by extracting the visual content of its representative frame.

In the paper, the extraction algorithm of visual scene representative frame is mainly based on the average hue of the scene. That is to say, first we calculate the average hue of a visual scene and then select the key frame closest to the average hue of the visual scene as the representative frame of the scene.

3.3. Visual Feature Extraction. On the basis of obtaining all the visual scenes of a Flash animation, the content feature extraction of each scene is completed. The properties and main contents of the scene are described by the visual features of the visual scene. The visual features are extracted and indexed in order to work for the recognition of emotional semantics in the later period. Hue, texture, movement, sound, text, scene switching (rhythm), and other low-level visual features are the factors that affect people's emotional changes. In this study, we mainly extract the visual features such as the main color and the texture.

3.3.1. The Main Color. Color plays an important role in arousing emotion, setting up the image, highlighting themes, attracting attention, and enhancing artistry. Colors are emotional, and there is a certain relationship between the color and the emotion it wants to express. For example, red makes people feel happy, warm, angry, and energetic. Blue makes people feel calm, rational, and fresh. Green is a feeling of harmony, quiet, health, and safety. And if there is slightly change in saturation and transparency, each color will produce a different feeling.

There are a variety of common color spaces, such as RGB, HSV, Luv, and Lab. They describe colors from different angles. Generally, we use visual consistency as the selection criteria of color space. In this way, the selected color space is more close to people's subjective understanding of color, so as to better reflect the influence of color features on emotional characteristics. The so-called visual consistency refers to calculate the distance between two colors in the selected color space, and when people have a big visual difference feeling between the two colors, the distance between the two colors is also large; when people have a small visual difference feeling, the distance between the two colors is small too. HSV color space model has visual consistency, and compared with RGB color space model, it is more in line with the human visual system. Similarly, the HSV color space model also has three dimensions, which can simultaneously represent three different attributes of color, namely, H (hue), S (saturation), and V (value). These three properties are exactly in line with people's subjective perception of color. Therefore, we use the HSV color space model in our study.

The representative frames we extracted from the visual scene are true color images, and we can extract the RGB color component value from them. Therefore, we need to convert the image description from RGB color space to HSV color space. We assume that $R, G, B \in [0, 1, \ldots, 255]$, and the conversion from RGB color space to HSV color space is as formula (1).

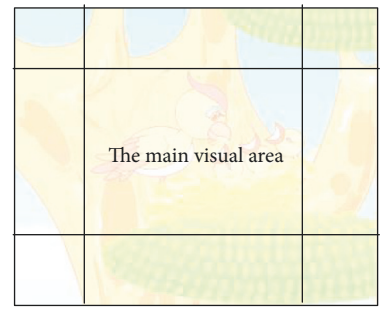

Figure 2: The main visual area.

Initially, we set $r = R/255$, $g = G/255$, $b = B/255$, and $m = \max(r, g, b)$, $n = \min(r, g, b)$

$$H = \begin{cases} 0, & m = n \\ \dfrac{60\,(g-b)}{m-n} + 120, & g = m \\ \dfrac{60\,(r-g)}{m-n} + 240, & b = m \\ \dfrac{60\,(g-b)}{m-n}, & r = m,\ g \geq b \\ \dfrac{60\,(g-b)}{m-n} + 360, & r = m,\ g < b, \end{cases} \tag{1}$$

$$S = \begin{cases} 1 - \dfrac{n}{m}, & m \neq 0 \\ 0, & m = 0, \end{cases}$$

$$V = m,$$

$$H \in [0, 360]\ \ S \in [0, 1]\ \ V \in [0, 1].$$

In this study, we mainly extract the five main colors of the representative frame and their percentages, the average color of the image, and the local color of the image. The local color refers to the average color of the main visual area (as shown in Figure 2) in the representative frame of the visual scene, which represents the color tendency of the whole image.

3.3.2. The Texture. Texture direction, curvature, and the significant degree are also important components of visual features. Texture features have certain regularity and can better reflect the emotional characteristics of image. The factors such as the line grainy, smoothness, and direction will affect people's emotional feelings. For example, the strong sense of edge particles gives people a positive and bright feeling, the smooth edge gives a gentle and relaxed feeling, and the multidirection superposition line gives a complex feeling.

