Journal of Theoretical and Applied Information Technology
31st July 2022. Vol.100. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A HYBRID PLACESNET-LSTM MODEL FOR MOVIE TRAILER GENRE CLASSIFICATION

DAYOU JIANG [1]

[1]Anhui University of Finance and Economics, Department of Computer Science and Technology, Bengbu, 233000, China

E-mail: ybdxgxy13529@163.com

ABSTRACT

Video analysis technology has always been an essential branch of computer vision. Analyzing the movie genres is beneficial to pushing relevant and exciting content to target customer groups to achieve precision marketing. There are some researches on movie trailers to classify movie genres. However, most of them are based on movies' auditory and visual content using various machine learning models or neural network models for classification. This paper considers the features learned using scene-based neural network models in movie genre classification. This paper proposes a hybrid PlacesNet-LSTM (long short-term memory) model for movie trailer genre classification. To compare the performance, the paper also studies two schemes using various video and audio features based on multiple machine learning models and LSTM, respectively. The experimental results show that the PlacesNet-LSTM model on scene recognition achieves the best classification performance in various combinations.

Keywords: *Movie, Genres Classification, Machine Learning, Long Short-Term Memory Networks, Scene Recognition*

1. INTRODUCTION

With the exponential growth of video in sharing platforms, many applications such as video recommendation, movie trailer making, digital storytelling, video retrieval, and social media for relevant and exciting content for specific target users are attracting the attention of major manufacturers. As an essential branch of computer vision, video analysis has important practical significance for solving various problems. As one of the most common manifestations to satisfy people's entertainment, movies play an increasingly important role in both commercial economy and cultural and artistic fields. Large-capacity data sets about the film, such as Internet Movie Database (IMDB) [1], MMTF -14K [2], also came into being.

A movie trailer needs to condense the movie's story, showing the best parts of the film in a short period to attract the audience to the greatest extent and induce the audience to go to the theater. The use of movie trailers to identify movie categories has also become a video analysis. There are mainly recognition methods based on text, audio, and video alone or in combination [3-4]. With computer technology and artificial intelligence development, many new mining methods have transitioned from machine learning to deep learning. The CNN-MoTion [5] model proposed by Wehrmann uses SVM to perform subsequent classification learning and training on CNN features. After that, the team proposed a time-space neural network model based on CTT-MMC [6], in which the CTT module adopts an ultra-deep residual network. Sivaraman [7] verified that the performance of LSTM is better than the VGG16 neural network in the classification of movie subjects. Ben-Ahmed [8] used two different neural network models to extract auditory and visual features of movie clips, respectively, and utilized LSTM [9] learning on visual elements and SVM learning on acoustic features for genre prediction. The method proposed by Shambharkar [10] selects segments of a video sliding window to train a 3D neural network in time-space, which improves the recognition rate compared with the neural network based on the 2D image. R. B. Mangolin [11] extracted multi-modal information from movie-related audio, posters, subtitles, introductions, and movie images, trained with various machine learning methods, and fused the integrated results.

Existing video genre classification methods use machine learning and deep learning in feature engineering and classification problems. Still, they

rarely consider the possibility of scene recognition for subject classification. This paper studies two schemes based on machine learning and LSTM, which combine multimodal video content features and diverse machine learning techniques to compare and analyze the movie trailer genre recognition.

The main contributions of this paper are in two aspects:

1) A movie trailer genre classification scheme based on the hybrid PlacesNet-LSTM model is proposed, which uses the scene-based video features extracted from pre-trained PlacesNet [12], and achieves the best movie genre classification performance compared to other combinations.

2) The performance of movie trailer genre recognition combined with multiple machine learning techniques and multimodal features is studied. The recognition performance using CNN-based visual features is significantly better than thing low-level visual features. The performance of Random Forest (RF) [13] and Extreme learning machines (ELM) [14] also holds an advantage.

2. BASIC THEORY AND TECHNOLOGY

2.1 Feature Extraction

Color Histogram [15] describes the frequency of pixels with the same color in a given image. The color composition distribution of an image is less sensitive to changes in image quality (such as blur) relative to image rotation and magnitude. Geometric transformations such as small translation and scaling are also insensitive and computationally inexpensive, so they are often used in image retrieval.

Local Binary Pattern (LBP) [16] consists of relative values by comparing each pixel with its neighbors. It is commonly used to classify textures in computer vision and has the characteristics of low computational cost and resistance to shifting gray image values.

Oriented gradient histogram (HOG) [17] counts the gradient orientation in the local part of the image. Unlike edge orientation histograms, etc., it is computed on a dense grid of evenly spaced cells and uses overlapping local contrast normalization to improve accuracy.

