Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-16691-1

A survey of micro-video analysis

Jie Guo[1] · Rui Gong[1] · Yuling Ma[1] · Meng Liu[1] · Xiaoming Xi[1] · Xiushan Nie[1] · Yilong Yin[2]

Received: 10 October 2022 / Revised: 17 May 2023 / Accepted: 27 August 2023

Abstract
As opposed to traditional video, a micro-video is a short video that is spread on social platforms. As user-generated contents, micro-videos have stronger social attributes compared to ordinary videos. Research on micro-video analysis has been conducted in both industry and academia and includes venue classification, tag prediction, popularity prediction, action prediction, click prediction, and recommendation. In this paper, we first review the studies on these tasks in terms of micro-video classification, prediction, and recommendation. Thereafter, we present an overview of the methods, features, datasets, and evaluation metrics relating to these studies. Finally, we analyze the challenges of micro-video analysis. Because of the limited research work on micro-video analysis, we can not summarize some aspects of micro-video analysis, such as micro-video classification.We believe that this survey will aid in enhancing the knowledge of researchers and practitioners who are interested in micro-video analysis.

Keywords Venue recognition · Popularity prediction · micro-video recommendation

1 Introduction

Micro-video refers to video clips that are captured by Internet users using mobile phones, computers, cameras, and other video-recording devices, with a duration ranging from a several seconds to a several minutes. Various micro-video platforms have rapidly emerged since 2016. Micro-videos are spread on social platforms and have stronger social attributes compared to ordinary videos. Furthermore, a substantially greater number of micro-videos than ordinary videos exists. As micro-videos are user-generated contents, they are generated by users, not platforms. Therefore, it is necessary for platforms to group micro-videos according to their contents and recommend micro-videos according to user interests.

Research on micro-video analysis has been conducted in both industry and academia and includes venue classification, tag prediction, popularity prediction, action

✉ Xiushan Nie
niexiushan@163.com

Jie Guo
guojiesdu@163.com

[1] Shandong Jianzhu University, Jinan, China

[2] Shandong University, Jinan, China

Published online: 20 September 2023

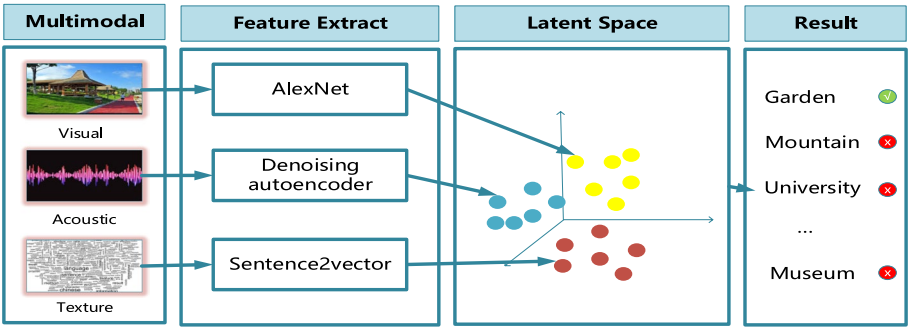

Fig. 1 Micro-video classification framework based on common subspace learning

prediction, click prediction, and recommendation. As micro-videos consist of real-world data, many challenges arise in micro-video analysis: 1) Noise. The noise in micro-video analysis includes vision noise that originates from shaking devices and audio noise. 2) Long-tailed distribution. The number of micro-videos in hot topics is large and the number of micro-videos in non-hot tags is small, which results in long-tailed distributions. 3) Incomplete data. Data are missing in multi-modalities owing to privacy protection or other reasons. 4) Inadequate intra-class compactness. The intra-class compactness may be inadequate as a result of the multiview and hierarchical structure. 5) Inconsistent semantic strength. The semantic strength of different modalities in the same sample is inconsistent as a result of the influence of noise or user privacy protection.

Compared with traditional video, micro-video classification tasks are mostly based on micro-video feature learning representation, so it can better represent micro-video semantics. Visual modalities contain a wealth of semantic information, so some studies on the representation of visual modalities are based on the characteristics of micro-video. Micro-video contains multi-modality information, such as visual modality, acoustic modality, textual modality, etc., and multi-modality information representation can provide richer semantic representation for micro-video. Therefore, most of the existing micro-video classification methods are based on multi-modality fusion, focusing on how to learn the consistency and complementarity between multi-modalities. Taking the venue recognition task as an example (popularity prediction, tag prediction, micro-video recommendation and other tasks are similar to the venue recognition), some researches learn the consistency between the multiple modalities through common subspace learning, as shown in Fig. 1. The learning method of common subspace can learn the consistency between multiple modalities well, but it cannot achieve the balance of consistency and complementarity between the multiple modalities well. In order to better learn the consistency and complementarity between multiple modalities, some studies carried out micro-video classification tasks by connecting features in series, as shown in Fig. 2.

In recent years, research on micro-video analysis has focused on real-world data. Several studies have considered the above challenges and proposed corresponding methods. Most of the studies focused on micro-video venue classification, recommendation, and other specific applications. Therefore, in this paper, we review the existing studies on micro-video classification, prediction, and recommendation.

Sections 2–6 extensive discuss micro-video analysis. The existing relevant studies, experiments conducted, and limitations have been surveyed well. Section 7 is the conclusions of micro-video analysis.

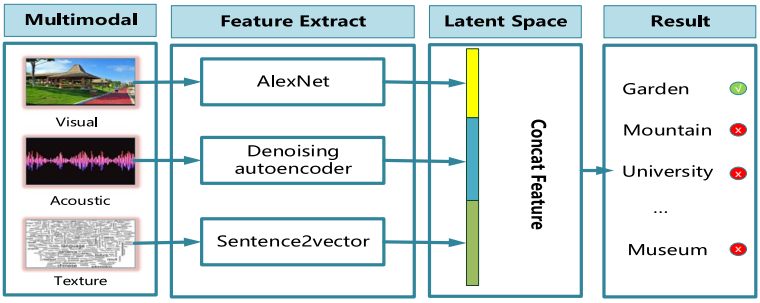

Fig. 2 Micro-video classification framework based on tandem feature learning

2 Micro-video classification and prediction

Micro-video analysis has important research and commercial value. Thus far, research has been conducted on micro-video classification and prediction, including venue recognition, tag prediction, popularity prediction, sentiment analysis, and circularity estimation.

2.1 Venue recognition

The venue category information of micro-video is an important clue for social network applications such as location application and personalized service, and the venue recognition task is to judge the venue information of micro-video. as shown in Fig. 3. While traditional videos usually have venues from multiple locations, micro-videos usually only record content from a specific location due to their short duration, usually only 6-15 seconds. In addition, micro-video is generated by mobile phone users and has the characteristics of randomness, subjectivity and diversity, so micro-video have problems such as low quality and information loss. According to the statistics of some researchers on micro-videos, only 1.22% of micro-videos are related to venue information, which greatly hinders location applications and personalized services.