Cooccurrence matrix extraction algorithm is a kind of statistical method of texture feature which is recognized by people, and it is a relatively mature method. We extracted 8 texture features of the representative image, such as ASM (angular second moment), ENT (entropy), CON (contrast),

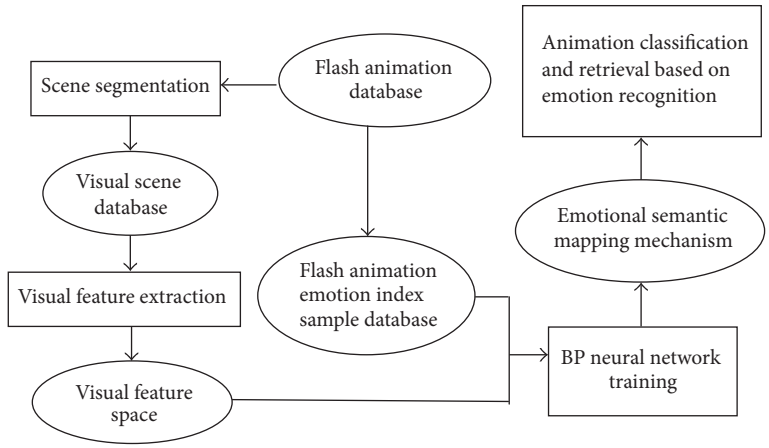

Figure 3: Emotional semantic recognition framework.

and COR (correlation). ASM is the sum of squares of the elements values of the GLCM, which reflects the uniformity degree of the image gray distribution and the coarseness of texture. ENT is the measure of the amount of texture information in the image, which represents the nonuniformity or complexity of the texture in the image. CON directly reflects the brightness comparison of a pixel value and its neighborhood pixel value and also reflects the clarity of the image and the degree of the depth of the texture. COR is used to reflect the consistency of the image texture. It describes the degree of similarity between rows and columns in a matrix, and it is a measure of the linear relationship of gray scale.

We obtained the average value of them and their standard deviation. The process is as follows: firstly, we convert each color component to gray scale; secondly, in order to reduce the amount of calculation, we compress the original image gray level into 16 levels; thirdly, we calculate four cooccurrence matrixes, take the distance of 1 and the angles 0, 45, 90, and 135; fourthly, we normalize the cooccurrence matrix; fifthly, we obtain the cooccurrence matrix and calculate 4 texture parameters, such as the energy, entropy, inertia moment, and the correlation; finally, we calculate the average value and the standard deviation of the above 4 texture parameters and have a total of 8 values as the final texture features.

4. Emotional Semantic Recognition

In the process of emotional semantic recognition of Flash animation visual scene, we first need to establish an emotional classification model and then use neural network to obtain the mapping relationship between the low-level visual features and the high-level emotional semantics. The emotional semantic recognition framework is as shown in Figure 3.

4.1. Emotional Classification Model. According to the interpretation of the "modern Chinese dictionary," emotion is the positive or negative psychological reaction to the stimulation from the outside world, such as liking, anger, sadness, fear, admiration, and disgust. There are many ways to express the emotional characteristics, which are generally divided into discrete and continuous [18]. The continuous emotion model focuses on describing the process of the development and change of the object emotion. The discrete emotion model focuses on describing the emotional state of an object at a certain time. The discrete emotion model is more explicit for emotion classification, and it uses emotional adjectives to describe different emotions, which is easy to be accepted and understood by people and is easy to recognize, calculate, and match. This method of using emotional words to describe human emotions has been going on for a long time. "The doctrine of the mean" divides the emotion into four kinds such as happiness, anger, sorrow, and joy. Foreign scholars also make different statements about the emotional description. Krech divided all emotions into four basic emotions such as happiness, sadness, anger, and fear and gave the definition of each emotion [19]. Through a series of experiments, Shaver et al. divided the emotions into six kinds of basic emotions such as fear, sadness, angry, surprise, joy, and love [20]. Izard divided the emotions into basic emotions and complex emotions and proposed that people total have 8 to 11 basic emotions, namely, fear, interest, disgust, joy, surprise, pain, anger, sadness, shame, contempt, and self-guilt [21–23].

