Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 7862343, 8 pages
https://doi.org/10.1155/2022/7862343

Research Article

Detection of Shot Transition in Sports Video Based on Associative Memory Neural Network

Wanli Ke

Henan Institute of Economics and Trade, Zhengzhou, Henan 450046, China

Correspondence should be addressed to Wanli Ke; kewanli@henetc.edu.cn

Received 9 January 2022; Revised 28 January 2022; Accepted 31 January 2022; Published 28 February 2022

Academic Editor: Xin Ning

Users must quickly and effectively classify, browse, and retrieve videos due to the explosive growth of video data. A variety of shots make up the video data stream. The most important technology in video retrieval is shot detection, which can fundamentally solve many problems, resulting in improved detection effects and even directly affecting video retrieval performance. This paper investigates the shot transition detection algorithm in digital video live broadcasts based on sporting events. To solve the problem of shot transition detection using a single training sample, an AMNN (Associative Memory Neural Network) model with online learning ability is proposed. Experiments on a large football video data set show that this algorithm detects shear and gradual change better than existing algorithms and meets the application requirements of sports video retrieval in most cases.

1. Introduction

Among the many new forms of communication, video has the most content, and it can provide a wide range of information that is more specific, rich, and vivid than text, sound, or image [1]. Because of the variety of information contained in the video, there is a lot of video data. Automatic analysis of broadcast sports videos is an important application field in video retrieval technology. Shot transition detection in sports video is the most difficult application field [2, 3]. Because a video stream is made up of a series of image frames, extending the method of image-based retrieval to content-based video retrieval is a natural idea. However, the method of still image retrieval cannot be directly applied to video because videos often contain many video frames with varying degrees of similarity [4].

AMNN (Associative Memory Neural Network), as a shot change detection method, has certain advantages. It can acquire the implicit expression of the rules and rules of shot change detection through learning. It has strong adaptability and is easy to realize and is not constrained by a single training sample. It is a nonlinear technology that does not require any data distribution and can effectively store and recall information [5, 6]. In shot transition detection, some studies have proposed a unified model using histogram features, which works well on large data sets [7]. In the keyframe extraction and content-based video retrieval, it is necessary to switch the detection shot. Obviously, people always want the network to have multimode recall ability. A shape recognition system based on associative memory is proposed in reference. For feature extraction, a discrete Hough transform based on a feedforward neural network [8] is used, and for recognition and classification, a third-order AMNN based on mean-field theory is used. However, the BPNN (BP neural network) connection weight has a direct impact on the detection effect of shot transition in sports video [9]. Currently, the initial value of connection weight is mostly determined by a random method, which is very subjective, making the detection stability of shot transition in sports video insufficient, and the detection of shot transition in sports video requires additional research.

Because the essence of shot transition is the semantic discontinuity of image sequence features, existing algorithms' underlying features are insufficient to solve problems in sports videos [10]. Although object-level features (high-level features) are the best way to express semantics, analyzing videos at the object level is still a difficult task today. Shot transition detection methods include the histogram

method, pixel-based method, motion feature method, and others. This paper examines the detection algorithm for shot transition in sports video based on sporting events.

2. Related Work

Many research institutions and universities have done a lot of in-depth analysis and research on the detection of shot transition in sports videos. In recent years, many literature studies related to the detection of shot transition in sports videos have appeared, and many effective research results of shot transition detection in sports videos have been obtained [9, 11]. Reference [12] samples the video frame sequence locally in the horizontal, vertical, and diagonal directions to form a spatiotemporal slice, which is a two-dimensional picture made by orderly splicing multiple frames of sampling information in the same direction. Comparing the two methods of pixel value and histogram, it is found that histogram can meet the requirements of speed and accuracy of video boundary detection [13]. Reference [14] put forward the idea of block histogram. Reference [15] uses the Canny operator to detect the edge of each frame of the image, then calculates the in-and-out degree of the edge pixels of the image to get the edge change rate, and uses the edge change rate to detect the shot boundary [16]. By detecting the edge features of candidate shot boundaries, the false detection of shot boundaries caused by flash can be avoided and the calculation time can not be greatly increased. Reference [17] gets the discrete cosine coefficients of B and P frames through motion compensation in video coding, which solves the problem that the boundary between B and P frames cannot be detected. In [18], the boundary coefficient of shot is constructed by calculating the frame difference of multiple frames with an interval of n in a fixed sliding window. Experiments show that the boundary coefficient has strong anti-interference ability to noise.

