Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 1020044, 10 pages
https://doi.org/10.1155/2021/1020044

Research Article

Video Scene Information Detection Based on Entity Recognition

Hui Qian[1], **Mengxuan Dai**[1], **Yong Ma**[1], **Jiale Zhao**[1], **Qinghua Liu**[1], **Tao Tao**[1], **Shugang Yin**[2], **Haipeng Li**[3] and **Youcheng Zhang**[3]

[1]*School of Computer and Information Engineering, Jiangxi Normal University, Nanchang 330022, China*
[2]*Siji Network Security Technology (Beijing) Co., Ltd., Beijing 102209, China*
[3]*Nanjing Unary Information Technology Co., Ltd., Nanjing 210002, China*

Correspondence should be addressed to Yong Ma; may@jxnu.edu.cn

Received 19 June 2021; Revised 12 September 2021; Accepted 18 October 2021; Published 31 October 2021

Academic Editor: Chi-Hua Chen

Video situational information detection is widely used in the fields of video query, character anomaly detection, surveillance analysis, and so on. However, most of the existing researches pay much attention to the subject or video backgrounds, but little attention to the recognition of situational information. What is more, because there is no strong relation between the pixel information and the scene information of video data, it is difficult for computers to obtain corresponding high-level scene information through the low-level pixel information of video data. Video scene information detection is mainly to detect and analyze the multiple features in the video and mark the scenes in the video. It is aimed at automatically extracting video scene information from all kinds of original video data and realizing the recognition of scene information through "comprehensive consideration of pixel information and spatiotemporal continuity." In order to solve the problem of transforming pixel information into scene information, this paper proposes a video scene information detection method based on entity recognition. This model integrates the spatiotemporal relationship between the video subject and object on the basis of entity recognition, so as to realize the recognition of scene information by establishing mapping relation. The effectiveness and accuracy of the model are verified by simulation experiments with the TV series as experimental data. The accuracy of this model in the simulation experiment can reach more than 85%.

1. Introduction

With the development of computer network and multimedia technology, the way characters receive information has shifted from traditional words and pictures to video stream. Taking China as an example, in the first half of 2020, the number of online audiovisual users has reached 901 million, with a year-on-year growth of 4.87% (https://new.qq.com/omn/20201014/20201014A05GLY00.html), which also leads to a sharp increase in video data. With the development of 5G technology, video's share of worldwide mobile data traffic will climb from 60% in 2018 to 74% in 2024 (https://blogs.cisco.com/sp/mobile-vni-forecast-2017-2022-5g-emerges). In such an environment with a large amount of video data, understanding video content is an important step for the intelligent system to approach human's understanding ability. It also has a great application value in social services, national security, and industrial development. However, video data is characterized by nonstructuration, strong redundancy, high dimension, deep information hiding, and understanding difficulties. How to map the complex video information into the semantic space in line with human cognitive habits is a challenge for video information extraction.

In recent years, the extraction and analysis of video information has become an important research content in video processing, which is of great significance in video semantic extraction, video query, and other aspects. Character detection and background detection, which are similar to scene information detection, have been deeply studied and widely applied [1–9]. However, there are not many in-depth researches on video situational information. At present, most of the proposed model is targeted to the recognition of face [10], character [11, 12], or background content [14] of the video, by extracting key frames and recognizing

the character information or the scene information in the frames to realize the extraction of the relationship between characters [13–15] and video scene classification [16, 17]. Zheng and Yu [10] combined the squeeze-and-excitation network (SEN) and residual network (ResNet) to accurately detect the face information in each frame, extract the position of the target face, and then extract face features from adjacent frames through the RNFT model to predict the position of the target face in the next frame. Gong and Wang [16] extracted background audio signals from match shots and recognized the sound of cheering and hitting from the audio signals of each match shot. By combining background audio signals and shot image information, this method realizes a more accurate video classification. Ding and Yilmaz [14] used to analyze whether characters appear in the same video scene, so as to extract the relationship network of the characters in the video. Tran and Jung [15] counted the cooccurrence of characters in video images to extract their relationship. However, most of these methods only take the global character/scene features at the camera level into consideration, ignoring the local features with more information and the relations that exist among them.