The classification and recognition of emotion have great subjectivity, and its concept is continuously enriched with the development of psychology and cognitive science. Different scholars put forward different views. Whether the emotion model is accurate or not directly affects the mapping mechanism of the visual features and emotional characteristics of the visual scene in this study and then affects the emotional characteristics tagging of computer animation. In this study, we chose the discrete affective model to describe the visual scene of Flash animation. In the discrete emotion model, if the emotion types are many, the neural network will become complex and will reduce the training effect; but if the emotion types are less, it will make the neural network loss information and cannot fully describe the emotion of the images. Therefore, it is necessary to find a reasonable emotional category to balance the relationship between the complexity of neural networks and the type of image emotion.

Table 1: Comparison of emotion types and accuracy rate.

Number of emotions	16	8	6	5	4
Accuracy rate of the maximal main emotion	21.02%	30.43%	32.13%	32.88%	38.74%
Accuracy rate of the first two main emotions	55.56%	76.43%	80.18%	84.53%	89.34%

Table 2: Sample of the emotional feature value of a representative frame.

Visual scene representative frame	Type	Warm	Cheerful	Exaggerated	Interesting	Desolate	Dull	Messy	Illusory
	MTV	3	4	3	3	0	0	0	1

According to the objects of computer animation and its emotional focus, we initially divided the emotion involved in the animation library into 16 kinds, which can be divided into two aspects: positive emotions, and negative emotions. Positive emotions include warm, quiet, cheerful, lively, funny, exaggerated, humorous and interesting; negative emotions include desolate, boring, dull, messy, illusory, thrilling, terror, and fierce. We used neural network to classify emotions of computer animation images, and tested the accuracy of the experimental results corresponding to 16 emotions, 8 emotions, 6 emotions, 5 emotions and 4 emotions to determine the emotional types. The experimental data are shown in Table 1.

It can be seen from Table 1 that with the decrease of the number of emotion categories, the accuracy of sentiment classification is gradually improved, which is the predictable result. When the number of emotions is reduced to 8, the accuracy rate of the two main emotions has reached more than 70%, which basically meets the requirements of emotional classification. When the number of emotions is less than 8, the increasing range in accuracy is not obvious. We know that if the number of emotions is too less, it will lead to incomplete description of the images in the computer animation.

After comprehensive analysis and according to the characteristics of the computer animation and the emotional focus, referring to all kinds of literature, educational needs, and retrieval needs of users, we finally use 8 kinds of emotion to describe the visual scene in the Flash animation database, including 4 kinds of positive emotions: warm, cheerful, exaggerated, and interesting; 4 kinds of negative emotions: desolate, dull, messy, and illusory. Through the statistical analysis of the Flash animation emotion in the experimental database, it shows that these 8 kinds of emotions can describe most of the emotional tendency of the visual scene in Flash animation and can basically meet the needs of users.

4.2. Visual Feature Data Preprocessing. In the content above, we extracted the visual features of the visual scene and obtain the characteristic component of the color and texture of the image. Because the color features and texture features are description of different dimensions of scene, the two are not comparable. And based on the analysis of the neural network, the range of input data should be (0, 1) or (−1, 1). So we need to convert the feature data to normalized range.

The normalization formula is as follows:

$$x_i = \frac{Q_i - Q_{\min}}{Q_{\max} - Q_{\min}} a + b. \quad (2)$$

We set $a = 0.9$ and $b = (1 - a)/2$, where $Q_{\max}$ is the maximum value of the normalized data, and $Q_{\min}$ is the minimum value. Q_i is the value before the normalization, and x_i is the value after the normalization.

It can be seen from formula (2) that the data range after the normalization is (0, 1). At the same time, using the parameters of a and b, we can adjust this data range. Doing this, we can avoid the saturation of the data, which will lead to the reduction of the training effect.

Because the influence of color feature on human emotion is greater than that of the texture feature, we need to give different weight coefficients of the two feature value. The final feature data is $D_{\text{color}} = \alpha \cdot d_{\text{color}}$ and $D_{\text{texture}} = \beta \cdot d_{\text{texture}}$. Among it, $\alpha + \beta = 1$. d_{color} is the color feature value and d_{texture} is the texture feature value after the internal normalization. While D_{color} and D_{texture} are the color feature value and the texture feature value after the external normalization. The normalization of different data is the integration of color feature and texture feature, so as to affect human's emotion together. Through many experiments, we determined the value of α to take 0.6 and β to take 0.4 and formed the visual feature value of the image finally.