The GIST descriptor [18] summarizes the gradient information of image scale and orientation, which can provide a rough description of the scene. By convolving four scales, eight orientation images with 32 Gabor filters, and 32 feature maps of the same size as the input image can be generated. By dividing each feature map into 16 regions by a 4×4 grid and averaging the features within the areas, a GIST descriptor of length 512 can be obtained.

The dual-complex wavelet transform based on singular value decomposition (DTCWT-SVD) [19] method not only maintains the translation invariance of the dual-complex wavelet transform and has good direction analysis ability; at the same time, the combination of singular value decomposition not only simplifies the dimension of the feature but also removes the noise.

Mel-scale Frequency Cepstral Coefficients (MFCC) [20] is a linear transformation of the logarithmic energy spectrum based on a nonlinear Mel scale of sound frequency. The frequency band division of the Mel-frequency cepstrum is equidistantly divided on the Mel scale, which approximates the human auditory system more closely than the linearly spaced frequency bands used in the normal logarithmic cepstrum, so it is widely used in speech recognition functions.

2.2 K-Means Clustering Algorithm

The K-Means algorithm [21] clustering divides N instances into K disjoint “sample clusters” so that each point belongs to the cluster corresponding to the nearest mean. It is a simple algorithm that can quickly and efficiently cluster such data and be used in data analysis, customer segmentation, recommender systems, search engines, image segmentation, semi-supervised learning and dimensionality reduction, etc. It is often used as a preprocessing step for other algorithms. Compared with the Mean Shift clustering algorithm, the K-Means algorithm is also suitable for large datasets. K-Means is not perfect, and it is challenging to avoid suboptimal solutions within a limited number of times. In addition, the number of clusters k needs to be pre-specified as an input parameter, and choosing an inappropriate value of k may lead to poor clustering results.

2.3 Classification Algorithm

Support vector machine (SVM) [22] In addition to linear classification, SVM can also use the kernel trick to perform a nonlinear sort, implicitly mapping its input into a high-dimensional feature space. When the training set is vast, when the size is large, or if there are many features, the linear sum function is often used. Otherwise, the Gaussian RBF kernel is often used. SVMs can achieve significantly higher search accuracy than traditional query optimization schemes for image classification

problems. SVM is very sensitive to the scaling of features and needs to be normalized.

K-nearest neighbors algorithm (KNN) [23] is a kind of learning based on local instance approximation. In K-NN classification, the classification of an object is determined by a "majority vote" of its neighbors, and the most common type among the K nearest neighbors determines the class assigned to the thing. If K=1, the closest node directly assigns the class of the object. Although the algorithm is simple in principle, it is computationally intensive and time-consuming. Furthermore, the presence of noisy and non-correlated features or feature scales inconsistent with their importance can seriously degrade the accuracy of the K-Nearest Neighbors algorithm.

The logistic regression classification algorithm [24] is a predictive analysis algorithm based on the concept of probability, which uses the Logistic Sigmoid function to transform its output to return a probability value. The Logistic Sigmoid function is a function that resembles an "S"-shaped curve when drawn on a graph. It takes a value between 0 and 1 and "squeezes" it to the top and bottom margins, marking it as 0 or 1. Although logistic regression is best suited for instances of binary classification, it can also be applied to multiclass classification problems by combining multiple binary classifiers.

Softmax regression [25], also known as multiple logistic regression, extends the logistic regression model to multi-classification problems. Given an instance, the Softmax regression model first computes the coefficients for each class, then applies the softmax function to these coefficients to estimate the probability of each class.

Artificial Neural Networks (ANNs) [26] are vaguely inspired by biological neural networks that constitute animal brains. Artificial neural networks consist of a series of simulated neurons. The simplest types have one or more static components, including the number of units, layers, unit weights, and topology. Artificial neural networks can replicate and model nonlinear processes.

Extreme learning machines can be used for classification, regression, clustering, sparse approximation, compression, and feature learning tasks with single or multiple layers of hidden nodes. In most cases, ELMs are used as single hidden layer feedforward networks (SLFNs), including but not limited to sigmoid networks, RBF networks, threshold networks, fuzzy inference networks, complex neural networks, wavelet networks, Fourier transforms, and Laplacian networks. In the classification task, given any non-constant piecewise continuous function as the activation function in SLFN, adjusting the parameters of the hidden nodes can make the SLFN approximate any objective function. The SLFN with the hidden map can be of any shape and divided into arbitrary disjoint regions.

Decision Trees (TREE) [27] are constructed using a heuristic method called recursive partitioning (also known as divide and conquer). Because it splits the data into subsets, then repeatedly splits it into smaller subsets, and so on, until the algorithm determines that the data within the subsets is sufficiently uniform, until the process stops, or some other stopping condition is met. In practice, it is sometimes necessary to limit the depth of the tree to prevent overfitting.

Random Forest is an ensemble learning method that constructs many decision trees during training and outputs them as classification or regression. Random decision forests correct the habit of decision trees to overfit their training set. It is a weighted average where each node has a weight equal to the number of training samples associated with it, making it easier for decision trees to measure the relative importance of each feature.

