International Journal of Information Technologies and Systems Approach
Volume 16 • Issue 2

Recognition and Analysis of Scene-Emotion in Photographic Works Based on AI Technology

Wenbin Yang, Shandong University of Arts, China*

ABSTRACT

Emotional effect is highly subjective in people's cognitive process, and a single discrete emotional feeling can hardly support the description of the immersion scene, which also puts forward higher requirements for emotional calculation in photography. Therefore, this article first constructs a photographic scene recognition model, and then establishes a visual emotion analysis model which optimizes the basic structure of vgg19 through CNN, extracts the user's photography situation information from the corresponding image metadata, establishes the mapping relationship between situation and emotion, and obtains the low-dimensional dense vector representation of the situation features through embedding. The authors divided eight emotional categories; accuracy of the model is compared and the feature distribution of scene-emotion in different works is analyzed. The results show that the accuracy of the scene-emotion recognition model of photographic works after multimodal fusion is high, reaching 73.9%, in addition, different shooting scenes can distinguish the emotional characteristics of works.

KEYWORDS

Emotion, Feeling-CNN, Photographic, Scene, Vgg19

1. INTRODUCTION

Through the analysis and modeling of information such as the content identification of real-time photography scenes and the emotional preferences of users, combined with cutting-edge technologies such as image understanding and text generation in deep learning, the emotional state of users and the content of photography scenes can be accurately analyzed (Wei et al, 2022). while the existing sentiment analysis mostly starts with the texts produced by users on the Internet, which uses natural language processing and other technologies for analysis (Yu et al, 2022; Chatterjee, 2019). With the great improvement of the ability of convolution neural network to process image information, there are more researches on analyzing users' emotions through photographic works, which achieves good emotion classification results (Rao et al, 2016; Meng et al, 2021). Different from the tasks of object recognition and scene recognition, the task of visual emotion analysis involves more complicated factors, in addition to the image, due to the influence of individual factors (including

DOI: 10.4018/IJITSA.326055 *Corresponding Author

growth environment, cultural background, social background, etc.), different people have diverse emotional understandings for the same image (Burkitt, 2002). Therefore, it is necessary to consider richer elements as much as possible in visual emotional analysis.

The deep learning model can input the texture, color and other information of the image, automatically extract the emotional features of the image, and make use of the dependency relationship between the features of different levels to model the learning representation of them, which has achieved good results in the classification of photographic emotions (Li, 2019; Zhu et al. 2019). However, learning photographic features from a global perspective makes it difficult for the convolutional neural network to determine which region or law in a photographic work to fully express the emotion expressed in the shooting process, and to define the influence of regional information on the overall emotion of the work. users' emotion towards photography is subjective, and people's emotion is related to many factors, such as people's environment at that time, the photographed content, etc. The above method of emotion classification only from the image level ignores the contextual information behind the image, but there is abundant emotional information hidden in the scenes. Therefore, it is difficult to accurately capture the fine-grained emotion of users(Bhunia et al, 2022; Li et al, 2020).

Traditional CNN can only analyze a single feature domain, ignoring the contextual information behind the photography. Deep learning methods such as convolution neural network, embedding feature and multi-feature fusion can be used to improve the effect of emotion recognition. Based on the above problems, the main contribution of this study is that

(1) This paper optimizes the basic structure of vgg19 through CNN, builds a photographic scene recognition model and the visual emotion analysis model, which establishes the mapping relationship between the scene and emotion;
(2) The information of user photography situation is extracted from the corresponding image metadata, and the mapping relationship between situation and emotion is established. The low-dimensional dense vector representation of situation features is obtained through embedding;
(3) The features of photographic works and the contextual features behind them can be fully integrated, so as to expand the feature domain of data.

2. RELATED WORKS

2.1 Analysis of Photographic Emotion

Emotion analysis is to analyze and identify emotion polarity in subjective information carriers (text, image, voice, etc.) with various modern information processing. In recent years, with the popularity of multimedia information, the traditional single modal sentiment analysis method is not effective in the recognition of complex scenes (Liu et al, 2020). In the analysis of photographic emotion, the current research can be roughly divided into two categories, one is based on the extraction of photographic visual feature, and the other is based on artificial intelligence (AI).

Extraction of low-level features based on photographic vision, including features without semantic information such as color, line, texture and shape. Zhou et al. (2016) researched on landscape images, according to different color features and emotional features, established the relationship between image color features and user evaluation, and extracted the color features of landscape images by combining the color histogram of images, where SVM was used to classify emotions, and achieved good accuracy. Cai et al. (2019) considered the differences between the whole and local photographic images to highlight the emotional color, and proposed a method of embedding the whole image features and local image features to predict the emotional polarity of photography.

A series of deep learning methods, represented by convolutional neural networks, have achieved great success in photography content classification and sentiment analysis. Yan (2018) confirmed

that the features extracted by neural network perform better than manual design, because neural network can learn how to express features on specific tasks, and the result of this feature selection is the process of network learning. Acar et al. (2016) studied how to apply CNN to the task of visual emotion prediction by visualizing local image patterns related to emotion. Xu et al. (2014) used the idea of transfer learning, trained convolution neural network on Image-Net data set, and then fine-tuned the model parameters on the photographic emotion classification task, which accelerated the training speed of the model.

It can be seen from the above researches that the low-level features are limited by the judgment of subjective consciousness and that the extraction of design features is tedious, and the classification effect is not as good as that of the end-to-end neural network method. At the same time, the deep learning method only analyzes a single feature domain at present, which ignores the contextual information of photography.