In response to the above challenges, some venue recognition work has been carried out. Nie et al. [47] constructed a dataset and proposed several venue classification methods based on multimodal learning. Zhang et al. [7] initially proposed micro-video venue

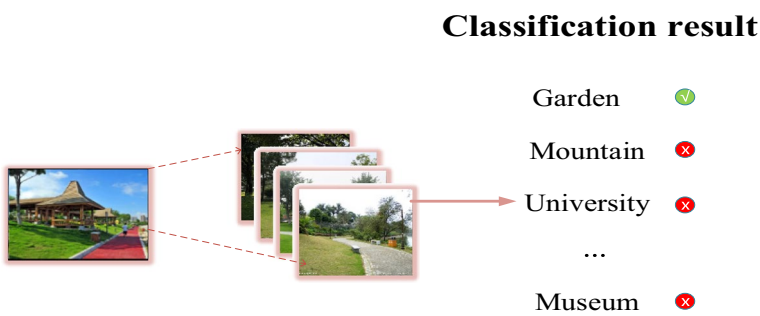

Fig. 3 Venue recognition task structure

recognition and a method of novel tree-guided multi-task multimodal learning method is presented, to label bite-sized video clips with venue categories. Nie et al. [8] proposed a deep transfer model that can deliver external audio information to enhance micro-videos with lower quality sound modalities. Liu et al. [21] proposed an end-to-end deep learning framework based on sparse constraint method. In order to effectively strengthen the generalization ability of the model and classify micro-videos into tree-like structures, the same author [23] developed a bran-new structure-guided multi-modal dictionary learning framework and an online learning algorithm for micro-video organization. Chen et al. [9] leveraged the check-in record of users in location-based social networks to overcome the limitations of using solely visual signals. They presented a universal framework based on matrix decompose to acquire the reciprocal action between the visual substance and temporal modality, thereby improving the accuracy of site-specific predictions of social images. Liu et al. [58] proposed a studying model that unitedly learns LSTM-CNNs with a archetype for micro-video scene recognition. Later, the authors [10] proposed an modified neural network structure known as NNeXtVLAD and combined NNeXtVLAD, CNNs, and context gating to produce a bran-new end-to-end united studying framework for application. Moreover, they [40] formulated micro-video scene recognition as a multimodal serial modeling issue and proposed a multimodality sequence model with gated completely convolutional blocks. Wei et al. [19, 42] presented a nervous multimodal coadjutant studying framework to divide the consistency and complementarity using a innovative Relationship between perceived attention mechanism. Guo et al. [57] outlined the challenges of micro-video scene recognition and proposed a new method based on a multilayer nervous network. On this basis, they [49] developed a combined multilayer neural network and supervised hash learning method. Moreover, a deep multimodal fusion network for scene recognition was proposed [34]. The authors [33] also presented attention-based consistent semantic learning. This method combines attention mechanism and double branching structure to solve the problem of intra-class semantic inconsistency.

Compared to traditional video venue recognition, most of the above studies are based on the characteristics of micro-videos to learn venue recognition methods, including problems such as noise, sparse concepts, inconsistent semantic strength between modalities, poor compactness within the same venue class, and missing data. However, for these problems, the relatively simple methods such as mean filling and strong modalities' semantic enhancement to weak modalities are basically adopted, without considering the distribution of modal noise, uncertainty of data, and adaptive auxiliary enhancement of strong and weak modalities, which cannot fully represent the original Semantic information.

2.2 Popularity prediction

Popularity prediction refers to the prediction of people's affection for a certain video. as shown in Fig. 4. It has great commercial potential in many aspects, such as network marketing and brand tracking. The popularity prediction of traditional video has obtained a good theoretical basis, but there are still great challenges for the popularity prediction of micro-video. The production purpose of micro-videos is to spread and share quickly among users. Compared with traditional videos, micro-videos have more internal connections with social networks in addition to the time, noise, poor video quality and other features. Therefore, it is an important task to predict the popularity of micro-video. The number of clicks is an indicator to predict the popularity of micro-videos, and the number of shared micro-videos can also indicate its spread in the social field. The playing time of micro-videos can

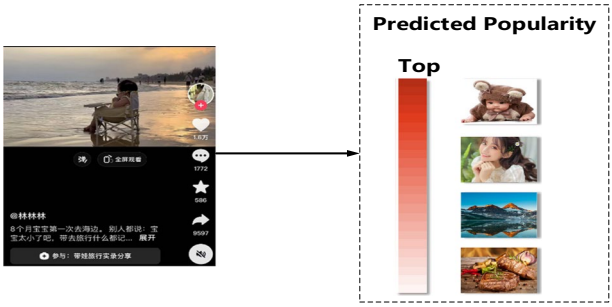

Fig. 4 Popularity prediction task structure

measure users' participation in the video content, and can also serve as a supplement to the number of clicks. Under the influence of external factors, the popularity of micro-videos is often not fixed, so the task of popularity prediction is full of challenges.

In view of the above challenges, some researchers put forward relevant solutions. Chen et al. [5, 6] proposed a novel transductive multimodal learning strategy to predict accurately the popularity of micro-videos. In order to reduce the interconnection, heterogeneity and noise problems in micro-video assessment of popularity and make micro-video popularity prediction more accurate, Jing et al. [22] proposed a low-rank multi-view implant structure. Ding et al. [16] demonstrated the use of publicer-related and video description-related functions to build a model to identify microvideos with large deviations between clicks and shares. Chen et al. [17] developed a micro-video strategy of click-through prediction based on deep network by modeling the historical behaviors of users. Ma et al. [54] proposed a coarse-to-fine method for the joint learning of the click-through and playtime of micro-videos. Su et al. [39] presented a feature-recognition transductive framework, the micro-video is divided into unlike popularity standard by attribute features, and the popularity rating are predicted by accurate underlying features. Xie et al. [45, 46] proposed a multimodal shifty encoder-decoder framework for the task of micro-video popularity prediction. Han et al. [26] developed an emotion capsule model for micro-video click-through rate forecast base on positive and negative feedback.

The traditional popularity prediction method may only use the machine learning method of support vector regression, and can not make full use of the relationship between different modalities. However, in the task of popularity prediction, in addition to visual, acoustic and textual features, social features of micro-video can not be ignored. Therefore, some studies consider extracting heterogeneous information among features from different views, comprehensively considering the intrinsic semantic structure of heterogeneous features, so as to complement each other and better capture the semantic information of features.

2.3 Tag prediction

Tag prediction refers to the prediction of the categories of micro-videos. as shown in Fig. 5. Compared with traditional video, micro-video is usually shot by handheld mobile devices, and contains a third-person view, egocentric view and self-facing view, video frames as shown in Fig. 6. Moreover, micro-video will be continuously released on network platforms, reflecting the temporal evolution of label themes. These factors make micro-video become an important data for the model of video understanding. As an effective metadata of multimedia content management, tags can be used to index multimedia content and

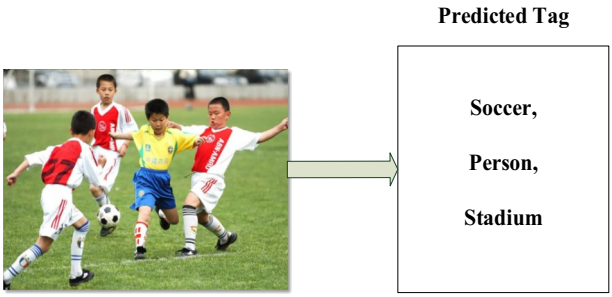

Fig. 5 Tag prediction task structure

greatly improve the search performance. However, statistics show that most users are not used to providing tags when uploading micro-videos, which results in unbalanced distribution of labels and low quality of micro-videos. Therefore, the tag prediction of micro-videos has become an important research topic.

The accuracy of hashtag prediction is important for micro-video research. Nguyen et al. [2] introduced a real-world dataset of micro-videos and developed viewpoint-specific time-evolving video understanding models, which were defined by state-of-the-art motion and profound visual features. Huang et al. [3] proposed a label refinement method to deal with unbalanced and low-quality micro-video labels. Su et al. [41] presented a low-rank regularized deep collaborative matrix factorization strategy and subsequently proposed a depth matrix decomposition algorithm with latent correlation assessment to tackle micro-video multilabel classification tasks more effectively. Yang et al. [35] presented the senTiment enhanced multi-mOdal Attentive haShtag recommendaTion (TOAST) model to perform micro-video lable recommendations and creatively aggregated the sentiment and content features for enhanced performance He et al. [28] proposed a novel solution known as TagPick to bridge the categories and micro-video hashtags.

Tag information is usually defined manually, which is easy to lead to poor compatibility with training data. The deep structure of low-dimensional representation is encoded in the feature space, ignoring the complementarity of label representation. Therefore, it is important to consider the low-rank constraints of each representation to ensure that the lowest order inherent representation of the micro-video is learned. In addition, a tag is not only a

Fig. 6 Third-person, egocentric, and self-facing views video frame

tag, but also has semantic information. Therefore, in the task of tag prediction, it is of great significance to effectively mine and utilize the semantic information of the tag.