3. EXPERIMENTAL SCHEME OF VIDEO GENRE CLASSIFICATION

This section will introduce the scheme based on the combined multimodal features and multi-class machine learning models used in the comparative research on video genre classification

3.1 Solutions based on machine learning

The comparative experimental scheme based on machine learning training is as follows:

1) Video preprocessing is divided into video transcoding, removal of irrelevant content, size adjustment, and keyframe extraction.

2) Video feature extraction extracts the visual and auditory features of the video, respectively. Visual elements are further divided into shallow optical parts and CNN-based in-depth components.

3) Use the K-Means clustering algorithm to divide the features into different numbers of “sample clusters” ranging from 10 to 200 with an interval of 10.

4) Generate histogram features based on each video's corresponding number of categories

according to the number of different clustering categories.

5) Use the category histogram feature of the video to train under different supervised learning classification methods, and use the trained classification model to detect the videotape for verification.

The experimental program flow using machine learning is shown in Figure 1:

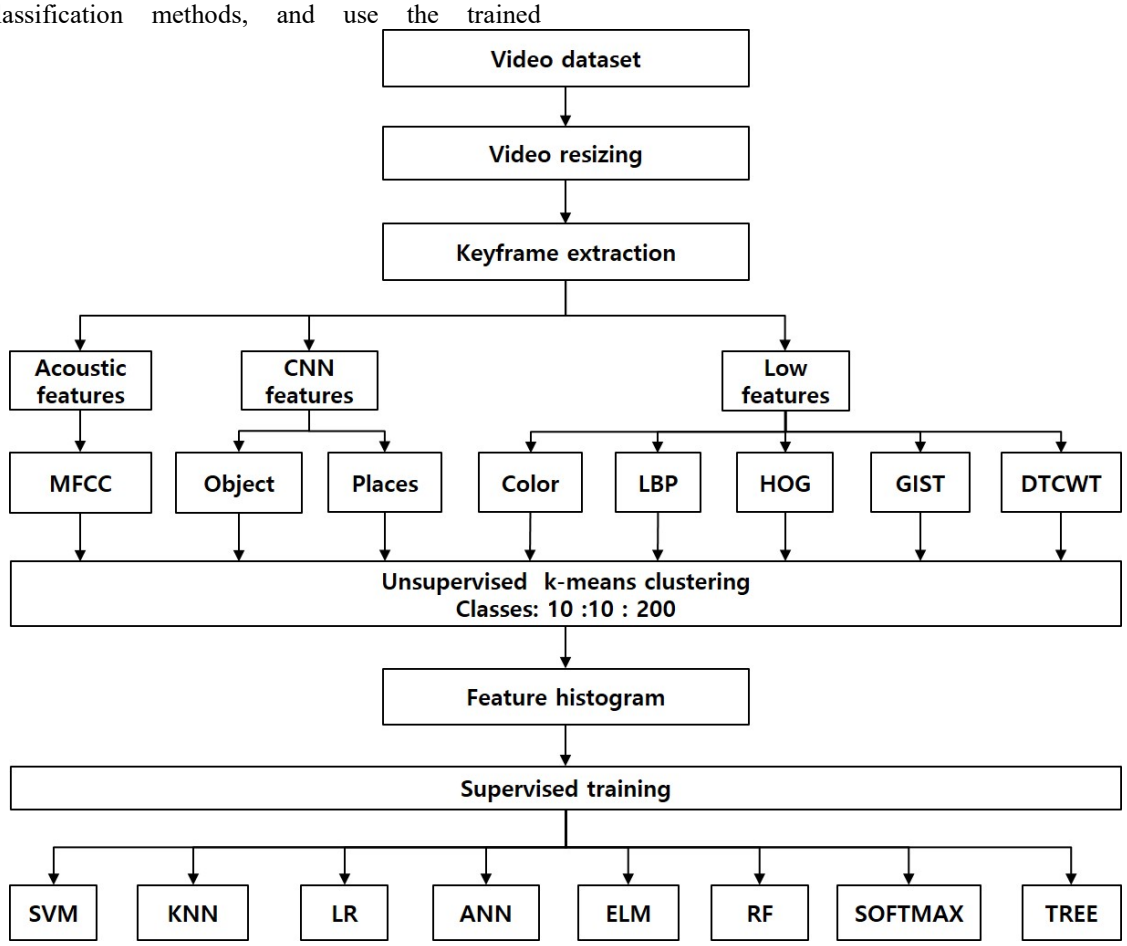

Figure 1: Flowchart of Video Genre Classification Based on Machine Learning

3.2 Solutions Based on LSTM Network

A bidirectional LSTM network is a recurrent neural network (RNN) that learns long-term dependencies between time steps of sequence data. In the experiment, the feature sequence data of keyframes in the video is input into the LSTM network for training.