The fuzzy neural network is a hybrid of artificial neural networks and fuzzy logic systems that has both advantages and the ability to explain fuzziness and has emerged as a key modeling tool for fuzzy intelligent systems. As a result, many researchers began to focus on neural network architecture. According to the findings, the real biological neural system is a complex network with sparse interconnections [19]. To train connected networks, [20] used fuzzy rules. Later research discovered that this algorithm had a subset recall feature and could not guarantee the safe storage of multiple fuzzy pattern pairs. Reference [21] proposed an orthogonal coding learning algorithm, in which the network can store multiple pattern pairs if the input pattern pairs meet the regular and max-min orthogonal conditions. However, meeting the orthogonal condition is difficult in general. Reference [22] demonstrates that AMNN with a small-world system can achieve the same retrieval performance as a random interconnection network while using a quarter of the connections. The combination of fuzzy theory and neural networks in [23] has both advantages. Reference [24] introduces the small-world system to fuzzy AMNN, builds fuzzy self-AMNN based on small-world architecture, optimizes the network structure, reduces the network's time and space complexity, and maintains optimal network performance while lowering network connection costs.

3. Research Method

3.1. Analysis of Shot Transition in Sports Video. Shot segmentation is the basis of hierarchical video structure, which requires that all kinds of complicated edited shot boundaries can be correctly detected, and the motion changes in shots can be effectively distinguished, so as to eliminate their interference with shot boundary recognition. Keyframes are the representatives of shots and the important basis of video retrieval. Keyframes should be able to reflect the main movements and changes in shots. The lens is the continuous image frame recorded by the camera from opening to closing, and it is the smallest physical unit in the video. Inside the shot, the features of adjacent and similar video frames are similar, with little change. However, at the shot transition, the features of video frames often change obviously.

Another popular method is to use the histogram of each frame as a feature. The main idea behind this method is that the histograms of successive frames in a shot have similar global visual properties, so the difference between their histograms is smaller than the difference between the histograms of two frames at the shot boundary. Because the histogram only considers the global distribution, it is unaffected by local object motion; however, when the video shot contains global motion, the histogram will change dramatically. The basic algorithm of a histogram-based method is to divide the color space into discrete color cells one by one and then count the number of pixels that fall into each cell.

Assuming that the color space is divided into n intervals, and K_{ki} is the number of pixels falling into the i th color interval in the k frame, the interframe difference can be expressed by the following formula:

$$D_{k,k+1}(I) = \sum_{i=1}^{n} \left| H_{ki} - I_{(k+1)i} \right|. \tag{1}$$

When two images have completely different structures, their color histograms may be very close, causing the shot transformation to be missed occasionally, which is a drawback of the histogram-based method. The edge detection method takes advantage of the video shot's edge features. The basic idea is that when the video shot changes, the new edge should be far away from the old edge, and the position where the old edge vanishes should also be far away from the new edge.

Firstly, the algorithm extracts the edge map E_k and E_{k+1} of the two frames of video images k and $k+1$, and the difference between the two frames of video images is calculated by the following formula:

$$\text{diff} = \max\left(d_{in}, d_{out}\right), \tag{2}$$

where d_{in} is the proportion of incoming pixels, that is, new pixels far away from the existing edge, and d_{out} is the proportion of outgoing pixels, that is, new pixels far away from the new edge.