2.4 Other studies

Studies on creativity analysis, circularity evaluation, event detection, and sentiment analysis also hold great research significance for micro-video analysis. Redi et al. [1] created a novel dataset of innovative micro-videos and a set of features is proposed to simulate the novelty and aesthetic value of micro-video. Sano et al. [4] presented a cyclicity evaluation strategy to classify cyclic/non-cyclic videos on the basis of exploring the space and time statistics of various visual features. In order to deal with the issue of micro-video event detection, Zhang et al. [36] proposed a multi-modal low-rank representation algorithm with label relaxation, which used the complementarity between modalities to obtain the potential public representation of micro-video. Liu et al. [20] developed a technology for exploring the emotion of the comment for online micro-videos. This algorithm contrasts positive and passive words that are displayed in comments. The proportion and degree value of adverbs of degree in positive and passive assessment is used to calculate the affective value. Gu et al. [37] proposed an unsupervised sentiment keyframe extraction model based on low-rank and sparse representation, in which the problem is transformed into an optimization problem for low-rank sparse representation.

3 Micro-video recommendation

Micro-video recommendation algorithms also offer important research and application value. Recommendation methods for micro-videos can be designed according to factors such as the user interest, attention mechanism, multimodal fusion, and user interaction with micro-videos.

With the rapid rise of micro-video platforms, a large number of videos appear on various micro-video platforms, making it difficult for users to find the videos they are interested in. Meanwhile, it is also difficult for video publishers to continuously attract users to watch their videos. Therefore, the recommendation algorithm of micro-video becomes an important research topic. The micro-video recommendation task is designed to recommend videos to users that they like. as shown in Fig. 7. The traditional recommendation algorithm can reflect whether the user interacts with the video, and there are only two values of 0 and 1, which cannot reflect the user's interaction intention and interest preference. Compared with traditional video, micro-video has short duration, noise, missing data and other problems. If only the interaction between users and micro-video is considered, its effect will be worse than traditional video. Therefore, it is very important to analyze and process micro-video separately.

Micro-video sharing platform is playing an increasingly important role. This is due to the recommendation system in reducing the user search time and reduce information overload powerful ability. In the pursuit of high quality micro-video recommendations, it is crucial to discover users' interests and provide micro-videos tailored to their tastes. Jin et al. [52] proposed the multimodal interest development method to solve problems such as cold-start and mismatch in video recommendation. This method extracts user interests based on original video and audio content. Considering the category problem of micro videos,

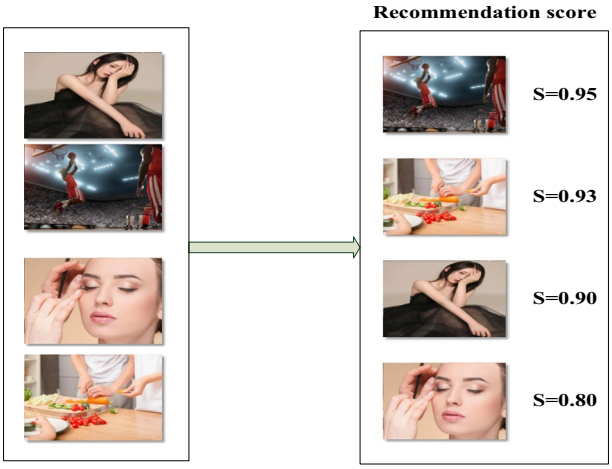

Fig. 7 Micro-video Recommendation task structure

Chen et al. [55] proposed two implicit scoring methods according to the interest preferences of the classes. Jiang et al. [44] developed a multiscale time-aware user interest modeling framework to learn user pleasure from fine-grained interest groups. These approaches have achieved great success, but modeling only historical behavior reduces the ability of user modeling to capture diverse and dynamic user interests. Because user interests are dynamic, new items cannot be well recommended based on history alone, thus reducing the diversity of recommendations. To this end, Lu et al. [32] proposed the explicit dynamic modeling of multiple trends in the current user preferences and made predictions based on historical and future potential trends. The key problem of micro-video recommendation is to establish user interest model and the relationship between user interest and micro-video content. At present, most methods only focus on the popularity of micro-video, ignoring the diversity of user interest. In order to solve this problem, Huang et al. [15] presented the layered modeling of user interests based on multimodal features. Yi et al. [31] proposed a cross-modal variational autoencoder, which is a hierarchical Bayesian generative model for content-based background music recommendation of micro-videos. The specific implementation method is to match the related background music with the micro-video by projecting the two multimodal inputs into the shared low-dimensional latent space. To strengthen and identify user preferences in the domain of microvideo, Lei et al. [30] proposed a sequential multimodal message transmission network. Subsequently, they established a cross-field contrastive learning pretraining strategy to learn the sequence behavior encoders for bridging the gap between the two domains.

It is difficult to build a personalized recommendation system. Firstly, users' interests will change with time. For example, they like swimming in summer and skiing in winter. Secondly, users may have different ways of interaction on micro-videos, including liking, commenting and following. "Likes" and "comments" indicate that the user is attracted to the work, and "likes" indicate that the user wants to continue seeing the author's work in the future. How to integrate different levels of interest into personalized recommendations becomes a big challenge.

The use of information exchange between users and projects to strengthen user representation is another method for micro-video recommendation. Wei et al. [11, 12] presented

a multi-modal graph convolution network framework, which is based on the conceptual design of information transmit of graph neural networks, and uses the multi-modal information exchange between users and micro-videos to generate fine-grained preference information of users for micro-videos. Li et al. [13] proposed an LSTM method based on time graph. In this method, the diachronic interaction list is encoded as time graph and a multi-level interest modeling layer is constructed to reinforce the user preference expression. Li et al. [50] presented a graph-based multiview representation interactive embedded model for micro-video hashtag recommendation based on multiview learning, tag correlation, and video user tag information. This model comprehensively considers sequence feature learning, video user tag interaction and tag correlation, and improves the performance of micro-video tag recommendation. Cao et al. [38] proposed muLti-mOdal-based hashtaG recOmmendation to recommend micro-video tags using multiple modalities. Two key questions are effectively solved under this framework: sequential frame modeling and multimodal fusion. Liu et al. [29] presented a conceptual perception denoising graph neural network, which connects micro-video, user, and concept nodes to form a tripartite heterogeneous graph. User and object representation for micro-video recommendation can be obtained through the graph neural network.

Although micro-videos contain such interactive methods as “like” and “comment”, statistics show that the vast majority of users just slide to the next video, with less feedback on micro-videos, which cannot be a good judge of whether users like this video. In addition, the average life cycle of micro-video is relatively short. Generally, two days after the release of micro-video, the number of user interaction drops sharply. These problems lead to sparse user interaction and poor performance of micro-video recommendation.

To solve these problems, Ma et al. [18] presented a deep neural network recommendation model with user items, contextual content, and visual content as inputs to solve the micro-video recommendation problem. These authors [48, 51] later proposed a method for the joint modeling of multisource content project information, multitype customer networks, and latent item classes. Liu et al. [53] developed a new transformer that alleviates the tendency of the RNN to compress all history into a fixed hidden representation. Chen et al. [25] used deep network learning to integrate multiple user preference information. The user interest was represented through potential, item-level, neighbor-assisted, and category-level representations. Liu et al. [56] presented a user video attention network framework, which uses an attention pattern to study the multimodal message from customers and micro-videos. A user-guided layered multicipital attention structure [43] was also proposed by them, which takes into account both micro-video content and user interests. Yao et al. [27] proposed a self-over co-attention model to strengthen the user preference expression. Shang et al. [14] improved the traditional recommendation algorithm and used the parallel computing algorithm of MapReduce and the Hadoop framework to process big data.

Compared with traditional videos, micro-videos are shorter but more focused on certain interesting topics. Micro-video interaction sequences are more closely related and have more stable preference patterns. Through the joint analysis of visual content and semantics, more specific types of micro-video can be found technically. Traditional recommendation methods are mostly based on Collaborative Filtering (CF), which collects the historical behaviors of multiple users and learns the cooperative similarity between users or items to recommend. The user-item interaction matrix is extremely sparse and does not depend on content analysis. Therefore, there are many problems such as sparse data and cold start. Some studies consider using heterogeneous information for analysis, integrating potential representations of multiple information sources into a unified model to facilitate personalized recommendations.