Scene detection is also widely used in real life. For example, in the novel coronavirus epidemic which started from 2020, the mode of online meeting and online teaching has become more and more popular, and the video data of meeting and course have also increased. When we process these video data, we find that there is a kind of application condition, that is, in a video, we usually only pay attention to the state of a target person/object under a specific situation. For instance, if a student participates in two consecutive classes in the same classroom, and the surveillance camera in the classroom will shot a video of these two classes. And we would like to analyze the student's attendance in one of the classes to ensure whether he was late or left early or returned after leaving for a period of time. When using the video information processing model mentioned above to analyze it, we found the following problems:

(1) Without more information, it is difficult for the computer to directly judge whether the student is in a changed course or not

(2) The computer is able to recognize all the parts when the student was absent in the whole video, but the process of determining whether the absence occurred in the course we are concerned about usually needs to be done manually

Lei et al. [21] proposed the SSCD method. It realizes the recognition of changing objects in a fixed scene and judges the change of street scene. However, it can not solve the above problems. In the case of lens movement or a large number of personnel changes, the error rate of the model will increase greatly, and it is difficult to deal with the processing of human-centered video. Similarly, there is the method proposed by Santana et al. [22], which can realize the rapid recognition of moving objects from a fixed perspective and judge the scene changes based on the results. However, this method can only obtain the contour map of moving objects and still can not well solve the above problems. The method proposed by Huang and Liao [23] can realize the scene detection task from the perspective of motion, but it has certain requirements for the consistency of video. At the same time, the method compares frame by frame, which has high requirements for the performance of the machine and insufficient processing speed.

To solve the above problems, a video scene information detection model based on entity recognition is proposed in this paper. This model makes use of more information including global information at video level and partial information at entity level for more information to get more accurate results. Similar to this example, there are many application conditions, such as the situational judgment of meeting process and the abnormal judgment of security video, etc., but existing video processing models are not able to handle such application conditions well.

According to the spatiotemporal features of the video scene, this paper selects the state of the video object as the characteristic to help us analyze and understand the video scene, combines with the state feature of the video subject, and determines the scene feature of the video subject. In this paper, the innovations can be summarized as the following three points:

(1) This paper proposes a new situational information detection model, which can recognize the changes of video situational information with high efficiency

(2) This paper establishes situational features by combining the spatiotemporal continuity between the subject and the object in video content, which enables the model to recognize situational information without semantic information of the video object and achieves good results.

(3) The accuracy of the model proposed in this paper reaches 80%

In this paper, we will explain and verify the above research contents. Section 2 will briefly introduce the existing entity recognition models, such as Yolo, and some mature face recognition models, such as face recognition. At present stage, these models are the premise for the test in this study. In Section 3, we will introduce the models, including their establishment, mathematical basis, and partial content of the pseudocode. Section 4 will present our experimental results and summarize the failed parts, which are also what need to be further discussed in our subsequent research work. In Section 5, we will summarize the research content and briefly introduce the main research directions in the future.

2. Relevant Work

2.1. Yolo. Yolo is a new target detection method [18], which is characterized by rapid detection and high accuracy. Redmon regarded the target detection task as a regression problem of target region prediction and category prediction. In this method, a single neural network is used to directly

predict item boundary and category probability to achieve end-to-end item detection. Yolo is widely used in target detection [19], target tracking [20], and other applications. Zhang et al. [19] used the deep separable convolutional method to optimize the convolution layer of the tiny Yolo model and divided a complete convolution operation into deep convolution and point-by-point convolution, thus reducing the parameters of CNN and improving the operation speed. Mohammed et al. [20] combined the neural network, image-based tracking, and Yolo V3 to solve the problem of intelligent vehicle tracking.