2.2 Analysis of Photographic Scene

Situation is an objective context information that describes the location and environment of things (including temperature, noise, light, etc.). The purpose of modeling scenes is to establish a unified digital representation of user context in this environment. This work can be traced back to a long time ago that when mobile devices were not popular, they mainly relied on some manual rules, and triggered by GPS, etc. With the enrichment of mobile situational data, some scholars use machine learning model to model the situation. Wu (2016) adopted a situational modeling framework based on representation learning in the scenario of user behavior prediction. While Luddecke et al. (2018) have developed context modelling language (CML) by using object-oriented representation method to quickly model contextual data from database. There are also some applications and researches on situational modeling by using image metadata. Subudhi et al. (2019) analyzed tourists' emotional activities based on the location information of photos and drew "emotional maps". Hossain et al. (2020) considered the characteristics of people recalling photos (scene-like memory), and fused metadata such as image content and photo location, so as to create a scene-like search platform, which users can describe through scenes.

It can be seen that the modeling method of photography scene is flexible, but different scenes and data formats have different emphases, which makes it difficult to describe the scene. The data in this photographic scene comes from Exif image metadata, which is helpful to enhance the effect of photographic emotion analysis.

3. FEATURE RECOGNITION IN PHOTOGRAPHIC SCENE

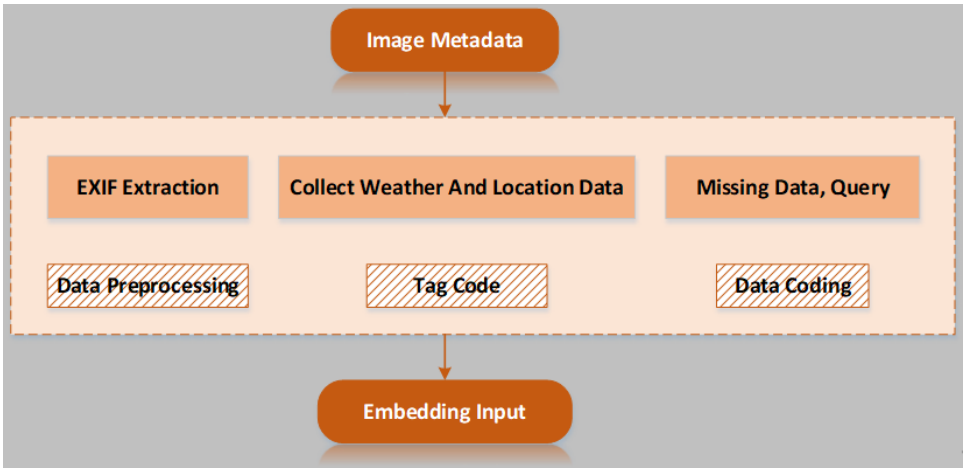
In this paper, two dimensions of the photo scene are selected, namely, user behavior features and emotional description text features. The formalization of information can be expressed as: UC = {Behavior _ info, Context _ info}

The selected features can be obtained directly or indirectly through mobile devices or social software. Exif is a kind of image metadata, which can tell us when and where this photo was taken, and what parameters and equipment it was used. It records the data of resolution, shooting time, ISO, shutter time, focal length, exposure and so on. Through latitude and longitude, we can also find out specific location information, some of which are continuous values, such as ISO values, which need to be coded discretely when extracting. The modeling process of photographic scene is shown in Figure 1.

3.1 Information Representation of Photography Behavior

Behavior depicts the photography mode, aperture size, and filter style of the user when shooting images. For example, photos with small aperture are usually used in the perspective of shooting natural scenery, which can speculate on the content and emotions that the user wants to express when shooting. If the manager has negative emotions such as sadness, it can be inferred that the light is

Figure 1. Modeling process of photographic scene



dark and the ISO sensitivity is high when shooting. While if the filter style is warm colors, it usually indicates that the user may feel more warm emotions at this time. The informatization expression of user behavior is Behaviour_info. = { Photo_model,ISO,Aperture,Exposure-time,Filter_model},where Photo_model represents the camera shooting mode (six common modes are selected, such as wide angle, night scene, slow shooting, portrait and macro), and ISO,Exposure_time and Aperture respectively represent the sensitivity, exposure time and aperture size; Filter_model indicates the category that users decorate images with filters. Because there are too many styles of filters, this paper simplifies the classification of filters into two categories: cold and warm.Including eight emotional categories: Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear and sadness., the distribution of these features in emotional images is different, for example, the aperture value of landscape images with awe-inspiring is lower than other categories. Because these features need to be transformed by embedding layer, the continuous value features are discretized, and then carry out category coding.

3.2 Basic Characteristics of Photographers

Photography features also include the basic features of users, such as the user's name, age, gender, etc. The label information of user basic information is expressed as Userinfo. = { Uuid,Name,Sex,Age } , where Uuid is the unique identifier of the user, Name is the user's name, Sex is the user's gender, and Age is the user's age (divided into five intervals: below 18, 18~25, 25~40, 40~60 and above 60 in this paper). Therefore, the coding of basic information features is divided into three layer and embedded respectively, and the dimension after embedding features is 3.

3.3 Characteristics of Shooting Behavior

3.3.1 Information of Time

From the date "year, month, day" to the early and middle of the day, for example, people are more likely to show their true feelings in the middle of the night. Formally expressed as: Timeinfo = { Datetime,Season,Parttime } , where datetime represents the specific date (XX year, XX month, XX day), Season represents the season (Spring, Summer, Autumn and Winter), and Parttime represents the time period of the day (morning, noon, afternoon and night).