4 Feature representation

The scene or visual concept conveyed by visual features is the visual signal of tasks of micro-videos. It is particularly important to extract high-level semantics from visual modalities to represent micro-videos[66–70]. For example, when we observe "table", "chair", "coffee cup" and so on from a micro-video, we can easily predict that the micro-video is shot in a coffee shop, which prompts us to extract rich features from the visual modalities to represent the micro-video. Convolutional neural network (CNN) has been used as a powerful model to extract the representation of visual features. We summarized the existing methods of feature extraction for micro-video tasks, most of which used deep convolutional neural networks to extract visual features. The main models include: method based on AlexNet and LSTM, method based on VGG16 and Attention, method based on ResNet and Visual Transformer, etc.

4.1 Based on AlexNet and LSTM combined method

AlexNet[60] was proposed and achieved excellent results in ImageNet Large-scale Visual Recognition Challenge in 2012, which promoted the development of convolutional neural networks. AlexNet is composed of 5 convolution layers + maximum pooling layer and 3 fully connected layers. The first 5 convolution layers + maximum pooling layer are used for feature extraction, and the last 3 fully connected layers are used as classifiers. It has 60 million parameters and 650,000 neurons. To make training faster, it used unsaturated neurons and GPU implementation. The network structure is shown in Fig. 8.

AlexNet contains Conv convolution layer, which main function is to extract the features of input images. The LRN layer (local response normalization) is proposed to normalize the data between 0 and 1, create a competition mechanism for the activity of local neurons, make the values with large responses become relatively larger, inhibit other neurons with small feedback, and enhance the generalization ability of the model. Using ReLU activation function, the gradient dispersion of Sigmoid in deep network is solved successfully. MaxPool is the maximum pooling layer. Its main function is to downsample features, reduce the size of feature map, and retain the main information of features. AlexNet uses the maximum pooling layer to avoid the blurring effect of average pooling and improve the richness of features.

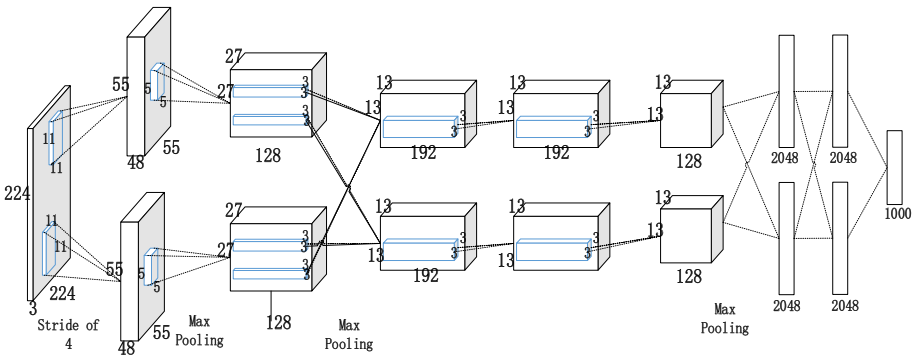

Fig. 8 AlexNet network structure

The AlexNet model is pre-trained on 1.2 million clean images on ILSVRC12 and can provide robust initialization for recognizing semantics. Before feature extraction, key frames are extracted for each micro-video, and the average pooling strategy is adopted for all key frames of the video. Then AlexNet is used to extract the visual features of each frame, which can better extract the high-level semantic information in the visual modalities and capture the relationship information between video frames, which is conducive to the classification of micro-videos.

Some researchers use AlexNet to extract visual features. On this basis, feature sequences can be independently captured by combining with Long Short Term Memory (LSTM) [61]. Using LSTM to extract frame-level features, the time structure of a video can be captured into a single representation. Its network structure is shown in Fig. 9. LSTM is a Recurrent Neural Network (RNN) designed to solve the problem of gradient disappearance or gradient explosion caused by long-term dependence on recurrent neural networks.

4.2 Based on the combination of VGG16 and attention

VGG network [62] was proposed in 2014 and has become a popular convolutional neural network model due to its relatively simple network structure and excellent performance in network training. VGG16 contains 13 convolutional layers, 3 fully connected layers, 5 pooled layers, and a Softmax layer. The network structure is shown in Fig. 10.

VGG16_places365 is a pre-trained VGG16 network for image recognition in the common dataset Places365 for image scene recognition. VGG16 network is composed of small convolution kernel, small pool kernel and ReLU, with relatively simple structure. By increasing depth, the performance can be effectively improved. Convolution can replace full connection and adapt to pictures of various sizes. VGG16_places365 network is used to extract original visual features, which simplifies the structure of convolutional neural network, improves the fitting ability of training, and can better retain high-level semantic information. In the process of feature extraction, VGG16 uses convolution kernel with size of 3*3, which makes the training results better. In the process of convolution operation, the value in the middle position is extracted many times, while the feature extraction of the boundary value is relatively less. Padding is added to make better use of the boundary value and facilitate calculation. The feature information extracted after the convolution operation may have information redundancy. The pool layer is used to continuously reduce the space size of the data, so that the number of parameters and the amount of calculation are constantly reduced, and overfitting can be controlled to a certain extent.

Some researchers use VGG16_places365 to extract visual features. On this basis, more effective feature information can be focused by combining the self-attention mechanism method, and its network structure is shown in Fig. 11. Self-attention mechanism a resource

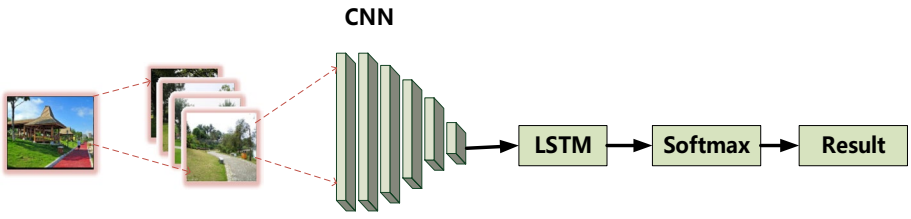

Fig. 9 CNN + LSTM network structure

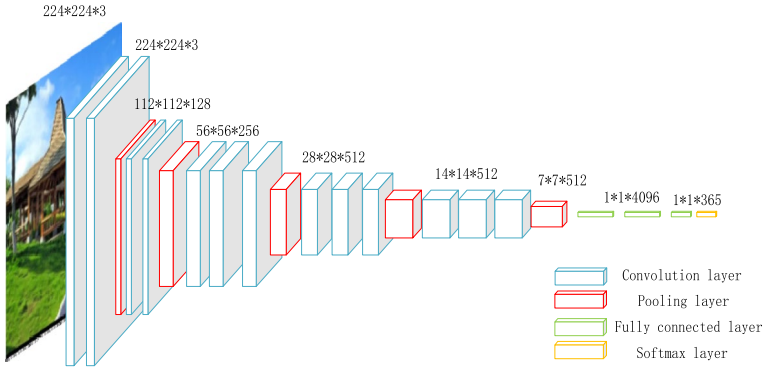

Fig. 10 VGG16 network structure

allocation scheme that allocates computing resources to more important tasks while solving the problem of information overload in the case of limited computing power. In the feature extraction stage, global images are scanned to obtain the target region that needs to be focused on, and then more attention is paid to this region to obtain more valuable details, so as to extract richer semantic information.

4.3 Based on ResNet and visual transformer combined method

In the process of micro-video visual feature extraction, the researchers found that with the increase of the number of layers, the performance of the neural network becomes lower and lower, showing the degradation problem. In order to solve this problem, ResNet was proposed in 2015 ImageNet competition and achieved excellent results [63]. ResNet designed a residual structure using skip connection, which enabled the network to reach a deep level and improved the performance. The residual structure is shown in Fig. 12. Taking ResNet50 as an example, it contains 49 convolution layers and one fully connected layer. The network structure can be divided into seven parts. The first part mainly carries out the calculation of input convolution, regularization, activation function and maximum pooling; the second, third, fourth and fifth parts contain residual blocks, each of which contains three layers of convolution. The input of the network is $224 \times 224 \times 3$, and after the convolution calculation of the first five parts, the output is $7 \times 7 \times 2048$, which will be converted into a feature vector by the pooling layer. Finally, the classifier will calculate and output this feature vector.