In this paper, Yolo V4 can be used as the target detection network in the entity detection stage. On the basis of Yolo V3, Yolo V4 has made a lot of innovations. The innovation of the input end is mainly the improvement of the input end during training, including Mosaic data enhancement, CMBN, and SAT self-confrontation training. Backbone network combines all kinds of new ways, including CSPDarknet53, Mish activation function, and Dropblock. The neck target detection network often inserts some layers in the backbone and the final output layer, such as the SPP module in the Yolo V4 and FPN+PAN structure. The anchor frame mechanism of the output layer is the same as that of Yolo V3. The main improvement is in the loss function CIou-Loss during training, and the NMS screened by the prediction box is changed into DIOU-nms. Yolo V4 is a major update of the Yolo series, with average accuracy (AP) and frame per second (FPS) in the COCO dataset improved by 10% and 12%, respectively.

2.2. Face Recognition Algorithm. In the model proposed in this paper, it is also feasible to directly use the face recognition algorithm to replace the target detection network. This method will reduce the accuracy of the model to some extent, but meanwhile, the computing efficiency will be better than the complete target detection network. When only the face recognition algorithm is used for scene information detection, the target object will be replaced by face recognition results, which greatly reduces the computational load of the model.

Face recognition is a powerful, simple, and easy-to-use face recognition open-source project, equipped with integrated development documents and application cases, and compatible with the Raspberry Pi system. You can use Python and command line tools to extract, recognize, and manipulate faces. Face recognition is a deep learning model based on C++ open-source library dlib. The face dataset Labeled Faces in the Wild is used for testing with a 99.38% accuracy. But the recognition accuracy of children and Asian faces has yet to be improved.

SeetaFace2 is a face recognition project written in C++ that supports Windows, Linux, and ARM platforms and does not rely on third-party libraries. This project includes face recognition module FaceDetector, face key point locating module Face Landmarks, and face feature extraction and comparison module Facerecognizer. FaceDetector can achieve a recall rate of over 92% under the condition of 100 false detections on FDDB, it also supports 5-point and 81-point localization of face key points, and its 1-to-N module supports face recognition applications with a base of thousands of characters.

3. Model

3.1. Model Description. The steps of video scene information extraction are as follows: Firstly, the input video is analyzed and preprocessed to obtain the entity target in each frame of the video. The main purpose of this work is to lay a good foundation for the subsequent subject-object labeling and the establishment of spatiotemporal relationship. Secondly, according to the input subject picture, the entity targets are compared and labeled, and the remaining entity targets are labeled as the object. Then, the video subject labeling results are used as scene nodes to extract and analyze the spatiotemporal relationship between the objects and the subjects in the video, so as to judge whether the scene is continuous or not. Finally, the attributes of scene nodes, namely, the scene information of the subject, is determined in the continuous scene.

This paper mainly focuses on the following:

(1) How to establish the relationship between subjects and objects?

(2) How to judge the attributes of scene nodes?

The model in this paper completes the above research contents through three stages of information processing.

3.1.1. The First Stage: Establish the Spatiotemporal Relationship between the Subject and the Object. In this stage, we lock the current situational information by establishing the relationship between the subject and the object, which is also the situational feature introduced in the model. The spatiotemporal relationship is mainly based on the randomness of object selection. Under the same condition, the mathematical probability that a certain number of randomly selected entities in the initial image of the scene will simultaneously appear abnormal in this period of time and space is very small.

According to the Bayesian probability formula, let the subject be X, and the object set $Y = \{y_1, y_2 \cdots, y_n\}$, $y_1, y_2 \cdots, y_n$ are independent of each other and random; the anomaly probability of Y can be shown as P_y; and the probability of n object anomalies and subject anomalies at the same time can be shown as P.

$$P = P_y{}^n * P_{\text{subject anomaly}}, \tag{1}$$

can be concluded. As shown in Figure 1, when $P_y = 0.3$, the probability of misrecognition is less than 5% when the value of n is greater than 3.

The spatiotemporal relationship is also reflected in the spatiotemporal continuity of the object. In the same scene, the mathematical probability of continuous abnormal occurrence of an entity randomly selected in the initial image of the scene in this period of time and space is also very small.