The comparative experimental scheme based on LSTM network training is as follows:

1) Video preprocessing is divided into video transcoding, removal of irrelevant content, size adjustment, and keyframe extraction.

2) Video feature extraction extracts the visual and auditory features of the video, respectively. Visual elements are further divided into shallow optical parts and CNN-based in-depth components.

3) According to the time sequence, the features of the keyframes are combined into a time series.

4) Use the LSTM network model to train the time-series features of each video and use the trained network model to detect the tapes used for verification.

The experimental program flow is shown in Figure 2:

Journal of Theoretical and Applied Information Technology
31st July 2022. Vol.100. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

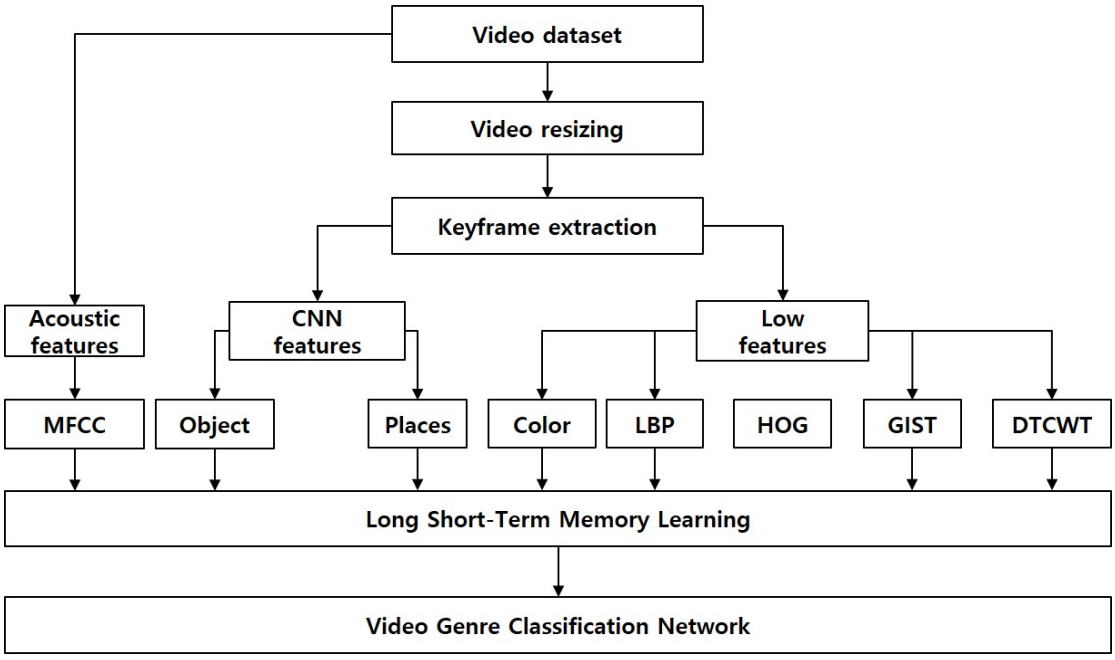

Figure 2: Flowchart of Video Genre Classification Based on LSTM.

4. EXPERIMENTS AND RESULTS

This section tests the dataset using the two proposed movie trailer genre classification recognition schemes.

4.1 Video Preprocessing

4.1.1 Building the dataset

Movies can often be categorized based on setting, theme, emotion, format, target audience, and budget. This experiment locks the genres of movie trailers into four categories: "action," "comedy," "drama," and "horror." Find the videos that do not have overlapping categories in the four genres on the genre search page of the IMDb website, record the title, production year, and genre category, and then obtain the corresponding download address of the movie trailer from the YouTube website, and link the video to the address Add it to the database, and finally use the you-get [28] tool to download videos in batches according to the video link address. The data contains 800 video trailers with a single genre label and 200 videos in each category. These 800 videos have 2,541,329 frames and a total duration of 28.615 hours (here is the time of the video clipping). The histogram of video duration distribution is shown in Figure 3.