3.3.2 Information of Location

Location features are divided into six categories, namely scenic spots, restaurants, sports grounds, amusement parks, etc. If users come to some magnificent scenic spots, they may be full of awe of nature, and they are more likely to have positive emotions in amusement parks. The formal expression of information label is: Location_info. = { City,Position,Pos_type } , where City stands for city, Position stands for a certain area in the city, which can be a specific scenic spot or area, and Pos_type indicates which type the location belongs to.

3.3.3 Information of Weather

Weather features include specific weather, temperature and humidity. For example, in the summer with cloudy days, 32°C and 60% relative humidity, the possibility of users feeling annoyed will increase. The formal expression is: Weather_info = {Weather, Temperature, Humidity}, where the categories of Weather are sunny, cloudy, rainy and snowy. The Temperature is divided into five grades from low to high, and the Humidity is divided according to the same rules.

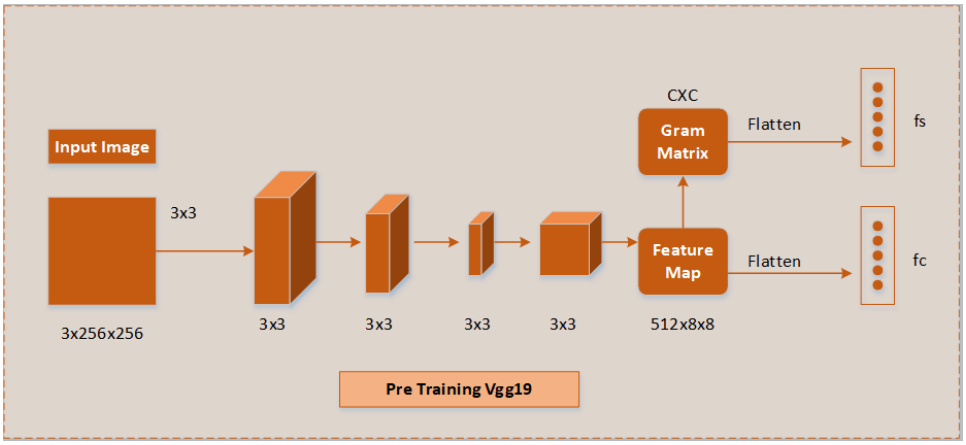
4. SCENE-EMOTION RECOGNITION MODEL IN PHOTOGRAPHY

4.1 Visual Emotion Analysis Based on CNN

CNN has achieved unprecedented results in the field of photography through convolution kernel sharing. The most commonly used vgg19 network consists of three parts: feature extraction layer, adaptive pooling layer and full connection layer. Content features and texture features are fused by Concatenate, and the input data is propagated forward through the convolution layer, the downsampling layer and the full connection layer to get the output value. Its implementation process is that fix its parameters through the feature extraction layer of vgg19 network pre-trained on ImageNet, reconstruct it according to the requirements and then fine-tune the parameters of the full connection layer. This paper also adopts this strategy. Figure 2 shows the architecture of emotion recognition model improved by vgg19.

As can be seen from the above structure, emotion recognition model includes two stages. The first stage is feature extraction. By fixing the model parameters of vgg19' s feature extraction module, which has been pre-trained on ImageNet, we can get the feature map, which represents the high-level semantic features of the image. In the second stage, feature map is processed in two ways, one is texture feature extraction module,Through channel-wise matrix operation of feature map, the gram

Figure 2. The architecture of emotion recognition model



matrix is obtained, which represents the texture features of the image. Then feature map is originally the representation of the image content features. Finally, the content features and texture features are fused by vector stitching, and the fusion coefficient is added to make two basic structures. The method of joint learning is adopted for CNN training.

In the second stage, the gram matrix is calculated by multiplying the matrix of channel-wise dimension by feature map, whose shape is $\mathrm{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$, in which C indicates the number of channels, W and H represents width and height, the real meaning of C is the number of convolution kernels on the upper layer, which represents different characteristics. Therefore, the feature map can be re-expressed as shown in the following Formula:

$$f = \left[\alpha_1, \alpha_2, \alpha_3, \cdots, \alpha_c\right] \tag{1}$$

where, α_c represents the N-dimensional vector composed of each neuron in channel C, N is equal to the product of H and W, and gram matrix is expressed as follows:

$$G\left(f\right) = \begin{bmatrix} \alpha_1 & \alpha_1 & \alpha_1 & \alpha_2 & \cdots & \alpha_1 & \alpha_c \\ \alpha_2 & \alpha_1 & \alpha_2 & \alpha_2 & \cdots & \alpha_2 & \alpha_c \\ \cdots & & \cdots & & \ddots & \cdots & \\ \alpha_c & \alpha_1 & \alpha_c & \alpha_2 & \cdots & \alpha_c & \alpha_c \end{bmatrix} \tag{2}$$

Among them, α_i, α_j represents vector inner product, which is a scalar; According to the calculation formula of Gram matrix, the size of Gram matrix is $\mathrm{R}^{\mathrm{C}\times\mathrm{C}}$, which is a symmetric real matrix; $\mathrm{G}\left(\mathrm{i,j}\right)$ represents correlation between channel i and channel j where different channels represent different filtering modes and represent different characteristics. The number of channels in Feature Map here is 512. In order to make network training convergence faster, Softmax is performed on the values of elements in each row and flatten operation is performed on normalized 0-1 matrix to shape it into the vector with the size of 512× 510/2. Finally, the vector representation of phote texture feature is obtained.