ResNet is widely used in many fields. In the task of micro-video classification, ResNet network can capture prominent visual features of micro-video, so that the

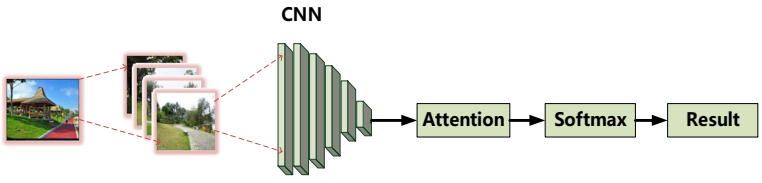

Fig. 11 CNN + Attention network structure

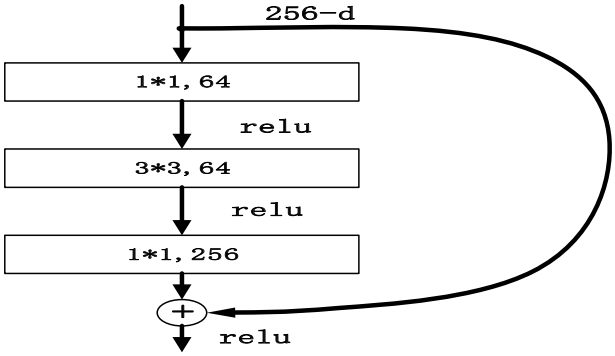

Fig. 12 ResNet residual structure

visual features can convey rich semantic information, and the classification accuracy can be improved by constantly increasing the depth of network. Thus, the micro-video classification method can achieve higher performance.

Some researchers used ResNet to extract visual features. On this basis, Wang et al. [64] proposed Resnet-DT, a domain adaptive network, to effectively improve the quality of features. In order to focus the scene information of micro-videos more effectively, they [65] proposed VT-ResNet to extract visual features and achieved good results. Its network structure is shown in Fig. 13.

5 Experiments in Micro-video analysis

5.1 Datasets

Datasets are highly significant in micro-video analysis research. Based on all references that were studied for this survey, several popular datasets are listed in Table 1. As indicated in the table, Vine and MicroVideo-1.7 m are frequently used datasets. Vine is a micro-video sharing website. The videos in this dataset include three modalities: visual, audio, and text.

Vine [7] This dataset consists of real data collected from the well-known Vine platform with micro-videos captured from Vine through its public API. The dataset only includes three modalities, location information, and exactly 6 s of micro-videos. Furthermore,

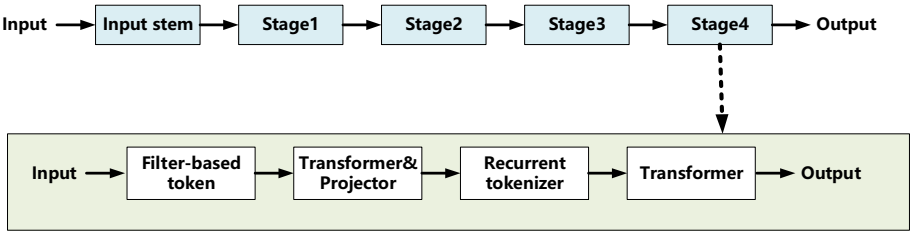

Fig. 13 VT-ResNet network structure

micro-videos with less than 100 tags were eliminated in the venue categories, resulting in 20,093 micro-videos collected from over 22 Foursquare revenue classes.

Microvideo-1.7 m [17] This dataset is based on real data collected from the well-know Chinese micro-video sharing platform. In this dataset, 10,986 users perform 12,737,619 interactions with 1,704,880 micro-videos. Inception V3 model is used to extract features from the cover image of each micro-video. The dataset contains 512 categories, with only one category per microvideo. Each interaction has an associated user ID and microvideo ID, as well as a timestamp.

TikTok [52] TikTok is a well-known micro-video-sharing app that permits customers to shoot and publish their favorite works. It consists of micro-videos users, and their interactions, such as clicks. This dataset includes 6293 users, 23,343 micro-videos, and 129,423 user behavior logs. Users with fewer than 10 records were deleted to avoid data sparsity.

MLSV2018 [24] This dataset is published by Meitu and contains a large dataset of multi-label micro-videos. MLSV2018 contains micro-videos in 63 hot categories with popular elements such as "diving," "hairdressing," and "painting." Each micro-video is 5 to 15 s in length and has 1 to 3 labels.

KuaiShou [13] This dataset is used to infer the possibility that a customer clicks on a new micro-video. Various actions of users and micro-videos are included in this dataset, The organizers released a visual embed of the 2048-day thumbnail for each micro-video. A total of 10,000 customers and their 3,239,534 interactive micro-videos were randomly selected from the large-scale dataset to construct the dataset.

Xigua [45] This dataset consists of real data that were collected from Xigua Video. In the crawling process, a partial quantity of active customers from Xigua Video were first tracked as seed

Table 1 Use frequency and application of datasets

Dataset	Frequency	Application
Vine	15	Venue Classification Hashtag Recommendation Popularity Prediciton
MicroVideo-1.7 m	5	Click-through Prediction Recommendation
TikTok	3	Click-through Prediction Recommendation
MLSV2018	2	Popularity Prediction
KuaiShou	3	Click-through Prediction Recommendation
Xigua	2	Popularity Prediction
Toffee	2	Recommendation
Twitter	1	Recommendation
YouTube Emotion Dataset	1	Emotion Prediction
Musical.ly	1	Hashtag Recommendation
MicrovideoSceneData_10	2	Scene Classification

users. The list of videos posted by every seed user every 6 min was scanned, and any new release was added to the tracking list. The view increment value of a video was recorded every 15 min. Eventually, 3,231,072 records were collected from 11,219 micro-videos posted by 2664 users.

Toffee [56] This dataset is a popular micro-video social platform in China. A user-centric approach was used to construct the dataset. First, a week of public timeline micro-videos and a random sample of active users were obtained. An active user is a user who has interacted with at least five micro-videos. Second, the seed audience was expanded by crawling their followers. A three-layer crawl was used to crawl the user ID and registration information, the user favorite micro-video ID list, and the content information of these micro-videos. This dataset contains 4011 users, 123,589 micro-videos, and 569,600 interactions.

Twitter [18] This dataset is the real data collected for micro-video recommendation. Firstly, Twitter Streaming APIs6 was used to obtain data for a year from 2015 to 2016. Subsequently, users who published less than five micro-videos were filtered out to alleviate data sparsity. Thereafter, a web crawler was used to collect all micro-videos in testing dataset. Finally, a dataset with 51,837 users and 147,378 micro-videos was constructed.

YouTube emotion dataset [59] The dataset consists of 1101 user videos, mainly from YouTube and Flickr. The dataset includes 844 micro-videos with a length within 200 s. They range in duration from 4 to 4500 frames, with an average of about 1583 frames and a standard deviation of 927 frames. These videos have rigorous logic and wonderful content, and there are no more Settings for video structure.

Musical.ly [43] This dataset was built by collecting data from Musical.ly. In the collected data, we removed several low-frequency tags that recorded fewer than 10 interactions with the microvideo. The final dataset contains 10,291 microvideos uploaded by about 6500 anonymous users. The dataset contains 669 unique tags, with an average of 3.42 tags per microvideo. There are three types of metadata for de-identified users: gender, age, and country. The user's history TAB is also included.