As a result, selecting only the features that are unaffected by these variables is still insufficient. Because a flash is typically a change in brightness, all of the color components will change in the event of some abrupt brightness changes with high intensity, resulting in higher discontinuous measurement values even if the histogram method is used, which is easily mistaken for a change of video shot.

The field of video classification is vast. In many applications, it is also known as video interpretation or video comprehension. Video in this context refers to a multimedia sequence that includes both auditory and visual elements and contains a series of images and sounds at the same time. There are many research contents in the field of automatic video classification, as shown in Figure 1.

Video classification is conceptually divided into two categories:

(1) Inherent characteristics of digital video: generally, this refers to editing effects, mainly including camera translation, zoom, rotation, and other editing effects. Although they are part of the video and have something to do with the story in an imprecise sense, they are not part of the story itself.

(2) Semantic classification: that is, video is classified from its inherent semantic content; for example, broadcast video can be divided into cartoons, news, movies, sports, and other types. The content of this paper is the widely used content-based video classification.

In the end, each event is composed of objects, which is the lowest level of classification, and it directly affects the semantic information of the video. The object that is often detected is the face [11]. The detection of objects needs good structural feature extraction. Generally, there are two ways to realize the first one. The rule-based method is used to introduce people's understanding and prior knowledge in the process of feature extraction, such as the structure of people's five senses. The second is to use the pattern learning method; that is, the sample of the object is first given for learning.

People have spent years researching video classification technology and have proposed a variety of content-based video classification methods and systems, ranging from low-level event detection methods with limited detection range to advanced type classification methods with a broad detection range. However, the most important technologies for these systems to research are video segmentation, video feature extraction and data processing, and video classification. There are some issues with the existing algorithms for detecting shot transitions in sports videos. Because the local color histograms for the first frame after the shot transition and the last frame before the transition are so similar, it is easy to miss the detection. Rapid shaking will occur when the camera tracks the running athletes in close-up. In today's sports video, most of the images corresponding to gradual changes in the distance feature sequence of the local color histogram no longer have simple peak shapes, but the patterns are more complicated, and most of the animation changes are missed.

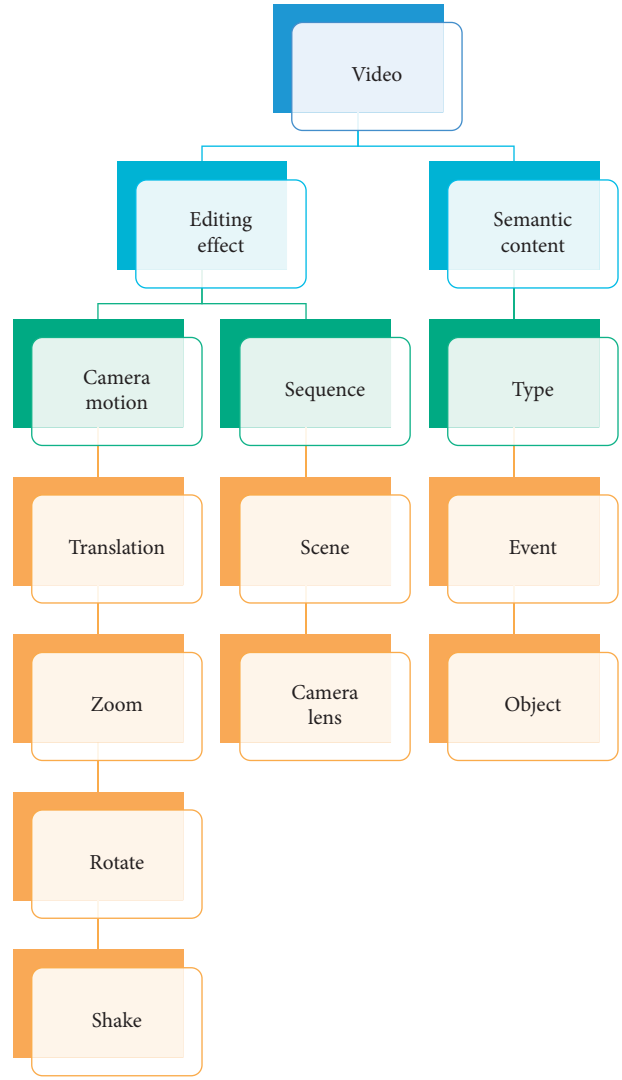

Figure 1: Generalized classification level of the video.