For the representation of photo content features, the Flatten operation is performed directly on the Feature Map, which is similar to the full connection layer of CNN model. All features are globally fused, and the shape of the vector is 512× 8× 8.

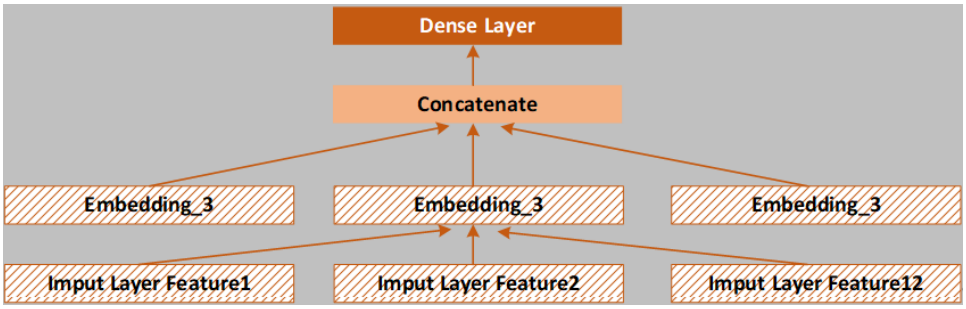
4.2 Scene-Emotion Model

Because deep learning has the problem of slow convergence for sparse feature training, using an embedding layer as feature mapping, and word vector as the representative product of embedding, shines brilliantly in many tasks in the field of Natural Language Processing. It can get the vector representation of each word in low dimension by learning, which is a distributed representation. In addition, it is also widely used in various fields of deep learning. As the input mapping layer of neural network, it can get better classification effect in many applications. Its core idea is to train and learn high-dimensional features to get low-dimensional vector representation. The features extracted from the metadata of photo make up 12 features, which are equivalent to a self-built dictionary. Then it is mapped to a dense low-dimensional vector through the following embedding layer. The structure of emotion-situation model is shown in Figure 3.

The formal representation of each feature is as follows:

$$e_i = l_i \times M_i^{k\times h} \tag{3}$$

Figure 3. Structure of emotion-situation model



where l_i represents the one-hot encoding vector of feature i, M_i is a matrix of $R^{k\times h}$, k represents the dimension of feature i, h represents the new dimension of feature i after embedding, e_i is an h-dimensional vector, that is, the new vector representation of feature i. Next, vector splicing of 12 features was done to form L_d, which was Concatenate, and finally L_d obtained is a 36-dimensional vector.

The role of scene-emotion embedding is to map the original sparse image metadata's one-hot scene features into dense low-dimensional vector forms by embedding. whose purpose is to make the training of neural network converge faster and improve the final emotion classification effect.

4.3 Emotion Recognition Based on Multimodal Fusion

The main data sources of photography scenes in this study include photos and captions. Usually, emotional analysis of photos is directly used as inference of users' emotions, but sometimes users take some positive emotional photos, which may have negative emotions when deliberately hiding or inspiring their own purposes.At this time, the essay can be used as an auxiliary sentiment analysis to identify the user's sentiment together. Figure 4 shows the emotion recognition model of multimodal fusion.

Among them, Feeling-CNN represents the model of visual emotion recognition. The input of this part is the user's photographic works, and BLSTM is the model of emotion analysis, whose input is a simple description of the user's mood or content of the taken photos. Finally, the probability distribution of emotion prediction of these two parts is fused and weighted, and the emotion category is obtained. Taking "Erhai Lake in July is very beautiful, I like this lifestyle that I have the mood to cook, have the ability to read, and have time to travel "as an example, its realization process is shown in Figure 5:

Figure 4. Emotion recognition model based on multimodal fusion

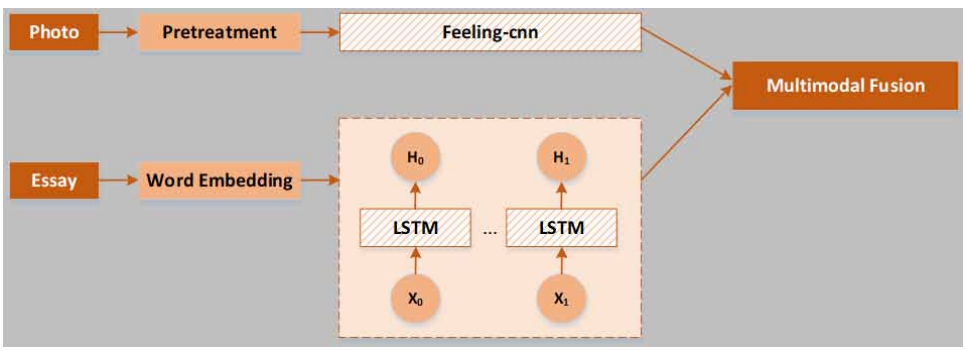
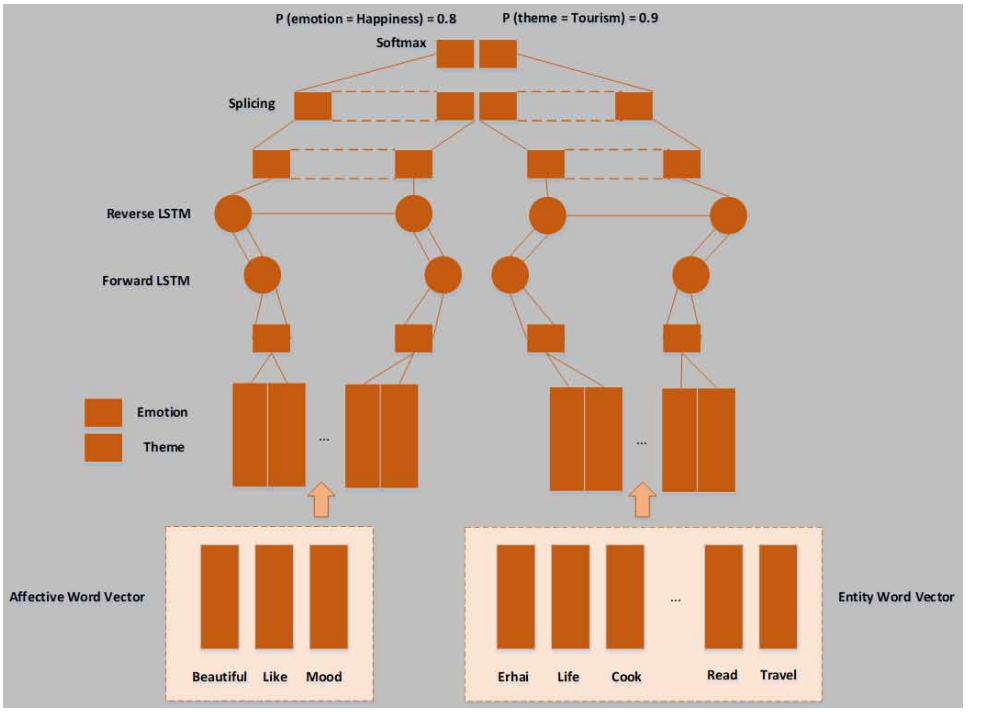