MicrovideoSceneData_10 [33] The new microvideo scene dataset, created by recombining the original dataset, was obtained from Vine and is publicly accessible on the website.[1] This dataset was created for micro-video scene classification [33, 34]. The new data set, with samples ranging from 100 to 2000, can be grouped into 10 scenario categories. Each video sample was taken by a moving camera and averaged about 6 seconds. In order to maintain the real distribution of micro-video, the new data set is not balanced within classes and has different characteristics in many classes. We preprocess the dataset to remove noisy data that is not helpful for scene semantic analysis.

5.2 Evaluation metrics

The area under the curve (AUC), Precision@K, Recall@K, F1-measure@K, and NDCG@K are used as evaluation metrics for micro-video recommendation. The AUC is

[1] www.acmmm16.wixsite.com/mm16

the area under the ROC curve, which is derived by plotting the relationship between the real and false optimistic rates. Precision@K is the percentage of actual user interest project in the top-k project in the recommended list, Recall@K is the recall data of the top-k project, and F1-measure@K is the average symphonious data of the precision and recall rate of the top-k items. NDCG is the Normalized Discounted Cumulative Gain and NDCG@K considers the order of recommended project in the top-k list.

The macro-F1,micro-F1 and mAP are used to calculate the micro-video classification performance. Macro-F1 assigns an same weight to each class label, and micro-F1 assigns an same weight to all examples. mAP is the percentage of all samples that predict the correct sample. The performance of the micro-video popularity prediction is determined based on the standardized mean square error (nMSE) between the predicted and actual prevalence. nMSE is an estimate of the total deviation between the predicted and measured values.

5.3 Performance analysis

We analyze the methods of venue classification, popularity prediction, tag prediction and micro-video recommendation, and evaluate the performance of different methods by performance criteria.

The methodological performance of the venue recognition task is shown in Table 2. From the performance analysis of methods on different datasets, it can be seen that the classification performance of various methods in Vine dataset is generally low. Micro-F1

Table 2 Performance analysis of venue recognition method

Method	Dataset	Evaluation Metrics	performance
TRUMANN[7]	Vine	Micro-F1	25.27%
		Macro-F1	5.21%
Deep transfer model[8]	Vine	Micro-F1	31.21%
		Macro-F1	16.66%
NMCL[19]	Vine	Micro-F1	40.04%
		Macro-F1	26.78%
EASTERN[21]	Vine	Micro-F1	59.51%
		Macro-F1	30.57%
Jointly learning model[58]	Vine	Micro-F1	62.73%
		Macro-F1	32.93%
Multi-modality sequence model[40]	Vine	Micro-F1	63.23%
		Macro-F1	33.84%
NNeXtVLAD[10]	Vine	Micro-F1	**66.87%**
		Macro-F1	**41.88%**
Multi-layer neural network[57]	Vine	mAP(@50)	45.04%
		mAP(@100)	44.68%
Combinational fusion method[49]	Vine	mAP(@50)	46.9%
		mAP(@100)	47.7%
MESL[34]	MicrovideoSceneData_10	Accuracy	**98.26%**
ACSL[33]	MicrovideoSceneData_10	Accuracy	75.3%
INTIMATE[23]	Public dataset	Micro-F1	**6.60%**
		Accuracy	6.28%

Bold entries are the best performance of methods in same dataset

is no more than 70%, Macro-F1 is no more than 50%. Because the data in Vine dataset conforms to the real data distribution, there is a large amount of data and noise problem. Part of the data is incomplete, there is a problem of missing data, resulting in the semantic strength of each modality is not consistent; At the same time, the dataset has some problems such as poor intra-class compactness and unbalance of data categories. Therefore, the micro-video venue classification task is more challenging. As can be seen from the MicroVideoSceneData_10 dataset, the performance of the multi-modality fusion method is higher than that of the single-modality ACSL method. The coordination problem of consistency and complementarity exists among the multi-modalities of micro-video venues. The multi-modality information is used to learn the consistency and complementarity between different modalities, so as to achieve the balance of consistency and complementarity between different modalities and learn richer semantic representation. Therefore, micro-video venue recognition based on multi-modality fusion has a very large research space.

The method performance of popularity prediction task is shown in Table 3. The better performance of TMALL suggests that connecting different modalities via a uniform potential space is better than directly imposing inconsistent penalties on the original space. The excellent results of MMVED show that a deep learning-based approach is superior to a machine learning-based approach.

Table 4 shows the performance of the method of tag prediction task, and Table 5 shows the performance of other methods such as sentiment analysis and event detection. It can be seen that the performance of such classification tasks is generally low, and problems such as low data quality and noise in the data set have a great impact on the classification tasks, which poses great challenges.

Table 6 shows the method performance of micro-video recommendation tasks. The performance of vine dataset is less than 40%, and that of Kuaishou dataset is more than 70%, which is higher than other datasets. This proves the importance of sequential characteristics of user interests, and it is necessary to use dynamic interest modeling and historical sequence modeling of users. The performance of current micro-video recommendation methods is generally not high, and the diversity and variability of users' interests make this task a great challenge.

6 Challenges in Micro-video analysis

Compared with the traditional video, the duration of micro-video is shorter, usually only 6-15 seconds, and most of them are generated by mobile phone users, featuring randomness, subjectivity and diversity, which brings some challenges to the task of micro-video

Table 3 Performance analysis of popularity prediction method

Method	Dataset	Evaluation Metrics	performance
TMALL[5]	SentiBank dataset	nMSE	97.9%
TLRMVR[22]	Vine	nMSE	93.4%
Ding et al. [16]	Anonymous dataset	AUC	90.2%
THACIL[17]	MicroVideo-1.7 M	AUC	68.4%
coarse-to-fine model[54]	MicroVideo-1.7 M	AUC	68.5%
FDTM[39]	Vine	nMSE	92.9%
MMVED[45]	Xigua dataset	nMSE	97.5%
ASCap[26]	MicroVideo-1.7 M	AUC	72.5%

Table 4 Performance analysis of tag prediction method

Method	Dataset	Evaluation Metrics	performance
Nguyen et al. [2]	Real-world dataset	MAP	28%
Huang et al. [3]	Vine	MAP	61.7%
LRDCMF[41]	MLSV2018	MAP	81.35%
TOAST[35]	Vine	Recall	73.60%

analysis: 1) there is a lot of noise information. Many micro-videos have blurred image, noisy background, dim light, special effects and other noises. 2) The intra-class compactness of the same scene is poor. As micro-videos are shot by many people, each person has a different shooting style, which leads to different styles of the same type of micro-videos, resulting in problems such as poor compactness within the class. 3) Unbalanced data categories. In a period of time, there are more discussions on hot topics, so the corresponding micro-videos will increase, which leads to the inconsistency of the number of micro-videos in each category. 4) Coordination problem of consistency and complementarity among the multi-modalities of micro-video. It is also a challenge to learn the consistency and complementarity of multiple modalities and achieve a better balance between the consistency and complementarity of different modalities.

Although substantial research has been conducted on micro-video analysis and a lot of work has been done to address the above challenges, there are still some challenges to be addressed in the future: 1) Incomplete data. The missing data problem has been considered in many works, and some studies have solved this problem using mean value filling. However, the filled vectors do not include accurate semantic information. Therefore, the accurate filling in of missing information or learning of semantic information from other modalities is an important problem to be solved. 2) Inconsistent semantic strength. Most micro-video analysis works have learned feature representations effectively through multimodal fusion. However, the semantic strength of different modalities in the same sample is inconsistent owing to the influence of noise or user privacy protection. Thus, the exploitation of multimodal information is another important issue to be addressed.