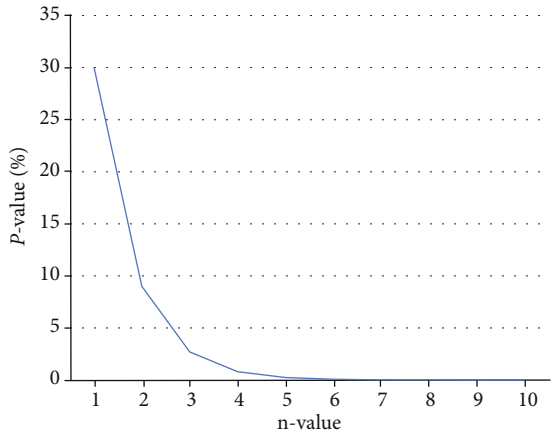

Figure 1: Influence curve of n value on model accuracy when $P_{\text{object}} = 0.3$.

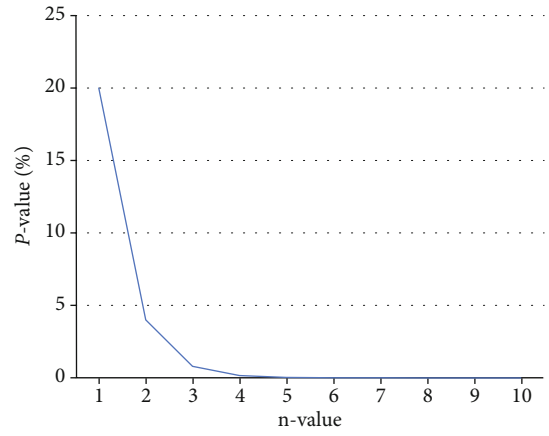

Figure 2: Influence curve of n value on model accuracy when P_{object}=0.2.

Since the occurrence of an anomaly in the same entity is an independent event, according to the principle of event independence, the object anomaly probability is assumed to be $P(y)$, and the probability P_{object} of object continuous abnormal n times.

$$P_{\text{object}} = P(y)^n, \tag{2}$$

As shown in Figure 2, in the case of $P_{\text{object}} = 0.2$, when n value is 2, the probability of misrecognition is 4%.

3.1.2. The Second Stage: Recognize whether the Subject and the object Are Abnormal. After establishing the subject-object relationship in the previous stage, we can realize the marking on the clips of the same scene in the video.

The main work in this stage can be divided into three steps. Firstly, each frame image is named according to the video frame order, and the same scene fragments are split one by one. Secondly, the features of the partial images of each entity in each frame of the same scene is extracted and compared with the target image feature to recognize whether the subject is in each video frame of the continuous scene. And the file names of each image that the subject exists are extracted as the subject recognition set. Finally, the features of the partial images of each entity are compared with the recognized object image features to recognize whether the object is in each video frame of the continuous scene, and the file names of each image that the object exists are extracted as the object recognition set.

The scene feature of a video V in a continuous scene is defined as

$$\widehat{V}(X, Y, t) = \frac{V(X, Y, t) - V(X, Y, t_l)}{\sigma(X, Y, t_l)}, \tag{3}$$

where $X \in \{1, 2, \cdots, M\}$, $Y \in \{0, 1, \cdots, N\}$ are scene indices and M and N are the number of subject and object of the video frame, respectively. $t \in \{1, 2, \cdots, T\}$ is the temporal index, and T is number of frames in a video. $V(X, Y, t)$ is the quantity of subject and object at time t. And t_l is the last time the situation changed.

$$\sigma(X, Y, t_l) = \frac{\sum_{t_l}^{t} V(X, Y, t)}{t - t_l}, \tag{4}$$

is the average value of V of each frame from scene change to present.

Since feature extraction and comparison are independent, tasks at this stage can improve detection efficiency through parallel approach. Similarly, nature video has high correlation among neighboring pixels both in space and time. In order to further improve the processing efficiency, we can also choose to extract a picture every few frames for comparison.