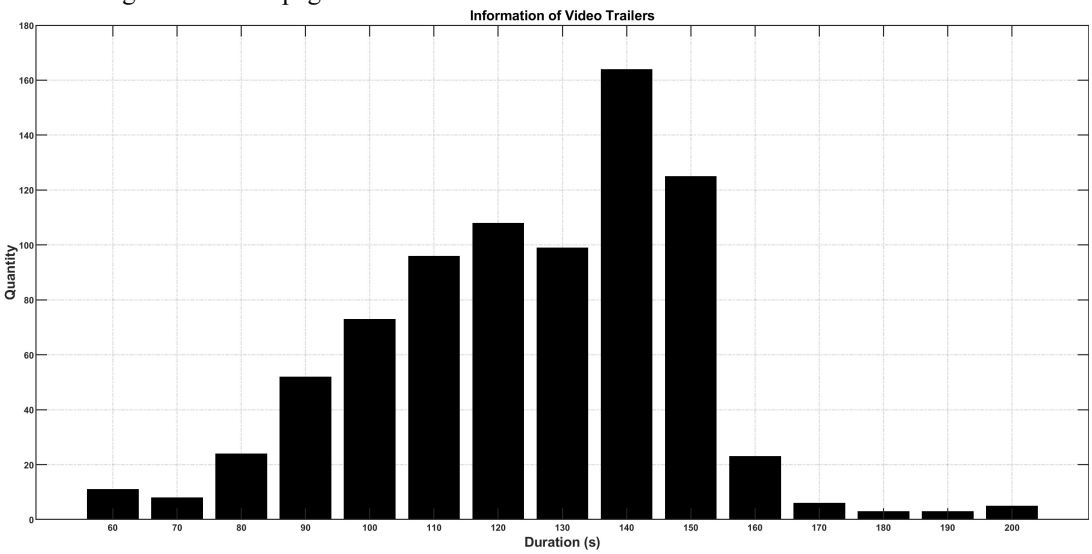

Figure 3: Duration Distribution of Trailer Video Dataset

4.1.2 Video transcoding, cropping and resizing

Use FFmpeg [29] software to transcode downloaded movie trailers that are not in MP4 format into MP4 format and resize the video in the dataset to the size of the CNN training network used. Considering the influence of movie image quality and training overhead, such as time and memory capacity, the CNN network model selected in the experiment is ResNet50 [30]. The input image size is 224×224×3. Video quality is reduced after resizing. Since the title and end of a movie trailer generally contain relevant information about the film company, it is also necessary to intercept the video and remove the label and ending parts.

4.1.3 Select keyframes

Keyframe extraction is commonly used in the field of video retrieval, and there are the following methods:

1) Based on the shot detection method, the first and last keyframes are selected from each shot, and it takes an average of 60 seconds to process each video on the dataset.

2) Select keyframes based on the interval. This method does not require lens analysis and directly selects a frame of images every 2 seconds. It takes 8 seconds to process each video on the dataset.

The number of keyframes selected by the shot detection method will directly affect the accuracy of genre classification because the number of frames chosen at this time is significantly less than the interval selection method. The early stage's experimental results on the small data set also showed that two kinds of keyframes were used. The classification accuracy of the selected methods differs very little. Considering the computational cost, interval selection of keyframes is adopted in the experiment, and each movie trailer set contains an average of 66 frames of images.

After video preprocessing, figure 4 shows the keyframes extracted from the action film "American Sniper." Although some unnecessary content can be removed after rough processing of the title and ending, there are also some irrelevant contents in the trailer, such as the screen when the movie content transitions are entirely black, or only text is displayed. Considering the computational cost factor, no removal processing is performed.

Figure 4: Sample Keyframes Extracted from The Trailer of American Sniper

4.2 Feature Extraction

Extracting visual and auditory features of videos is aimed at classification learning. Several shallow visual features are selected in the experiments, such as color histogram, local binary map, directional gradient histogram, Gist descriptor, DTCWT-SVD transform coefficients, etc. As a comparison, two CNN features based on objects and scenes are also used. The object-based network ResNet50 is trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset, and it can detect 1000 objects such as keyboards, mice, pencils, and many animals. The scene-based PlaceNet network proposed in previous research is mainly trained on the Places365 dataset to recognize scenes in images, such as food courts, cafeterias, canteens, fast food restaurants, etc. In addition, Mel-frequency cepstral coefficients are used to extract audio features from the video. The dimensions of each feature are listed in Table 1.

Table 1: The Name and Dimensions of Various Features.

Feature names	Dimensions
Color Histogram	256
Quantized Uniform LBP	58
HOG	31
Gist Descriptor	512
DTCWT-SVD	52
Fused shallow	909
Object-CNN Fc	1000
Places-CNN Fc	434
Fused CNN Fcs	1434
MFCC	800

4.3 Genre Classification Results of Movie Trailers

The experiment uses a 2-fold cross-validation method, alternately using 400 movie trailer videos for training and the remaining 400 videos for

testing, which requires 100 videos for each genre type during training and testing.

4.3.1 Results based on machine learning training

First, the features of each keyframe of the video are classified according to the K-means clustering algorithm to generate the overall feature category of the video data set. Then the feature category histogram is developed according to the number of features of the keyframes in each video. At this time, the feature category histogram of each video is used as the final input value of each video for machine learning training. The recognition accuracy of genre categories (the percentage of the total number of predicted types being actual categories in the total number of detections) obtained by 2-fold cross-validation using various classification methods are shown in the table. The minimum and maximum values in Table 2 refer to the accuracy rate of genre classification of movie trailers. The classification method and K-Means category length results when the accuracy rate takes the maximum value.