7 Conclusions

This survey has reviewed the important research on micro-video analysis, including venue classification, micro-video recommendation, popularity prediction, and click-through prediction. We have also provided an summarize of the benchmark datasets, features, and evaluation metrics. As opposed to traditional video, many challenges arise in micro-video analysis, including noise, long-tailed distributions, incomplete data, inadequate intra-class

Table 5 Performance analysis of other method

Method	Dataset	Evaluation Metrics	performance
Redi et al. [1]	D-100 dataset	Accuracy	80%
DoL[4]	Vine	Accuracy	87.2%
Zhang et al. [36]	MED dataset	MAP	84.4%
SKFE[37]	YouTube8 dataset	Accuracy	34.52%

Table 6 Performance analysis of micro-video recommendation method

Method	Dataset	Evaluation Metrics	performance
MMIE[52]	Tik Tok	NDCG	26.25%
MTIN[44]	MicroVideo-1.7 M	AUC	72.9%
	KuaiShou	AUC	75.2%
DMR[32]	MicroVideo-1.7 M	AUC	73.1%
	KuaiShou	AUC	74.2%
Huang et al. [15]	Vine	HR	35.6%
Yi et al. [31]	MNIST	Accuracy	82.02%
SEMI[30]	Taobao dataset	AUC	73.96%
		HR	49.97%
MMGCN[11]	Tiktok	Recall	55.20%
		NDCG	34.23%
ALPINE[13]	Kuaishou	AUC	73.9%
V2HT[50]	INSVIDEO	Recall	69.48%
		NDCG	54.98%
LOGO[38]	Public dataset	Recall	68.09%
		NDCG	48.54%
Conde[29]	Amazon	AUC	70.64%
LGA[18]	Twitter	Accuracy	33.9%
MSN[51]	Vine	Recall	32.57%
Collaborative Transformer[53]	TikTok	HR	52.38%
MUIR[25]	MicroVideo-1.7 M	NDCG	70.25%
UVCAN[56]	Toffee	HR	34.48%
UHMAN[43]	Musical	Recall	45.17%
SCAA[27]	Kuaishou	AUC	71.2%

compactness, and inconsistent semantic strength. Although existing works have considered some of these issues, in-depth research is still required for incomplete data and inconsistent semantic strength.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China (62176141, 62176139, 61876098), Major Basic Research Project of Natural Science Foundation of Shandong Province (ZR2021ZD15), Taishan Scholar Project of Shandong Province (tsqn202103088), Shandong Provincial Natural Science Foundation for Distinguished Young Scholars (ZR2021JQ26), Natural Science Foundation of Shandong Province (ZR2021QF119, ZR2022MF272) and special funds for distinguished professors of Shandong Jianzhu University.

Declarations

Not applicable.

References

1. Redi M, O'Hare N, Schifanella R, Trevisiol M, Jaimes A (2014) 6 seconds of sound and vision: creativity in micro-videos. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 4272–4279

2. Nguyen PX, Rogez G, Fowlkes C, Ramanan D (2016) The open world of micro-videos. arXiv preprint arXiv:1603.09439
3. Huang L, Luo B (2017) Tag refinement of micro-videos by learning from multiple data sources[J]. Multimed Tools Appl 76(19):20341–20358
4. Sano S, Yamasaki T, Aizawa K (2014) Degree of loop assessment in micro-video. In: 2014 IEEE International Conference on Image Processing (ICIP). IEEE, pp 5182–5186
5. Chen J, Song X, Nie L, Wang X, Zhang H, Chua T-S (2016) Micro tells macro: predicting the popularity of micro-videos via a transductive model. In: 2016 ACM international conference on Multimedia (ACM MM). ACM, pp 898–907
6. Chen J (2016) Multi-modal learning: Study on a large-scale micro-video data collection. In: 2016 ACM international conference on Multimedia (ACM MM). ACM, pp 1454–1458
7. Zhang J, Nie L, Wang X, He X, Huang X, Chua T-S (2016) Shorter-is-better: Venue category estimation from micro-video. In: 2016 ACM international conference on Multimedia (ACM MM). ACM, pp 1415–1424
8. Nie L, Wang X, Zhang J, He X, Zhang H, Hong R, Tian Q (2017) Enhancing micro-video understanding by harnessing external sounds. In: 2017 ACM international conference on Multimedia (ACM MM). ACM, pp 1192–1200
9. Chen J, He X, Song X, Zhang H, Nie L, Chua T-S (2018) Venue prediction for social images by exploiting rich temporal patterns in LBSNs. In: 2018 International Conference on Multimedia Modeling (MMM). Springer, pp 327–339
10. Liu W, Huang X, Cao G, Zhang J, Song G, Yang L (2019) Joint learning of nnextvlad, cnn and context gating for micro-video venue classification[J]. IEEE Access 7:77091–77099
11. Wei Y, Wang X, Nie L, He X, Hong R, Chua T-S (2019) MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In: 2017 ACM international conference on Multimedia (ACM MM). ACM, pp 1437–1445
12. Wei Y, Cheng Z, Yu X, Zhao Z, Zhu L, Nie L (2019) Personalized hashtag recommendation for micro-videos. In: 2019 ACM international conference on Multimedia (ACM MM). ACM, pp 1446–1454
13. Li Y, Liu M, Yin J, Cui C, Xu X-S, Nie L (2019) Routing micro-videos via a temporal graph-guided recommendation system. In: 2019 ACM international conference on Multimedia (ACM MM). ACM, pp 1464–1472
14. Shang S, Shi M, Shang W, Hong Z (2016) A micro-video recommendation system based on big data. In: 2016 IEEE/ACIS International Conference on Computer and Information Science (ICIS). IEEE, pp 1–5
15. Huang L, Luo B (2017) Personalized micro-video recommendation via hierarchical user interest modeling. In: 2017 the Pacific Rim Conference on Multimedia (PCM). Springer, pp 564–574
16. Ding J, Li Y, Li Y, Jin D (2018) Click versus share: A feature-driven study of micro-video popularity and virality in social media. In: 2018 SIAM International Conference on Data Mining (SDM). SIAM, pp 198–206
17. Chen X, Dong L, Zha Z-J, Zhou W, Xiong Z, Li Y (2018) Temporal hierarchical attention at category- and item-level for micro-video click-through prediction. In: 2018 ACM international conference on Multimedia (ACM MM). ACM, pp 1146–1153
18. Ma J, Li G, Zhong M, Zhao X, Zhu L, Li X (2018) Lga: latent genre aware micro-video recommendation on social media[J]. Multimed Tools Appl 77(3):2991–3008
19. Wei Y, Wang X, Guan W, Nie L, Lin Z, Chen B (2019) Neural multimodal cooperative learning toward micro-video understanding[J]. IEEE Trans Image Process 29:1–14
20. Liu Z, Yang N, Cao S (2016) Sentiment-analysis of review text for micro-video. In: 2016 IEEE International Conference on Computer and Communications (ICCC). IEEE, pp 526–530
21. Liu M, Nie L, Wang M, Chen B (2017) Towards micro-video understanding by joint sequential-sparse modeling. In: 2017 ACM international conference on Multimedia (ACM MM). ACM, pp 970–978
22. Jing P, Yuting S, Liqiang Nie X, Bai JL, Wang M (2017) Low-rank multi-view embedding learning for micro-video popularity prediction[J]. IEEE Trans Knowl Data Eng 30(8):1519–1532
23. Liu M, Nie L, XiangWang QT, Chen B (2018) Online data organizer: micro-video categorization by structure-guided multimodal dictionary learning[J]. IEEE Trans Image Process 28(3):1235–1247
24. Yuting S, Junyu X, Hong D, Fan F, Zhang J, Jing P (2021) Deep low-rank matrix factorization with latent correlation estimation for micro-video multi-label classification[J]. Inf Sci 575:587–598
25. Chen X, Liu D, Xiong Z, Zha Z-J (2021) Learning and fusing multiple user interest representations for Micro-video and movie recommendations[J]. IEEE Trans Multimed 23:484–496
26. Han Y, Pan G, Gao W, Guandong X, Jian W (2021) Aspect-level sentiment capsule network for micro-video click-through rate prediction[J]. World Wide Web 24(4):1045–1064