Figure 5. Emotion analysis model



5. EXPERIMENTS

5.1 Data Set Selection

To verify the model and algorithm in this paper, the following three data sets are compared.

(1) IAPS data is a picture database of International Emotional Picture System (IAPS), which contains 8 fine-grained emotional photos, with a total of 956 photos.
(2) Artphoto data set contains some art photos and abstract paintings, which are calibrated by peer rating, including 807 art photos and 228 abstract paintings.
(3) Flickr_LDL data set downloaded more than 30,000 photos from Flickr, and the photos with obvious emotions are calibrated as one of 8 emotions.

The scale of the first two data sets is too small to be suitable for deep learning model training. Therefore, the data set collated by photos on Flickr social networking site is selected as the image emotion training data set, and its emotion classification includes Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear and Sadness.

In this paper, photo metadata is important to obtain contextual information, but more users pay attention to privacy security on social networks. Therefore, as can be seen from Table 1, metadata is missing in many pictures. To solve this problem, we re-screened the data and eliminated the pictures that only contain EXIF. The number of photos in each emotional category has decreased by nearly 2/3. Therefore, for geographic location, weather and other information, they needs to be queried through a third-party API, Baidu map open platform-anti-Geocoding is used for obtaining location information in the experiment. Given the latitude and longitude parameters in the picture EXIF, weather information can be obtained through Baidu map open platform and weather query.

Table 1. Photographs of different emotional categories

Emotional category	Quantity	Quantity with EXIF
Amusement	270843	90271
Awe	362819	130050
Contentment	153920	61121
Excitement	128739	40002
Anger	234322	63020
Disgust	22210	5083
Fear	50302	12038
Sadness	54221	17902

5.2 Parameter Setting

According to the previous description of the model in this paper, the model is divided into three parts: emotion recognition model, scene-Emotion Model and emotion recognition based on multimodal fusion. Because the whole model contains many parameters and the data set is not large enough, it is a great challenge for the training of deep learning. To solve this problem, this paper uses the idea of transfer learning for reference, and transfers the model parameters of image classification to photo sentiment classification, so as to improve the generalization ability and speed up the convergence speed of the model. The specific method is to fix the parameters trained by vgg19 on Image-Net, remove its full connection layer, and then set up 1000 neurons in the hidden layer of sentiment recognition network. The whole connection layer is modified to 8 neurons, and the network parameters are initialized with the Gaussian distribution of zero mean and 0.01 standard deviation for the 8 new emotions. The batch of the model is 32, that is, 32 pictures are input in each batch for training, Relu is selected as the activation function. In order to prevent over-fitting, Dropout is adopted, the parameter of Dropout is set to 0.5, and pytorch is used as the deep learning framework. In addition, the training and testing of the model are carried out on the Googlecolab deep learning platform, and the graphics card is TeslaK80 and the memory is 11G.

6. RESULTS AND DISCUSSION

6.1 Model Comparsion

Different models are selected, including low-dimensional image features +SVM classifier method, Deep Senti Bank method (Ristoski et al, 2017), and various basic models and corresponding fine-tuning models based on CNN. Accuracy is selected as the performance index, and the results are shown in Figure 6.

As can be seen from the figure, the accuracy rate of the original model is 43.67%, and the accuracy rate of the scene-emotion recognition model of photographic works after multimodal fusion is higher, reaching 73.9%. Therefore, by adding the scene features and emotion features in the photography process, it is helpful to extract the emotion category of the image more accurately, with the most accurate boundary and the hierarchical change of correlation intensity, which is beneficial to obtain a model with better performance.