Table 2: The Movie Genres Classification Results of Trailers Using Different Dimensional Features Under Different Classification Methods.

Feature names	Min (%)	Max (%)	Method	Bins
MFCC	34.4	54.6	RF	190
Fused shallow	37.6	55.8	RF	90
Object-CNN Fc	47.1	69.7	RF	130
Places-CNN Fc	44.5	72.6	RF	130
MFCC+ Fused shallow	39.6	71	RF	170
MFCC+ Fused shallow + Object CNN Fc	40.6	71.1	RF	90
Fused CNN Fcs	45.4	71.8	ELM	180

Using this scheme, on the one hand, when the video material is single, and the number is scarce, the number of distinguishable feature categories that the K-Means clustering algorithm can generate is limited, directly affecting the final experimental results. On the one hand, the overhead of the whole process increases with the length of the class. In the experiment, only the fused shallow visual features are processed without considering the single external visual elements, saving the cost and making it more comparable to the parts generated by the CNN network. From the above table, it can be seen that:

1) The classification accuracy based on scene CNN coefficient features is the highest, and it is also higher than other fusion-based features.

2) The highest result obtained based on the Mel cepstral coefficient feature is less than 55%, and the length of the K-Means category used is also the largest, which is the worst in terms of efficiency.

3) The results obtained with features based on CNN coefficients are higher than those obtained with shallow visual features and auditory features, both minimum and maximum values. The required K-Means category length is also moderate.

4) Compared with other classification algorithms, the random forest algorithm shows better classification performance.

5) Fusion of multimodal features cannot improve the classification accuracy.

The above differences are mainly due to the number of clustering categories used and the classification methods. Even with the better performing random forests and extreme learning machines for training, the varying number of classes divided by clusters from 10 to 200 makes for a significant gap in the results.

In addition, the sound in the movie trailer contains various morphological signals such as speech, music, and sound effects, and different videos have unique speech and noise. This also leads to unsatisfactory experimental results based on the characteristics of Mel cepstral coefficients.

4.3.2 Results based on LSTM learning and training

First, the features of each keyframe of the video are combined into a time series in chronological order. If the number of keyframes in a video is 60, and the feature length of a single keyframe is 434, the size of the time series of this video is 434×60. When the input LSTM is based on the scene network PlacesNet feature, the parameters of the LSTM network are set to: the feature-length is 434, the target category is 4, the model is a bidirectional LSTM, the max-epochs is 100, the mini-batch size is 32, the optimizer uses Adam, the hidden unit number is 100.

Journal of Theoretical and Applied Information Technology
31st July 2022. Vol.100. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The results of this experiment are displayed in the form of a confusion matrix. In predictive analysis, the confusion matrix reports the number of false positives, false negatives, true positives, and true negatives, which allows for classification accuracy analysis. Accuracy is also not a reliable indicator of a classifier's actual performance. It can produce misleading results if the number of observations in the different classes in the dataset varies significantly. The confusion matrix obtained by extracting coefficient features using PlacesNet and feeding the feature sequence into the LSTM network is shown in Figure 5. The figure shows that the horizontal direction is the fundamental four target categories, and the vertical direction is the d corresponding category. Taking a drama film as an example, the number of true positives is 59, false positives are 25 (i.e., 11+13+1), and false negatives are 41 (i.e., 8+20+13). The number of true negatives is 275 (i.e., 400-100-25).