4.3. Emotion Feature Data Acquisition. Based on the emotional classification above, we quantify the 8 types of emotions in a visual scene. We use 6 integers to represent the degree of each emotion. "0" represents the irrelevant emotion, "1" represents a slight correlation, "2" represents the general degree, "3" represents the obvious degree, "4" represents the strong degree, and "5" represents the determined emotion. In accordance with this kind of emotion quantization method, we are looking for 4 to 7 people to evaluate each visual scene. An evaluation of a representative frame is as shown in Table 2.

We deal with the emotion evaluation value of each visual scene and calculate the quantization value of the different emotions of each visual scene in the way of getting rid

of the maximum score, getting rid of the minimum score, and calculating the average value of the remaining points. For example, there are seven people to evaluate one visual scene, and the scores for the emotion "worm" were 2, 3, 4, 3, 2, 2, and 1. We get rid of the maximum number 4 and the minimum number 1, and the average value of the remaining points is 2.4. So, we say the quantization value of the "worm" emotion of this visual scene is 2.4. We evaluated all the visual scenes in the sample database according to this evaluation method and then formed the final emotion feature database.

The value range of the emotion features in the database is between 0 and 5, which is more discrete, and therefore we need to deal with the original data. Because the scoring mechanism is independent for each emotion of each visual scene, we can directly divide all the emotion feature data by 5, so as to limit the value of the data in the range of 0 to 1. The compression of the data is beneficial to the accuracy of the output parameters of the neural network in the later period.

4.4. BP Neural Network Learning Process. Neural network is a kind of mapping mechanism with self-learning ability. Through this mapping mechanism, we can get more accurate emotion classification according to the existing data training. In this study, we choose the BP neural network with the strong self-learning ability and the classification ability to do the mapping work of emotional semantic. BP network is a kind of neural network with three or more than three layers of neurons, which is composed of input layer, hidden layer, and output layer. There is full connection between the upper layer and the next layer, while there is no connection between neurons in each layer. The topological structure of BP network belongs to simple forward network structure.

BP neural network learns the input data such as visual features and emotional features of a visual scene. The BP neural networks study and revise the weights and thresholds of the network and find the optimal link between them. And the optimal connection is saved as the universal mapping method for all data, so as to realize the automatic classification from the visual features to the emotion features of the visual scenes.

In the process of training with BP neural network, we first need to determine the number of layers of the network, mainly the determination of the hidden layers. Theoretically, it has been proved that a three-layer BP neural network containing one hidden layer can realize the mapping from any n-dimension to m-dimension, and only when learning a discontinuous function, it requires multiple hidden layers. But excessive number of hidden layers will lead to a more complex training network; the training time will be greatly increased; at the same time, the number of samples will be greatly increased. Therefore, in the actual design process, we should first consider the BP neural network with one single hidden layer. In our system, we choose a three-layer BP neural network with one hidden layer.

After the number of layers of BP neural network is determined, parameters of each layer need to be determined, such as the number of neurons, transfer function, and learning function.

4.4.1. Number of Neurons in the Input Layer. The dimension of the input feature vector affects the number of neurons in the input layer. Above, we introduced the composition of color features and texture features of the image, including main color (15 input items), the average color value (3 input items), local color (3 input items), mean value and standard deviation of energy, entropy, moment of inertia and correlation (8 input items), and a total of twenty-nine input parameters. Therefore, the number of neurons in the input layer of the BP neural network is determined to be 29; that is, the dimensions of the feature vector are 29. It turns out that 29 dimensional input vectors can cause slow computation problems. In order to solve this problem, according to principal component analysis (PCA), we gave up the last three of the five main colors and keep only the first two main colors. In this way, the input feature vector becomes 20 dimensions.

4.4.2. Number of Neurons in the Output Layer. The number of neurons in the output layer depends on the number of the emotion categories. Through analyzing the emotion model of computer animation image and comparing the experimental results, we finally determine the emotional model consists of eight categories, such as warm, cheerful, exaggerated, interesting, desolate, dull, messy, and illusory. Therefore, the output layer neurons vector dimension of BP neural network is determined to be 8.