Usually, when the number of image frames per second is more than 30, it means that the image processing has a relatively superior real-time. The analysis of Figure 7 shows that the proposed model processes an average of 31.81 f /s image frames per second, while other methods, such as VGG16 and ResNet 50, are all less than 15 f /s. This shows that the proposed model is effective

Figure 6. Accuracy of models under different methods

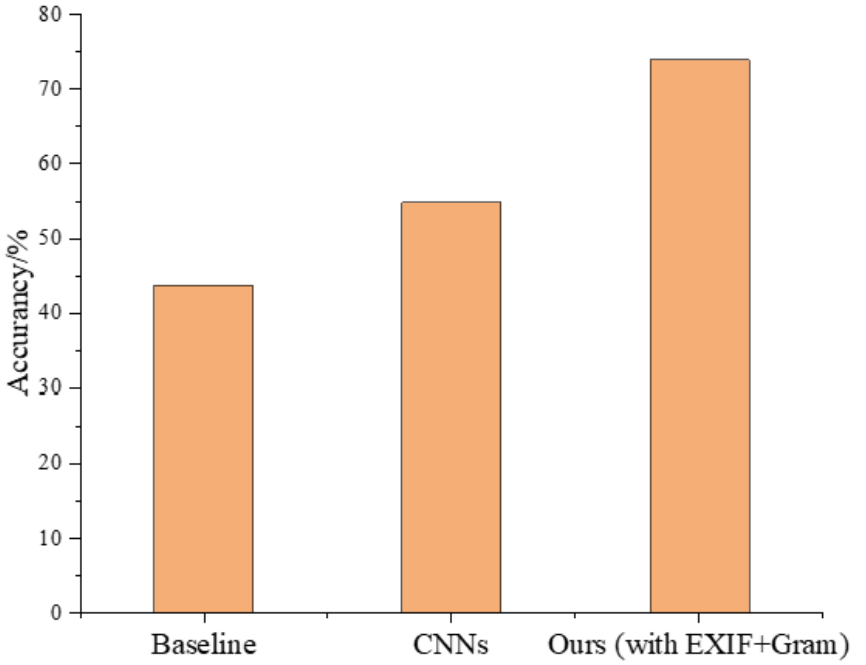
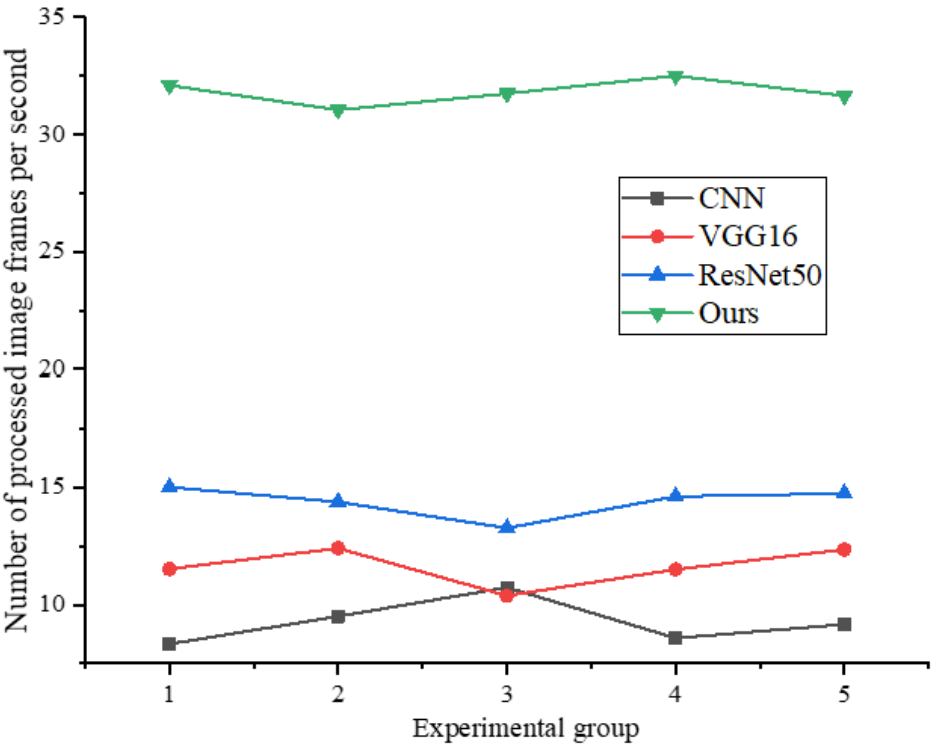