3.1.3. The Third Stage: Calculate the Results of Scene Detection. After the subject and object recognition set of the previous stage is obtained, we can integrate them to obtain scene detection results. In the above work, the method of renaming each frame image and taking each image file name as the result set is to reduce the computational load at this stage, so as to improve the efficiency of result integration.

The work in this stage is mainly divided into two steps. Firstly, the intersection part of each object recognition set is taken, and then, the intersection with the subject recognition set is taken. This part of image has two features: (1) the scene information does not change under the same scene and (2) there is no exception in the body. According to the above two features, we can get the video clips of the corresponding frame of intersection images and ensure that the scene of this video is unchanged and the subject is not abnormal.

$$R_{\text{normal}} = (A \cap B_1) \cup (A \cap B_2) \cdots \cup (A \cap B_n), \tag{5}$$

where A is the result frame set that recognizes the subject and B_n is the result frame set that recognizes the object.

Then, the image filenames of the intersection of the image filenames of the subject recognition set and the object

```
Input: entity target set, number of reference objects
Output: object target set
1:encodings is all entity target codes of first_frame
2:   for i in 0 to length of encodings – 1:
3:      if target code equal to the ith entity code of encodings:
4:      adds I to temp list
5:         end
6:   num is the number of reference objects
7:   for i in 0 to num – 1:
8:      rand is a random integer from 0 to length of encodings – 1
9:      while rand is in temp:
10:        rand is a random integer from 0 to length of encodings – 1
11:     adds rand to temp list
12:     adds the rand element of encodings to the lst
```

Algorithm 1: Entity relationship algorithm.

recognition set are compared to obtain the partial images containing the subject but not containing the object. According to the comparison results, we can get the video clips of corresponding frames and determine that the scene changes have been taken place in this video.

$$R_{\text{abnormal}} = \bigcup - R_{\text{normal}}. \tag{6}$$

3.2. Establishment of Spatiotemporal Relationship. A large number of existing models, such as Yolo and face recognition, have been able to realize fast entity recognition of image information. In this paper, such entity recognition results are directly seen as the entity target of video scenes.

After the entity target result set of the video is obtained by using the above models, the entity relationship algorithm is used to establish the relationship between the subject and the object target and establish scene features.

The entity object of the frame is recognized by the entity recognition method and is used as the input of the relationship algorithm. The number of reference objects is determined by the user, and the corresponding number of objects is arbitrarily selected from the entity object as the reference object of the current scene. According to the naive Bayes theory, there is a great similarity between the arbitrarily selected object and the subject in the same scene, and the more the selected objects, the stronger the relationship in the space and time.

3.3. Judgment of Scene Node Attributes. After the establishment of spatiotemporal relationship, the continuity of scene information is firstly detected. Only in continuous scenes can the judgment of scene node attributes have practical application attributes. After obtaining the continuous video clips of scene information, the attributes of scene nodes are determined according to the state of the main body of the video

The subject target and the output results of the relationship algorithm are taken as the input of the judgment algorithm. According to the relationship between the subject and the object, the entity target in the current frame is traversed, and the scene attributes are determined from the subject state and object state.

```
Input: subject coding, object target set
Output: Situational node attributes
1:i = 0
2:   encodings is all entity targets encode of frame
3:   for all the encode of scene's entity targets:
4:      if encode is not in encodings:
5:         i = i + 1
6:   if i equals to scene_entity:
7:      scene has been changed
8:   else:
9:      scene has not been changed
10:encodings is all object targets encode of frame
11:  if target_encode is in encodings:
12:  subject of video is in a particular situation
13:  else:
14:  subject of video is not in a particular situation
```

Algorithm 2: Scene attribute judgment algorithm.

3.4. Video Scene Detection Model. After the completion of entity relationship and scene node attribute judgment, the information of one scene can be detected. However, in general, a video contains multiple scene information. Therefore, on the basis of Algorithm 1 and Algorithm 2, this paper proposes Algorithm 3 to realize the detection of all scene information in a video data.

The content of the first frame of the video is taken as the initial scene information. Algorithms 2 and 3 traverse the video data. When the change of video scene information is detected, the time sequence of the scene change frame is recorded, and the content of the scene change frame is taken as the initial scene information of the subsequent video data to cycle to the end of the video.