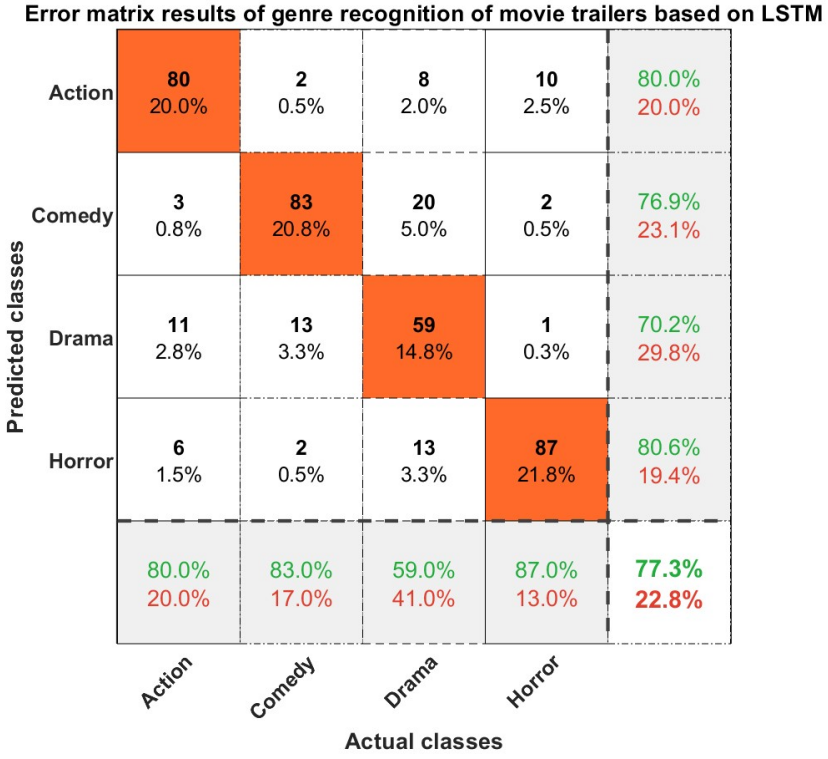

Figure 5: Confusion Matrix Diagram of Movie Genre Classification Using Trailer Based On Placesnet-LSTM

In classification problems, the commonly used judgment indicators are as follows:

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \quad (1)$$

$$recall = \frac{true\ positives}{true\ positives + false\ negatives} \quad (2)$$

$$specificity = \frac{true\ negatives}{true\ negatives + false\ positives} \quad (3)$$

$$accuracy = \frac{true\ positives + true\ negatives}{total} \quad (4)$$

Recall can be used as a quantitative indicator for avoiding false negatives, while specificity can be used as a quantitative indicator for false positives. The specificity of drama films is 91.6%, which shows that the proportion of other categories of movies misclassified as drama films is relatively low. However, its recall rate is only 59%, which shows that a considerable ratio of drama films is misjudged as other movie categories.

It can be seen from Figure 5 that:

1) The recall rates for classifying the four genres are 80%, 83%, 59%, and 87%.

2) The precision rates for the four genres are 80%, 76.9%, 70.2%, and 80.6%, respectively.

3) The number of videos predicted to be drama (84) is significantly lower than the other three categories (100 for action, 108 each for comedy and horror).

4) Ignoring the influence of true negatives, only considering actual positive individuals, and taking the recall rate as the classification accuracy, the average classification accuracy is 77.3% (if the true negatives are included, the average accuracy is 88.6%).

The relatively low recall rate of the plot rate is related to the characteristics of the trailer. Drama films mainly use the plot changes of the story or the development of characters to drive the progress of the whole movie. Compared with other categories, the rhythm will be slowed down, and the plot will be relatively compact. However, the trailer of a drama film also needs to be edited to more selling points in a short period, which disrupts the rhythm of the drama film and reduces the distinction from other categories, resulting in misjudgment.

Next, a comparative test is performed, and the features extracted by different feature extraction algorithms are combined into a time series for LSTM training and detection. The results obtained are shown in Table 3.

Table 3: The Movie Genres Classification Results of Trailers Using Different Features Based On LSTM.

Feature names	Accuracy (%)
Color Histogram	54.6
Quantized Uniform LBP	59.6
HOG	56.9
Gist Descriptor	52.7
DTCWT-SVD	51.7
Fused shallow	55.8
Object-CNN Fc	67.5
Places-CNN Fc	77.3
Fused CNN Fcs	72.5
MFCC	53.5

It can be seen from Table 3 that:

1) Under the training of the LSTM network, based on in-depth features, it has better video genre classification performance than shallow content features.

2) Using scene-based depth features has higher classification accuracy than object-based depth features.

3) Using fused shallow content features is not more accurate than using single external features. Using combined deep features is also less accurate than using scene-based in-depth features.

5. CONCLUSIONS

This paper proposes a hybrid PlacesNet-LSTM model for movie trailer genre classification and studies two video genre classification schemes using multimodal visual features and audio features based on multiple machine learning techniques. The other is based on LSTM architecture. Comparative experimental results show that the hybrid PlacesNet-LSTM model has significant advantages for movie trailer genre classification.

The articles [3] and [5] are based on the SVM model, which used low-level features and CNN features. Different from them, this paper tested eight other video and audio features (including low-level features such as Color, LBP, HOG, GIST, DTCWT, MFCC, and two CNN-based parts: Object-CNN, PlacesNet) and studied eight different machine learning model techniques (including SVM, KNN, LR, ANN, ELM, RF, SOFTMAX, TREE) and LSTM architectures for classification training. The same as [4, 8] is that both use the CNN-LSTM model; the difference is that this paper also studies the impact of scene-based CNN features on the classification structure and discusses the extraction of keyframes based on shot detection in the video preprocessing section.

However, the proposed model also has some limitations:

1) The proposed hybrid model is not an end-to-end scheme because the time-consuming image preprocessing is needed;

2) For complex, such as multi-label movie genre classification tasks in the article [7], there is still a large room for improvement in the model's performance.

Future work can start from the following aspects:

1) Optimize the video preprocessing link and establish an end-to-end video recognition model;

2) Integrate multiple information such as semantics for movie trailer genre classification.