4.4.3. Number of Neurons in the Hidden Layer. In BP neural network, the choice of the number of neurons in the hidden layer is very important, which will directly affect the performance of the network. Through experiments, we found that when the number of neurons in the hidden layer is 15, the accuracy rate has been increased to more than 78%, and the classification effect is better. At the same time, considering the complexity of the neural network, the selection of too many hidden layer neurons will increase the difficulty of training, and it is not conducive to the improvement of classification accuracy. Therefore, the number of neurons in the hidden layer is determined to be 15.

4.4.4. Function Selection. After determining the number of neurons in the input layer, the hidden layer, and the output layer, the transfer function and the learning function between the layers are also required. After processing the color feature and texture feature, the numerical range is (0, 1), so we select the logarithmic Sigmoid formula as the transfer function from the input layer to the hidden layer, that is, formula (3). The output value of the BP neural network is the emotional features, and according to the characteristics of various transfer functions, we use the Purlin formula as the transfer function from the hidden layer to the output layer, that is, formula

$$f(x) = \frac{1}{1 + e^{-\alpha x}}, \tag{3}$$

$$f(x) = kx. \tag{4}$$

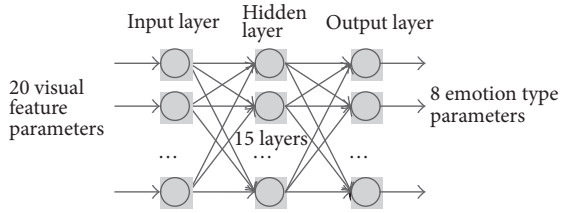

Figure 4: Three-layer mapping network structure of the emotional semantic of the visual scene.

Through the experimental tests of several learning functions, we find that when we select Trainlm learning function, the learning time is the shortest, the training times are the least, and the training error is the minimum, which has higher accuracy rate of emotional classification. Therefore, the paper decided to use Trainlm learning function to achieve faster convergence speed and better training results.

4.4.5. Parameters Selection. (a) We design a random generator program to generate a set of −0.5~+0.5 random numbers as the initial weights of the network; (b) the training rate value is 0.9; (c) the dynamic coefficient is 0.7; (d) the allowable error value is 0.0001; (e) the Sigmoid parameter value is 0.9.

Finally, the BP neural network structure we build in this study is shown in Figure 4.

The input matrix of our model is as formula

$$\mathbf{x} = \begin{vmatrix} x_{11} & \cdots & x_{1n} \\ \cdots & \cdots & \cdots \\ x_{m1} & \cdots & x_{mn} \end{vmatrix}. \tag{5}$$

The number of rows in the above matrix is the number of samples, $m = 3789$. And the number of columns is the number of the feature extracted from each sample, that is, the dimension of the input vector, $n = 20$.

The output matrix is as formula (6). In this formula, $i = 3789$ and $j = 8$. "j" is the feature number of the output of each sample, that is, the dimension of the output vector,

$$\mathbf{y} = \begin{vmatrix} y_{11} & \cdots & y_{1j} \\ \cdots & \cdots & \cdots \\ y_{i1} & \cdots & y_{ij} \end{vmatrix}. \tag{6}$$

Here, the basic mathematical expression of the above NN model is as formula

$$\begin{aligned} \mathrm{sim}_y = {} & w_{1\,1}^{(2,3)} * \mathrm{sigmoid}\left(w_{1\,1}^{(1,2)} * x_1 + w_{2\,1}^{(1,2)} * x_2 + \cdots \right. \\ & \left. + w_{20\,1}^{(1,2)} * x_{20} + b_1^{(2)}\right) + w_{2\,1}^{(2,3)} * \mathrm{sigmoid}\left(w_{1\,2}^{(1,2)} * x_1 \right. \\ & \left. + w_{2\,2}^{(1,2)} * x_2 + \cdots + w_{20\,2}^{(1,2)} * x_{20} + b_2^{(2)}\right) + \cdots \\ & + w_{15\,8}^{(2,3)} * \mathrm{sigmoid}\left(w_{1\,15}^{(1,2)} * x_1 + w_{2\,15}^{(1,2)} * x_2 + \cdots \right. \\ & \left. + w_{20\,15}^{(1,2)} * x_{20} + b_{15}^{(2)}\right) + b_8^{(3)}. \end{aligned} \tag{7}$$

There are two classes of parameters in the upper formula: the weight "w" and the threshold "b." Like, $w_{2\,1}^{(2,3)}$ represents the weight value of the second node in second layer to the first node in third layer. $b_2^{(2)}$ represents the threshold of the second node in second layer.