Figure 7. Comparison of image processing speed



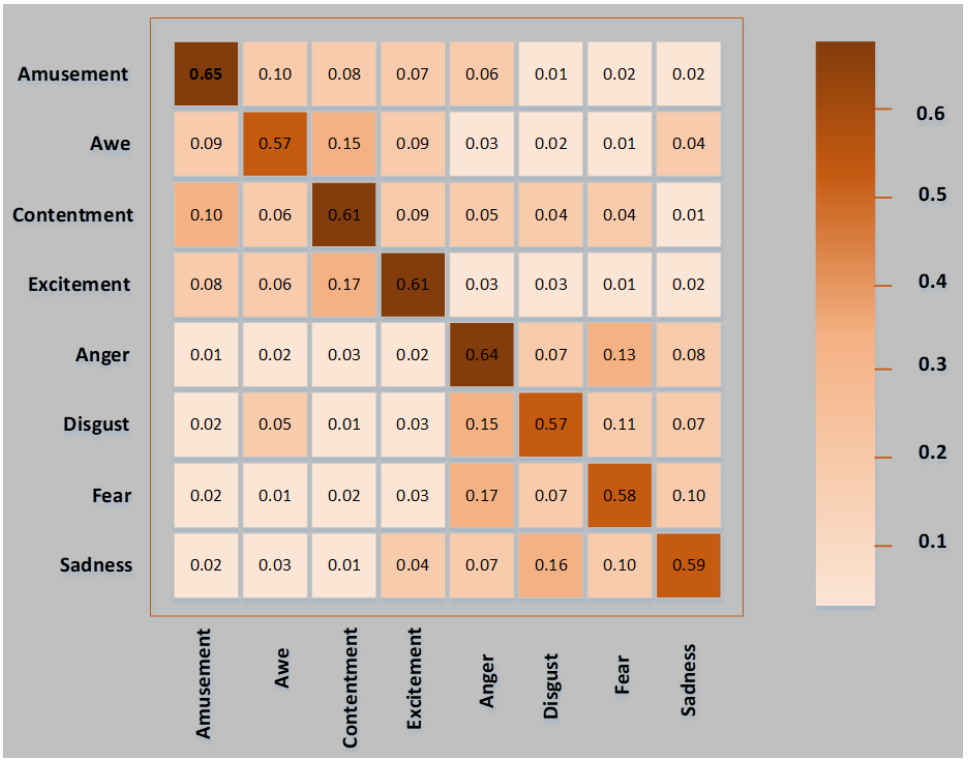
for feature extraction of complex networks. The one-hot situation feature of sparse image metadata is mapped to a dense low-dimensional vector form by embedding, which limits the computing threshold of the target and reduces the training or learning process through similarity comparison, thus reducing the time consuming of the training process and improving the computing efficiency of the target feature calculation.

6.2 Emotional Analysis of Photographic Works

The emotion classification confusion matrix based on the above data set is shown in Figure 8.

As can be seen from the figure, the error rates of Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear and Sadness in photographic works are 0.65, 0.57, 0.61, 0.61, 0.64, 0.57, 0.58 and 0.59 respectively. As a stable index of emotional health, positive emotion and negative emotion are formed in the acquired environment, and they are also indicators of individual adaptability on the level of consciousness, not spontaneous or unconscious behaviors driven by emotion. The balance between the two is called emotional balance. Individuals with a high degree of emotional balance often experience more positive emotions. However, individuals with low emotional balance often experience more negative emotions. Liang (2014) have reported that the higher an individual's satisfaction with the overall life, the more positive emotions people will experience, while the less negative emotions they will experience, and the stronger his happiness experience. Therefore, this model is helpful to analyze the emotional changes of users in the process of photography, and solve their internal emotional health problems.

Figure 8. Confusion matrix of emotion classification



6.3 Distribution of Scenes-Emotion Characteristics in Different Photography Works

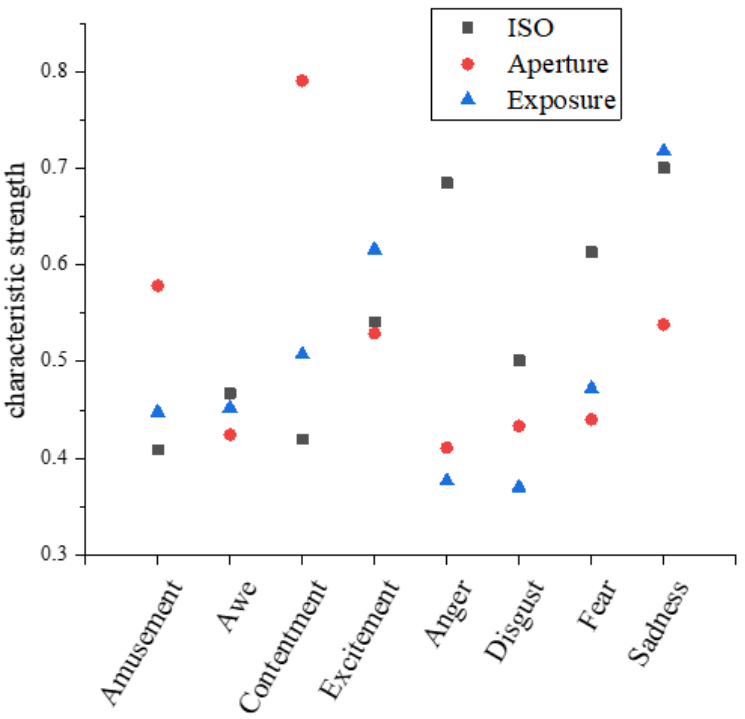
The results of distribution of scenes-emotion characteristics in different photography works are shown in Figure 9.

It can be seen from the data in the figure that the distribution of these features is different in photos with different emotions, which indicates that they have certain distinguishing ability for emotions. For example, awe works are mostly landscapes, so the aperture value is obviously lower than other categories; In a dark room or when shooting night scenes, the dark environment is easy to make people feel comfortable and safe.Therefore, under this category, the characteristic intensity of satisfaction class is larger, and the number of works is larger. In addition, the sensitivity has a strong correlation with the characteristics of sad emotions. For example, when shooting in foggy and snowy days, the whole environment may become gray and fuzzy, and the captured pictures are more textured, and it is easy to make people hesitate and feel sad.

7. CONCLUSION

Emotion analysis in the process of photography can build a bridge between measurable signals and perceived signals' expected emotional state to users, thus realizing the mapping between figurative photographic works' features and abstract human emotions. In this paper, the basic structure of vgg19 is optimized by CNN, and the photographic scene recognition model and visual emotion analysis model are constructed.The mapping relationship between scene and emotion is established. The model verification results show that the accuracy of the scene-emotion recognition model of photographic works after multimodal fusion is high, reaching 73.9%, and the analysis accuracy of positive emotions (such as fun) is higher than that of negative emotions. This model is helpful to analyze the emotional

Figure 9. Distribution of scenes-emotions characteristic in different works



changes of users in the process of photography, and judge their internal problems of emotional health through specific photography scenes and content, and adding the scene features and emotional features in the process of photography is helpful to extract the emotional categories of photos more accurately, with the most accurate boundary and the level change of the relevant strength, which is beneficial to obtain a model with better performance. However, this work still has some limitations. In the specific photography scene, there will be the operation of broadening or narrowing the field of vision. When the field of vision changes, it is necessary to ensure that the designed image processing method still maintains a relatively high precision recognition effect.