3.2. Design and Implementation of Shot Shear Detection Algorithm. By constructing the feature vector of each frame in the video, the shot boundary and nonboundary can be distinguished in features. The next step is to use some classification methods to classify the features. Considering the effectiveness and usability of each classifier, SVM (Support Vector Machine) classifier is used to detect shot boundary. Memory is an important part of neural network theory, and it is also an important function of neural networks for intelligent control, pattern recognition, and artificial intelligence. It mainly makes use of the good fault tolerance of neural networks, which can restore incomplete, defaced, and distorted input samples to a complete prototype, and is suitable for identification and classification.

The learning process of AMNN, like that of other neural networks, is the formation of network connection weight matrix teeth. AMNN is a constantly changing system. Once the network's connection weight matrix is learned and formed, as long as a specific pattern sample is an input, the network will evolve indefinitely until the system reaches the

state space's steady state (fixed attractor). The memory stage, also known as the storage stage or the learning stage, and the association stage, also known as the recovery stage or the recall stage, are the two stages that make up the working process of association memory. *Stage of Memory*. Set the initial network weights and adjust them using a learning algorithm until the network has many stable equilibrium states, also known as attractors, and attractors have a specific attraction domain. Lenovo has a stage. Association is a process in which a trained association-memory network converges to the attractor and associates the existing mode with a given input mode. Faulty self-associative memory has the same structure as general self-associative memory. It can be made up of a single-layer feedback network module, as shown in Figure 2.

The weight matrix whose memory matrix is transformed into the network module in the corresponding network module is a matrix determined by fuzzy relation, namely,

$$W = \max_k\left(A_k \circ A_k^T\right) = \left(w_{ij}\right),$$
$$w_{ij} = \max_k\left(a_{ki} \circ a_{kj}\right). \tag{3}$$

Here, the mark "$\circ$" indicates some fuzzy synthesis operator, such as taking the maximum and taking the minimum.

Attention is an important psychological adjustment mechanism in the process of human information processing. It can allocate limited information processing resources and make perception have the ability to choose. At present, the salient region detection algorithms are divided into two categories, one is based on local features, the other is based on visual contrast, and the latter is mainly proposed according to the process of visual perception, which has strong versatility. This paper extracts the features of visual attention on the basis of this algorithm. The specific visual feature extraction process is shown in Figure 3.

A 5×5 Gaussian filter is used to operate the low-level visual features of each frame image, a seven-layer pyramid model is obtained for each feature channel, and each frame image is constructed according to the following formula:

$$g_{l+1}(i, j) = \sum_{m=-2}^{0}\sum_{n=-2}^{0} w(m, n) g_l(2i + m, 2j + n), \tag{4}$$

where $g_{l+1}(i, j)$ represents the pixel point at (i, j) in L layer of pyramid, and $w(m, n)$ is the kernel function of Gaussian filtering.

After calculation, a Gaussian pyramid model is obtained. The bottom layer of the pyramid is the original image, and the ratio of the number of horizontal and vertical pixels in the upper layers to the number of pixels in the original layer is 1 : 1 (the original layer 0) to 1 : 64 (the highest layer 6), respectively.