5. Experimental Result

In our study, accuracy, recall and F measure are used to evaluate the quality of visual scene segmentation results. Among them, accuracy is defined as the ratio of the number of accurate visual scenes to the total number of visual scenes obtained by automatic segmentation. Recall is defined as the ratio of the number of accurate visual scenes to the total number of visual scenes obtained by artificial segmentation. F measure is the harmonic mean of the accuracy and the recall, that is, double the product of the accuracy and the recall, and then divided by the sum of them.

We downloaded 702 Flash animations from the network as the object of our experiment, including games, MTV, animation, courseware, and advertisement. To avoid subjectivity, we first asked several experimenters to manually annotate the visual scenes of the 702 Flash animations, and then we selected the scenes with high repetition rate as the correct scene. Finally, we obtained 5512 visual scenes that we consider to be correct. We use them to test the visual scene segmentation accuracy of edge density algorithm in this paper. Our experiments used VC++ compiler environment to complete the segmentation of visual scene and the extraction of visual features. Using edge density algorithm, we obtain a total of 5787 visual scenes, of which there are 5052 scenes which are correct. So the segmentation precision reaches 87.3%; the recall reaches 91.7%; and the F measure is 89.4%.

We manually annotated the emotion types of the 5052 correct scenes, automatically extracted the visual features of each scene, and established the visual feature-emotion database of the visual scene. Then 3/4 of the scenes were used as training samples of the BP neural network, a total of 3789 samples; and the remaining 1/4 of the scenes were used as test samples, a total of 1263 samples. In the experiment part of emotional semantic recognition, we first used training samples and reasonable training functions to train the BP neural network to obtain the best combination of parameters for visual scene emotional recognition. Then, we used the trained neural network to classify the test samples and analyzed the classification accuracy.

We use the neural network toolbox of MATLAB to learn the neural network. The main interface of the toolbox is shown in Figure 5.

In Figure 5, the "Import" button is responsible for importing the input data and the target data. The "Export" button is responsible for the export of data and neural network. The "Simulate" button is responsible for neural network simulation. The "Train" button is responsible for the training of neural network, and we need to set the maximum number of training steps and maximum training error. The "New Network" button is the core of the neural network toolbox. It is responsible for adjusting the range of input data,

Table 3: Neural network emotion classification results summary.

	Warm	Cheerful	Exaggerated	Interesting	Desolate	Dull	messy	Illusory
Goal	0.8	0.9	0.5	0.8	0.1	0.1	0.5	0
Result	0.7512	0.87032	0.51056	0.82135	0.07185	0.06393	0.30128	0.00693

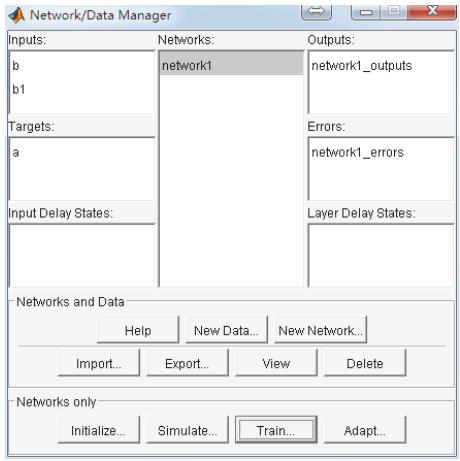

Figure 5: The main interface of the neural network toolbox.

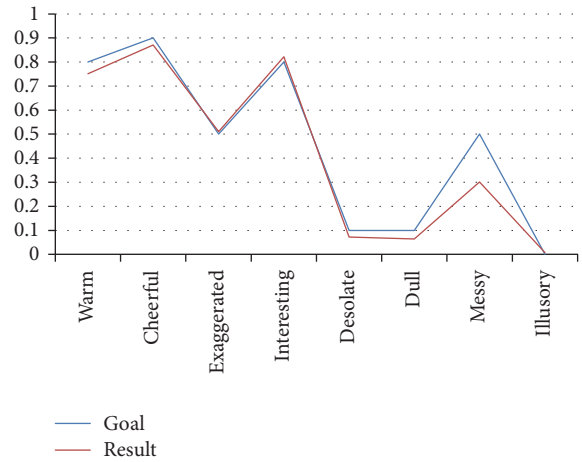

Figure 6: The emotional mean line graph.

selecting training function and performance function, and determining the number of neurons and transfer functions of each layer of neural network.