27. Dong Y, Zhang S, Zhao Z, Fan W, Zhu J, He X, Fei W (2021) Modeling high-order interactions across multi-interests for micro-video recommendation (Student abstract). In: 2021 AAAI Conference on Artificial Intelligence (AAAI). AAAI, pp 15945–15946
28. He L, Wang D, Wang H, Chen H, Guandong X (2021) TagPick: A system for bridging micro-video hashtags and e-commerce categories. In: 2021 ACM International Conference on Information and Knowledge Management (CIKM). ACM, pp 4721–4724
29. Liu Y, Liu Q, Yu T, Wang C, Niu Y, Yang S, Li C (2021) Concept-aware denoising graph neural network for micro-video recommendation. In: 2021 ACM International Conference on Information and Knowledge Management (CIKM). ACM, pp 1099–1108
30. Lei C, Liu Y, Zhang L, Wang G, Tang H, Li H, Miao C (2021) SEMI: a sequential multi-modal information transfer network for E-commerce Micro-video recommendations. In Proceedings of ACM SIGKDD conference 2021:3161–3171
31. Yi J, Zhu Y, Xie J, Chen Z (2021) Cross-modal variational auto-encoder for content-based Micro-video background music recommendation [J]. IEEE Trans Multimed 25:515–528
32. Lu Y, Huang Y, Zhang S, Han W, Chen H, Zhao Z, Wu F (2021) Multi-trends enhanced dynamic micro-video recommendation. arXiv:2110.03902v1
33. Guo J, Nie X, Ma Y, Shaheed K, Ullah I, Yin Y (2021) Attention based consistent semantic learning for micro-video scene recognition [J]. Inf Sci 543:504–516
34. Guo J, Nie X, Yin Y (2020) Mutual complementarity: multi-modal enhancement semantic learning for micro-video scene recognition [J]. IEEE Access 8:29518–29524
35. Yang C, Wang X, Jiang B (2020) Sentiment enhanced multi-modal hashtag recommendation for Micro-videos[J]. IEEE Access 8:78252–78264
36. Zhang J, Yuting W, Liu J, Jing P, Yuting S (2020) Low-rank regularized multimodal representation for Micro-video event detection[J]. IEEE Access 8:87266–87274
37. Xiaowei G, Lu L, Qiu S, Zou Q, Yang Z (2020) Sentiment key frame extraction in user-generated micro-videos via low-rank and sparse representation[J]. Neurocomputing 410:441–453
38. Cao D, Miao L, Rong H, Qin Z (2020) Liqiang Nie: hashtag our stories: hashtag recommendation for micro-videos via harnessing multiple modalities. Knowl [J] Based Syst 203:106114
39. Yuting S, Yang Li X, Bai PJ (2020) Predicting the popularity of micro-videos via a feature-discrimination transductive model[J]. Multimed Syst 26(5):519–534
40. Liu W, Huang X, Cao G, Zhang J, Song G, Yang L (2020) Multi-modal sequence model with gated fully convolutional blocks for micro-video venue classification[J]. Multimed Tools Appl 79(9-10):6709–6726
41. Yuting S, Hong D, Li Y, Jing P (2020) Low-rank regularized deep collaborative matrix factorization for Micro-video multi-label classification[J]. IEEE Signal Process Lett 27:740–744
42. Wei Y, Wang X, Guan W, Nie L, Lin Z, Chen B (2020) Neural multimodal cooperative learning toward Micro-video understanding[J]. IEEE Trans Image Process 29:1–14
43. Liu S, Xie J, Zou C, Chen Z (2020) User conditional hashtag recommendation for Micro-videos. In Proceedings of IEEE International Conference on Multimedia and Expo. 1-6
44. Hao Jiang, Wenjie Wang, Yinwei Wei, Zan Gao, Yinglong Wang, Liqiang Nie (2020) What aspect do you like: multi-scale time-aware user interest modeling for Micro-video recommendation. In Proceedings of ACM Conference on Multimedia 3487-3495
45. Xie J, Zhu Y, Zhang Z, Peng J, Yi J, Hu Y, Liu H, Chen Z (2020) A multimodal variational encoder-decoder framework for micro-video popularity prediction. In: 2020 International World Wide Web Conferences (WWW). W3C, pp 2542–2548
46. Zhu Y, Xie J, Chen Z (2003) Predicting the popularity of micro-videos with multimodal variational encoder-decoder framework. arXiv:2003:12724v1
47. Nie L, Liu M, Song X (2019) Multimodal learning toward micro-video understanding [M], San Rafael
48. Ma J, Wen J, Zhong M, Chen W, Li X (2019) MMM: multi-source multi-net Micro-video recommendation with clustered hidden item representation learning[J]. Data Sci Eng 4(3):240–253
49. Guo J, Nie X, Jian M, Yin Y (2019) binary feature representation learning for scene retrieval in micro-video. Multimed Tools Appl 78(17):24539–24552
50. Li M, Gan T, Liu M, Cheng Z, Yin J, Nie L (2019) Long-tail Hashtag Recommendation for Micro-videos with Graph Convolutional Network. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 509-518
51. Ma J, Wen J, Zhong M, Chen W, Zhou X, Indulska J (2019) Multi-source Multi-net Micro-video Recommendation with Hidden Item Category Discovery. In Proceedings of the 24th International Conference on Database Systems for Advanced Applications, 384-400

52. Jin Y, Xu J, He X (2019) Personalized micro-video recommendation based on multi-modal features and user interest evolution. In: 2019 International Conference on Image and Graphics (ICIG). SPIE, pp 607–618
53. Liu S, Chen Z (2019) Sequential behavior modeling for next micro-video recommendation with collaborative transformer. In: 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, pp 460–465
54. Ma S, Zha Z-J Wu F (2019) Knowing user better: jointly predicting click-through and playtime for micro-video. In: 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, pp 472–477
55. Chen J, Peng J, Qi L, Chen G, Zhang W (2019) Implicit rating methods based on interest preferences of categories for micro-video recommendation. In: 2019 International Conference on Knowledge Science, Engineering and Management (KSEM). Springer, pp 371–381
56. Liu S, Chen Z, Liu H, Hu X (2019) User-video co-attention network for personalized micro-video recommendation. In: 2019 World Wide Web Conferences (WWW). W3C, pp 3020–3026
57. Guo J, Nie X, Cui C, Xi X, Ma Y, Yin Y (2018) Getting more from one attractive scene: venue retrieval in micro-videos. In: 2018 Pacific Rim Conference on Multimedia (PCM). Springer, pp 721–733
58. Liu W, Huang X, Cao G, Song G, Yang L (2018) Joint learning of LSTMs-CNN and prototype for micro-video venue classification. In: 2018 Pacific Rim Conference on Multimedia (PCM). Springer, pp 705–715
59. Jiang Y, Xu B, Xue X (2014) Predicting emotions in user-generated videos. In: 2014 AAAI Conference on Artificial Intelligence (AAAI). AAAI, pp 73–79
60. Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet classification with deep convolutional neural networks [J]. Adv Neural Inf Proces Syst 25(2):1097–1105
61. Graves A, Graves A (2012) Long short-term memory [J]. In: Supervised sequence labelling with recurrent neural networks, 4th edn. Springer-Verlag, Berlin Heidelberg, pp 37–45
62. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556
63. He K, Zhang X, Ren S, et al. (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on Computer Vision & Pattern Recognition (CVPR). IEEE, pp 770-778
64. Wang B, Huang X, Cao G et al (2022) Hybrid-attention and frame difference enhanced network for micro-video venue recognition [J]. J Intell Fuzzy Syst 43(3):3337–3353
65. Wang B, Huang X, Cao G et al (2022) Attention-enhanced and trusted multimodal learning for micro-video venue recognition [J]. Comput Electr Eng 102:108127
66. Jian M, Wang J, Yu H et al (2021) Visual saliency detection by integrating spatial position prior of object with background cues[J]. Expert Syst Appl 168:114219
67. Jian M, Wang J, Yu H et al (2021) Integrating object proposal with attention networks for video saliency detection[J]. Inf Sci 576:819–830
68. Lu X, Jian M, Wang X et al (2022) Visual saliency detection via combining center prior and U-net[J]. Multimedia Systems 28(5):1689–1698
69. Jian M, Zhang W, Yu H et al (2018) Saliency detection based on directional patches extraction and principal local color contrast[J]. J Vis Commun Image Represent 57:1–11
70. Wan W, Wang J, Zhang Y, Li J, Hui Y, Sun J (2022) A comprehensive survey on robust image watermarking. Neurocomputing 488:226–247