4. Experiments

4.1. Experimental Data. The experimental dataset adopted in this paper is a public video dataset; the main content of

```
Input: video data
Output: Scene detection results
for i in 0 to total video frames − 1:
    ic = i + 1
    image is the icth frame of video
    if image is the first frame:
        determine if the subject of video is in the first frame
    else:
        if scene has not been changed:
            determine if the subject of video is in the frame:
                if result is True
                    adds ic to timeImage
                    set nextframe to False
                else:
                    if time lag between now and the last scene is more than 3 s:
                        adds ic to timeImage
                        set nextframe to False
                    else:
                        clear the last record in timeImage
                        set nextframe to True
        else:
            is_end = True
            for i in the number of frames in scene change:
                if scene back to original:
                    is_end = False
                end if
            end for
            if is_end is True:
                adds ic to timeImage
            end if
```

Algorithm 3: Video scene detection model.

which is TV play Ten Miles of Peach Blossom (the data comes from Tencent video, which is only used for academic research in this paper, and the copyright belongs to Tencent company), Hospital Playlist (the data comes from Netflix and is only used for academic research in this paper; the copyright belongs to Netflix company), Nirvana in Fire (the data comes from Tencent video, which is only used for academic research in this paper, and the copyright belongs to Tencent company), and It started with a Kiss. The average scenario switching time of each dataset is 7-10 seconds.

In this experimental environment, in order to analyze the performance of the algorithm proposed in this paper, the evaluation index used in this study is precision.

$$\text{precision} = \frac{\text{correctly recognized number}}{\text{total}}. \tag{7}$$

The values are between 0 and 1, and the closer they are to 1, the better the effect of the model will be.

The hardware configuration information used in the experiment is as follows: CPU R53600, graphics card GTX1660, internal storage 16G, operating system Win10, and development language Python3.

4.2. Experimental Results and Analysis. As mentioned in Section 3, the model proposed in this study is to process and calculate the entity recognition results in the video, so we will use the existing mature entity recognition algorithms in the experiment. Two existing character recognition algorithms are used to meet the needs of model operation. First, face recognition was used to extract environmental features and human face features; second, SeetaFace2 was used to extract environmental features and human face features. The evaluation criteria are whether the target disappears and whether the scene transforms. In this experiment, scene changes have been manually marked. The precision of marking is seconds, and the precision of model detection is video frames. Due to the inconsistency of precision between manual marking and model detection, when the time axes corresponding to the video frame contained in the detection results are the same as that of manual marks, the results are right.

Table 1: The experimental results based on face recognition.

Dataset	Correctly recognized	Total	Wrongly recognized
Dataset1	7	7	0
Dataset2	9	10	1
Dataset3	16	18	0
Dataset4	15	17	0
Dataset5	12	14	0

Figure 3: Incorrect results with face recognition.

4.2.1. Experiments Based on Face Recognition. In this paper, we use Face_recognition as the entity recognition part of the model. As the entity to establish scene features in the model, it is used to recognize the characters in the video image. After establishing the association between the entities in the video (Algorithm 1), the model calculates and determines the scene characteristics of each frame based on Formula (3), records the frame number of $\widehat{V}(X, Y, t)$ difference that is abnormal, and determines the corresponding time point on the time axis. Then, we compare it with the change time which is manually marked and get the test result.

These five datasets have been, respectively, tested, and the experimental results are as shown in Table 1.

Table 2: The experimental results based on SeetaFace2.

Dataset	Correctly recognized	Total	Wrongly recognized
Dataset1	7	7	0
Dataset2	9	10	0
Dataset3	14	18	0
Dataset4	16	17	1
Dataset5	12	14	0

We have found a part of the unrecognized images, as shown in Figure 3.

We have conducted separate recognition processing and found that some frames of the model could not recognize

Figure 4: Incorrect results with SeetaFace2.

their face features, leading to recognition errors during scene detection, which may be caused by decorations on the face.