After the network training is completed, we got the final connection weight and threshold matrix from input layer to the hidden layer and the final connection weight and threshold matrix from the hidden layer to the output layer. Then we use the BP neural network to simulate the emotional recognition of the visual scene in the test sample database. The maximum value of the 8 output parameters of the neural network is the emotion of the visual scene we recognized. Then we use this emotion to compare with the visual scene emotion in the test sample database, and if the comparison result is consistent, we think that the recognition is correct. In this paper, we use the recognition accuracy to describe the effect of visual scene emotion recognition, and we define the accuracy rate as formula

$$\text{Recognition Accuracy} = \frac{\text{the number of accurate recognition}}{\text{Total sample numbers}}. \quad (8)$$

To a certain extent, the formula can objectively reflect the quality of the training results of the system. At last, the recognition accuracy of the 1263 visual scene test samples is 84%.

In order to facilitate the analysis of the experimental results, we take the emotional mean values of all the visual scenes in the test sample database as the target values by category and take the emotional mean values obtained by using BP neural network as the result values of the experiment and then compare the two means to make a summary. The comparison results are shown in Table 3.

According to the emotional mean value given in Table 3, we draw the line graph, as shown in Figure 6.

To a certain extent, the graph above can reflect the quality of BP neural network for Flash animation emotional classification. It can be seen that BP neural network algorithm has better classification results in the emotional classification of visual scenes in Flash animation. The target curve and the experimental result curve are basically the same. There are only some deviations when in the "messy" identification. The analysis shows that the image content of "messy" emotion is too rich, and the change is not regular. So it will produce greater interference in the identification. In the future, we can use this BP neural network to classify the network Flash animations by emotion and establish index database to improve the retrieval efficiency and to meet the needs of network users.

6. Summary

With the coming of multimedia era, people pay more and more attention to Flash animation. Massive Flash animations on the network make its retrieval become a problem that must to be solved. The efficiency and accuracy of Flash retrieval based on text and metadata are generally not high, while the content-based Flash animation retrieval algorithm can improve the problem. From the low-level features of Flash animation to the high-level features, the content semantics of animation are more closer to the cognitive characteristics of people. Emotional semantic belongs to high-level feature. Based on the subjective needs and feelings of human beings, we classify and retrieve Flash animation according to the content of the emotional layer, and the obtained classification results can be more in line with human cognitive

habits, and it is the research direction of Flash animation classification and retrieval in the future. Visual scene, as the basic unit of describing Flash animation story, is suitable for the recognition of emotional semantics. In this paper, we mainly studied the mapping rules from the low-level visual features such as color and texture of the visual scene in Flash animation to the high-level emotional semantic features. The idea of this research is to segment and obtain the visual scenes of the 702 Flash animations downloaded from network by using edge density method and use the correct-segmented visual scene to establish the training sample database and test sample database and then extract the color features and texture features of visual scenes. After that, we choose discrete emotion classification model to carry on the emotion annotation to the samples. We use eight emotional adjectives to describe emotions: warm, cheerful, exaggerated, interesting, desolate, dull, messy, and illusory. We use the BP neural network to train and learn the samples of the training sample database and then obtain the mapping network from visual features of the visual scenes to the emotion adjectives. At last, we use the emotional semantic mapping network to recognize the emotional semantic of samples in the test sample database and analyze the recognition accuracy.

The research of emotional semantic analysis of Flash animation combines various fields such as visual scene segmentation, visual feature extraction, emotional model selection, and emotional mapping. And it is one of the most difficult and comprehensive research directions in content-based Flash animation retrieval system. Although our research has achieved good results, there are also shortcomings. For example, we can consider the combination of color or brightness to improve the segmentation accuracy and efficiency of Flash animation visual scene. And in the emotional semantic recognition process, we only considered low-level features such as color and texture of visual scene. In the future, we will consider combining more complex features such as shape, text, and audio for emotional semantic mapping. In addition, we used BP neural network to do semantic mapping work, and later we should study the emotional semantic recognition algorithm based on deep learning method. If so, we believe that the recognition accuracy can be greatly improved, and it is the future research direction.