ACKNOWLEDGMENT

I would like to thank Shandong University of Arts for its strong support for this work.

CONFLICTS OF INTEREST

The author has no conflict of interest.

FUNDING AGENCY

The work was not funded.

REFERENCES

Acar, E., Hopfgartner, F., & Albayrak, S. (2016). A comprehensive study on mid-level representation and ensemble learning for emotional analysis of video material. *Multimedia Tools and Applications*, *76*(9), 1–29.

Bhunia, A. K., Sain, A., Shah, P. H., Gupta, A., Chowdhury, P. N., Xiang, T., & Song, Y. Z. (2022, October). Adaptive fine-grained sketch-based image retrieval. In *Computer Vision–ECCV 2022: 17th European Conference Proceedings,* (pp. 163-181). Cham: Springer Nature Switzerland.

Burkitt, I. (2002). Complex emotions: Relations, feelings and images in emotional experience. *The Sociological Review*, *50*(S2), 151–167.

Cai, G., He, X., & Chu, Y. (2019). Visual emotion analysis of image embedding in whole and local areas. *Computer Applications (Nottingham)*, *39*(8), 2181–2185.

Chatterjee, A., Gupta, U., & Chinnakotla, M. K. (2019). Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, *93*, 309–317.

Hossain, N. Z., Jenia, T. N., Rahman, S., & Hasan, M. M. (2020). A context-based searching technique by extraction and fusion of meta data of digital photos. In *Proceedings of the International Conference on Computing Advancements* (pp. 23-34). IEEE.

Li, D., Rzepka, R., Ptaszynski, M., & Araki, K. (2020). HEMOS: A novel deep learning-based fine-grained humor detecting method for sentiment analysis of social media. *Information Processing & Management*, *57*(6), 102290.

Li, L., Zhu, X., & Hao, Y. (2019). A hierarchical CNN-RNN approach for visual emotion classification. *ACM Transactions on Multimedia Computing Communications and Applications*, *15*(3s), 1–17.

Li, Z., Xu, H., & Duan, B. (2019). Research on Image Emotional Feature Extraction Based on CNN Model of Deep Learning. *Library and Information Work*, *63*(11), 96–107.

Liang, H. (2014). *Attention Bias of Individuals with High Positive Emotion to Emotional Stimulation with Different Titers* [Master's thesis, Shaanxi Normal University].

Liu, J., & Wu, X. (2020). Multi-modal emotion recognition and spatial tagging based on long-short memory network. [Natural Science Edition]. *Journal of Fudan University*, *59*(05), 565–574.

Luddecke, D., Seidl, C., & Schneider, J. (2018). Modeling context-aware and intention-aware in-car infotainment systems. *Software & Systems Modeling*, *17*(3), 973–987.

Meng, X., Yang, W., & Wang, T. (2021). Review of sentiment analysis research based on image-text fusion. *Computer Applications (Nottingham)*, *41*(02), 307–317.

Rao, T., Li, X., & Xu, M. (2016). Learning multi-level deep representations for image emotion classification. *Neural Processing Letters*, *11*(12), 1–19.

Ristoski, P., Faralli, S., Ponzetto, S. P., & Weikum, G. (2017). Large-scale taxonomy induction using entity and word embeddings. In *International Conference on Web Intelligence* (pp. 81-87). IEEE.

Subudhi, B. N., Veerakumar, T., Esakkirajan, S., & Ramakrishnan, S. (2019). Context dependent fuzzy associated statistical model for intensity inhomogeneity correction from magnetic resonance images. *IEEE Journal of Translational Engineering in Health and Medicine*, *7*, 1–9. PMID:31281739

Wang, Y. (2023). Exploration on the Operation Status and Optimization Strategy of Networked Teaching of Physical Education Curriculum Based on AI Algorithm. [IJITSA]. *International Journal of Information Technologies and Systems Approach*, *16*(3), 1–15.

Wei, L., Yu, H., & Li, B. (2022). Advanced Artificial Intelligence Model for Financial Accounting Transformation Based on Enterprise Unstructured Text Data. [JOEUC]. *Journal of Organizational and End User Computing*, *34*(8), 1–15.

Wu, Q., Liu, Q., & Wang, L. (2016). Situational big data modeling and its application in user behavior prediction. *Big Data*, *2*(6), 110–117.

Xu, C., Cetintas, S., & Lee, K. C. (2014). *Visual sentiment prediction with deep convolutional neural networks*. arXiv preprint arXiv:1411.5731.

Yan, J. (2018). *Research on facial expression recognition in natural scenes based on deep learning* [Doctoral dissertation, Southeast University].

Yu, X. (2022). Global Multi-Source Information Fusion Management and Deep Learning Optimization for Tourism: Personalized Location-Based Service. [JOEUC]. *Journal of Organizational and End User Computing*, *34*(3), 1–21.

Zhou, Y., Guo, J., & Zhu, R. (2016). Landscape image classification based on multi-feature and support vector machine. *Computer System Application*, *25*(5), 135–141.