The experimental results show that face recognition can work well in this dataset. However, face recognition sometimes fails to extract character features because face features are currently used for scene features. The occurrence of the above problems has some serious impacts on the establishment of scene features in the proposed model. Due to the errors of character face recognition, the frame missing a certain entity is wrongly recognized as the changes of scene features during the establishment of scene features. On the whole, the model proposed in this paper does well in scene detection in the case of limited entities.

4.2.2. Experiments Based on SeetaFace2. The experiment is similar to the former one, but we make some changes that we use SeetaFace2 instead of face recognition as the entity recognition part of the model. SeetaFace2 is used to recognize the faces in the video image as the entity to establish scene features of the model and test the same datasets. The experimental results are as shown in Table 2.

We have found a part of the unrecognized images, as shown in Figure 4.

We have found that the SeetaFace2 model recognized faces very sensitively and even can achieve the recognition of supporting characters in the background of photos. And as the camera moves, the number of supporting characters

changes dramatically, leading to the misjudgment of a scene switch.

The experimental results show that the recognition effect of SeetaFace2 is very sensitive, and SeetaFace2 uses the model structure of ResNet50. In this network, multiple residual learning blocks are connected in series, and the deep representation of the deep learning image of the model is utilized, so the recognition effect is very sensitive thus the recognition error occurs. Different from the problems mentioned above, when SeetaFace is combined with the model proposed in this paper, entities will increase abnormally in some frames due to the recognition of background characters. This leads to the errors of the proposed model in establishing scene features, resulting in recognition errors.

4.2.3. Experiment Summary. In addition to the above open dataset experiments, we also used 20 self-made datasets for experiments, and the content of them is meeting recording videos. In order to produce situational changes, the videos have some situations such as characters leaving midway, characters joining midway, and meeting pausing. The results are as follows:

(i) Face recognition: 87%

(ii) SeetaFace2: 85%

The test results meet the expectations. The model proposed in this paper can achieve more accurate scene change detection. It can realize video scene change detection on the premise of using face recognition results as the main entity. The feasibility and universality of the model have been already proved in the experiment. We believe that the accuracy can be further improved if the result including object recognition is introduced as an entity. But in some special cases, such as too many characters in the background, characters turning back, and decorations on the face, it will lead to the failure of scene recognition. In the future, we plan to increase the correct rate of scene transformation recognition by using judgment logic, model recognition, adding background object feature recognition module, and other measures.

5. Conclusion

This paper proposes a video scene information detection based on entity recognition, which can achieve the task of video scene information detection on the premise of entity recognition of video pixel data. The proposed model has strong robustness, and the precision can reach more than 85%. At the same time, it can replace entity recognition with face recognition algorithm as the input of scene information detection without too many impacts on the results of scene information detection.

In this paper, we will explain and verify the above research contents. Section 2 briefly introduces the existing entity recognition models, such as Yolo, and some mature face recognition models, such as face recognition. At present stage, these models are the premise for the test in this study. In Section 3, we introduce the models, including their establishment, mathematical basis and partial content of the pseudocode. Section 4 presents our experimental results and summarizes the failed parts, which are also what need to be further discussed in our subsequent research work. In Section 5, we summarize the research content and briefly introduce the main research directions in the future.

We take the spatiotemporal relationship of video entities as the basis of situational information detection and creatively put forward the concept of situational features to ensure the logical accuracy of the model. In the process of experiments, we found some existing problems, such as overreliance on the accuracy of entity recognition and difficulties in screening noise information effectively. In the research, we focus on how to better combine the entity recognition model with the model proposed in this paper to improve the detection efficiency of the proposed video scene information detection model.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] C. Li, J. Wang, H. Wang, M. Zhao, W. Li, and X. Deng, "Visual-texual emotion analysis with deep coupled video and danmu neural networks," *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1634–1646, 2020.

[2] J. Zhang, K. Mei, Y. Zheng, and J. Fan, "Exploiting mid-level semantics for large-scale complex video classification," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2518–2530, 2019.