REFERENCES:

[1] https://www.imdb.com/

[2] Y Deldjoo, M G Constantin, B Ionescu, et al. "MMTF-14K: a multifaceted movie trailer feature dataset for recommendation and retrieval", *Proceedings of the 9th ACM Multimedia Systems Conference*. 2018, pp.450-455.

[3] Y F Huang, S H Wang. "Movie genre classification using SVM with audio and video features," *International Conference on Active Media Technology*. Springer, Berlin, Heidelberg, 2012, pp.1-10.

[4] C H Chou, P C Jen. "Image-Based Deep Learning Model for Movie Trailer Genre Classification," *2019 IEEE 8th Global Conference on Consumer Electronics (GCCE)*. IEEE, 2019, pp.1064-1066.

[5] G S Simões, J Wehrmann, R C Barros, et al. "Movie genre classification with convolutional neural networks", *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016, pp. 259-266.

[6] J Wehrmann, R C Barros. "Convolutions through time for multi-label movie genre classification", *Proceedings of the Symposium on Applied Computing*. 2017, pp.114-119.

[7] K Sivaraman, G Somappa. "Moviescope: Movie trailer classification using deep neural networks". *University of Virginia*, 2016.

[8] O Ben-Ahmed, B Huet. "Deep multimodal features for movie genre and interestingness prediction", *2018 international conference on content-based multimedia indexing (CBMI)*. IEEE, 2018, pp.1-6.

[9] Z Huang, W Xu, K Yu. "Bidirectional LSTM-CRF models for sequence tagging". arXiv preprint arXiv:1508.01991, 2015

[10] P G Shambharkar, P Thakur, S Imadoddin, et al. "Genre Classification of Movie Trailers using 3D Convolutional Neural Networks", *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2020, pp. 850-858.

[11] R B Mangolin, R M Pereira, A S Britto, et al. "A multimodal approach for multi-label movie genre classification". *Multimedia Tools and Applications*, 2020, pp.1-26.

[12] D Y Jiang, J W Kim. "Video searching and fingerprint detection by using the image query and PlaceNet-based shot boundary detection method". Applied Sciences, Vol. 6, No. 10, 2018, pp.1735

[13] T K Ho. "Random Decision Forests", *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal*, QC, 14–16 August 1995, pp.278–282

[14] G B Huang, Q Y Zhu, C K Siew. "Extreme learning machine: theory and applications". *Neurocomputing*. Vol. 70, No. 1, 2006, pp.489–501

[15] J Huang, S R Kumar, M Mitra, W J Zhu, R Zabih. "Image indexing using color correlograms", *In Proceedings of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, PR, USA, 7–19 June 1997, pp.762–768

[16] T Ojala, M Pietikainen, T Maenpaa. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns". *IEEE Trans. Pattern Anal. Mach. Intell*. Vol. 24, 2002, pp. 971–987

[17] N Dalal, B Triggs. "Histograms of oriented gradients for human detection", *In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 20–25 June 2005, Vol. 1, pp.886–893

[18] J Hays, A A Efros. "Scene completion using millions of photographs". *ACM Trans. Graph*. Vol. 26, 2007, pp.4

[19] D Y Jiang, J W Kim. "Texture Image Retrieval Using DTCWT-SVD and Local Binary Pattern Features". *JIPS*, Vol. 13, No. 6, 2017, pp.1628-1639

[20] J S Bridle, M D Brown. "An Experimental Automatic Word-Recognition System". *JSRU Report No. 1003*, Joint Speech Research Unit, Ruislip, England, 1974

[21] Forgy, W Edward. "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". *Biometrics*. Vol. 21, No. 3, 1965, pp.768–769

[22] C Cortes, Vladimir N Vapnik. "Support-vector networks". *Machine Learning*. Vol. 20, No. 3, 1995, pp.273–297

[23] E Fix, Joseph L Hodges. "Discriminatory Analysis". Nonparametric Discrimination: Consistency Properties (Report). *USAF School of Aviation Medicine, Randolph Field*, Texas, 1951

[24] J S Cramer. "The origins of logistic regression (Technical report)". 119. *Tinbergen Institute*. 2002, pp.167–178

[25] I Goodfellow, Y Bengio, A Courville. "Softmax Units for Multinoulli Output Distributions". *Deep Learning*. MIT Press. 2016: 180–184

[26] P Werbos. "Applications of advances in nonlinear sensitivity analysis". *System modeling and optimization*. Springer. 1982, pp.762–770

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[27] J R Quinlan. "Simplifying decision trees". *International Journal of Man-Machine Studies*. Vol. 27, No. 3, 1987, pp.221–234

[28] https://github.com/soimort/you-get

[29] FFmpeg, https://ffmpeg.org/

[30] K He, X Zhang, S Ren, et al. "Deep residual learning for image recognition", *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp.770-778