Conflicts of Interest

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

References

[1] "Kansei sessions," in *Proceedings of the IEEE International Conference on Systems Man and Cybernetics*, Tokyo Japan, 1999.

[2] M. Xia, D. Yukuan, and M. Tianyimi, "Analysis of image emotion characteristic and its harmonious feeling evaluation," *Electronic Journal*, S1, pp. 1923–1927, 2001.

[3] K. Yoshida, T. Kato, and T. Yanaru, "Image retrieval system using impression words," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, vol. 3, pp. 2780–2784, October 1998.

[4] C. Colombo, A. del Bimbo, and P. Pala, "Semantics in visual information retrieval," *IEEE MultiMedia*, vol. 6, no. 3, pp. 38–53, 1999.

[5] H. Jing, *Research on Clothing Image Emotional Semantic Classification Based on Feature Fusion*, Taiyuan University of Technology, Shan Xi, China, 2007.

[6] F. Yali, G. Na, and Z. Jiawei, "Image emotional semantic classification based on color and texture features," *Journal of Zhengzhou University of Light Industry*, vol. 23, no. 6, pp. 118–121, 2008.

[7] Q. Jiang, *Medical Psychology*, Chinese People's Medical Publishing House, Seijing, Chinea, 3rd edition, 2002.

[8] C. E. Osgood, J. G. Suci, and P. H. Tannenbaum, *The Measurement of Meaning*, University of Illinois Press, 1957.

[9] T. Hayashi and M. Hagiwara, "Image retrieval system to estimate impression words from images using a neural network," in *Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics*, vol. 1, pp. 150–155, IEEE, New York, NY, USA, October 1997.

[10] 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.

[11] J. Um, K. Eum, and J. Lee, "A study of the emotional evaluation models of color patterns based on the adaptive fuzzy system and the neural network," *Color Research & Application*, vol. 27, no. 3, pp. 208–216, 2002.

[12] W. Weining, Y. YingLin, and Z. Jianchao, "Image emotional semantic classification based on line direction histogram," *Computer Engingeering*, vol. 31, no. 11, pp. 7–9, 2005.

[13] Q. Ping and L. Chengmou, "The effects of positive and negative emotions on explicit memory and implicit memory," *Journal of Southwestern Normal University*, vol. 28, no. 1, pp. 143–148, 2003.

[14] Z. Juanjuan, *Related Technology Research on Image Visual Features And Emotional Semantic Mapping*, Taiyuan University of Technology, Shan Xi, China, 2010.

[15] W. Guojiang, W. Zhiliang, Y. Guoliang, W. Yujie, and C. Fengjun, "Review of research on artificial emotion," *Computer Application Research*, vol. 7, pp. 7–11, 2006.

[16] X. Lin, *Movie Emotion Recognition Based on Fuzzy Theory*, Beijing University of Posts and Telecommunications, BeiJing, China, 2009.

[17] X. Lin, X. Wen, Z. Lu, and W. Zheng, "Video affective content recognition based on film grammars and fuzzy evaluation," in *Proceedings of the 2008 International Conference on MultiMedia and Information Technology (MMIT '08)*, pp. 264–267, December 2008.

[18] Rosalind Picard.Affective Computing.United States.The MIT Press,.

[19] Krech, *Essentials of Psychology*, Cultural and Educational Publishing House, Beijing, China, 1980.

[20] P. Shaver, J. Schwartz, D. Kirson, and C. O'Connor, "Emotion knowledge: further exploration of a prototype approach," *Journal of Personality and Social Psychology*, vol. 52, no. 6, pp. 1061–1086, 1987.

[21] C. E. Izard, "On the ontogenesis of emotions and emotioncongnition relationships in infancy," in *The Development of Affect*, M. Lewis and L. A. Rosenblum, Eds., pp. 389–413, Plenum Press, New York, NY, USA, 1978.

[22] C. E. Izard, "Emotion-cognition relationships and human development," *Emotions, Cognition, and Behavior*, pp. 17–37, 1984.

[23] C. E. Izard, "Basic emotions, relations among emotions, and emotion-cognition relations," *Psychological Review*, vol. 99, no. 3, pp. 561–565, 1992.