[3] T. Prathiba and R. S. S. Kumari, "Eagle eye CBVR based on unique key frame extraction and deep belief neural network," *Wireless Personal Communications*, vol. 116, no. 1, pp. 411–441, 2021.

[4] M. M. Azab, H. A. Shedeed, and A. S. Hussein, "New technique for online object tracking-by-detection in video," *IET Image Processing*, vol. 8, no. 12, pp. 794–803, 2014.

[5] D. Tao, X. Li, S. J. Maybank, and X. Wu, "Human carrying status in visual surveillance," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 1670–1677, New York, NY, USA, 2006.

[6] Z. Jin, Z. Lou, J. Yang, and Q. Sun, "Face detection using template matching and skin-color information," *Neurocomputing*, vol. 70, no. 4-6, pp. 794–800, 2007.

[7] D. Tao, X. Li, X. Wu, and S. Maybank, "Elapsed time in human gait recognition: a new approach," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 2, II-II, Toulouse, France, 2006.

[8] B. Xiao, X. Gao, D. Tao, and X. Li, "A new approach for face recognition by sketches in photos," *Signal Processing*, vol. 89, no. 8, pp. 1576–1588, 2009.

[9] K. Hui, J. Wang, H. He, and W. H. Ip, "A multilevel single stage network for face detection," *Wireless Communications*

and Mobile Computing, vol. 2021, Article ID 5582132, 10 pages, 2021.

[10] G. Zheng and Y. Xu, "Efficient face detection and tracking in video sequences based on deep learning," *Information Sciences*, vol. 568, pp. 265–285, 2021.

[11] Y. Pang, Y. Yuan, X. Li, and J. Pan, "Efficient HOG human detection," *Signal Processing*, vol. 91, no. 4, pp. 773–781, 2011.

[12] P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic, "Finding actors and actions in movies," in *Proceedings of the IEEE international conference on computer vision*, pp. 2280–2287, Sydney, NSW, Australia, 2013.

[13] J. Lv, W. Liu, L. Zhou, B. Wu, and H. Ma, "Multi-stream fusion model for social relation recognition from videos," in *MultiMedia Modeling. MMM 2018*, vol. 10704, Springer, Cham, 2018.

[14] L. Ding and A. Yilmaz, "Learning relations among movie characters: a social network perspective," in *Computer Vision – ECCV 2010. ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds., vol. 6314 of Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2010.

[15] Q. D. Tran and J. E. Jung, "CoCharNet: extracting social networks using character co-occurrence in movies," *Journal of Universal Computer Science*, vol. 21, pp. 796–815, 2015.

[16] X. Gong and F. Wang, "Classification of tennis video types based on machine learning technology," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 2055703, 11 pages, 2021.

[17] Y. Li, "Research on sports video image analysis based on the fuzzy clustering algorithm," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 6630130, 8 pages, 2021.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, Las Vegas, NV, USA, 2016.

[19] S. Zhang, Y. Wu, C. Men, and X. Li, "Tiny YOLO optimization oriented bus passenger object detection," *Chinese Journal of Electronics*, vol. 29, no. 1, pp. 132–138, 2020.

[20] M. A. A. al-qaness, A. A. Abbasi, H. Fan, R. A. Ibrahim, S. H. Alsamhi, and A. Hawbani, "An improved YOLO-based road traffic monitoring system," *Computing*, vol. 103, no. 2, pp. 211–230, 2021.

[21] Y. Lei, D. Peng, P. Zhang, Q. Ke, and H. Li, "Hierarchical paired channel fusion network for street scene change detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 55–67, 2021.

[22] M. C. S. Santana, L. A. Passos, T. P. Moreira, D. Colombo, V. H. C. de Albuquerque, and J. P. Papa, "A novel Siamese-based approach for scene change detection with applications to obstructed routes in hazardous environments," *IEEE Intelligent Systems*, vol. 35, no. 1, pp. 44–53, 2020.

[23] Chung-Lin Huang and Bing-Yao Liao, "A robust scene-change detection method for video segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 12, pp. 1281–1288, 2001.