Assuming that N neurons are arranged in a one-dimensional circular array and numbered $(1, 2, \ldots, N)$ according to their positions in turn, then $D(i, j)$ represents the lateral feedback intensity of the $D(i, j)$ th neuron to the j neuron, and $d_{j,i}$ represents the distance between the i th

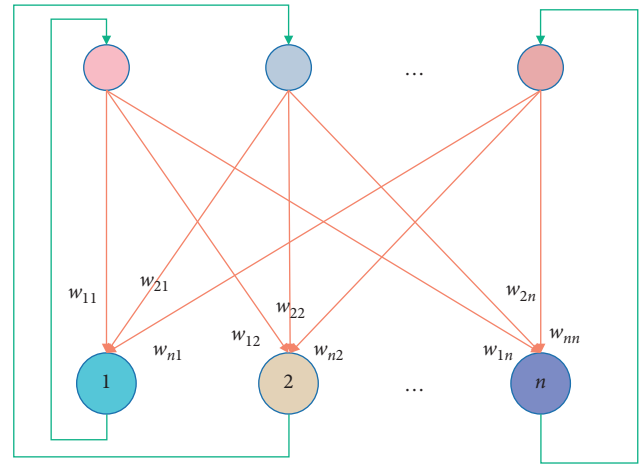

Figure 2: Fuzzy self-associative memory structure.

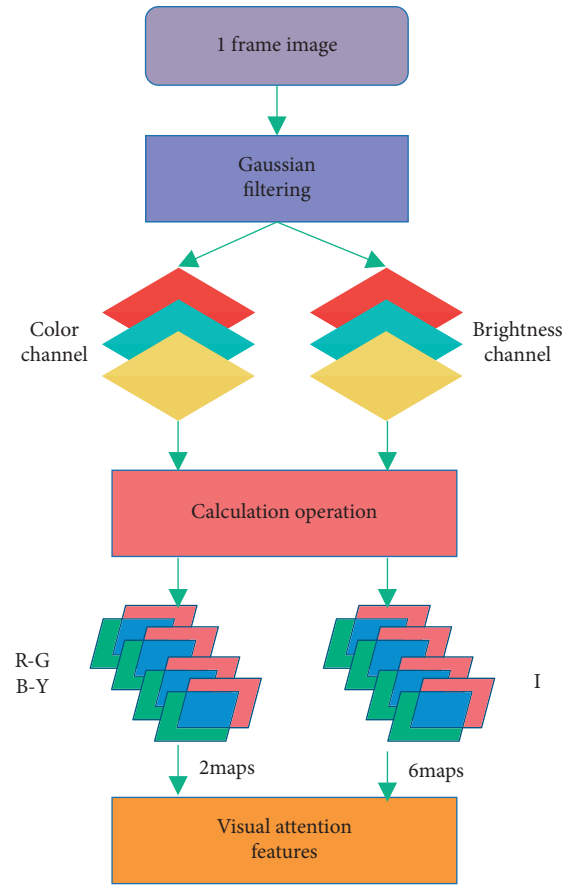

Figure 3: Flow chart of visual feature extraction.

neuron and the j neuron in the ring. The distance adopted in this paper is defined as follows:

$$d_{j,i} = \min\left(\mathrm{mod}\left((i - j), N\right), \mathrm{mod}\left((j - i), N\right)\right). \tag{5}$$

The network is trained by a learning algorithm based on MAX – T_{es}, and its connection weight matrix is obtained after training. Then, the weight matrix w_k is integrated by fuzzy minimization operation $\cap$, and the weight matrix of the final network is obtained:

$$W = \underset{k\subset K}{\cap} w_k = \left(\underset{k\subset K}{\wedge} \left(a_{ki} R_{T_{es}} a_{kj} \right) \right)_{n\times m}. \quad (6)$$

According to common sense, the objects or areas closer to the central part of the image are more important than the areas around the image, so the neighborhood contrast of each macroblock is weighted to calculate the static attention of the image. The formula for calculating the static attention of the whole image is

$$A_s = \frac{1}{N} \sum_{x=1}^{N} w_{i,j} * C_{i,j}, \quad (7)$$

where $w_{i,j}$ is the Gaussian weight value of the location of macroblock (i, j).

4. Results Analysis and Discussion

Mutual information can be used to determine how similar two things are. The mutual information between two things will be smaller if they are more similar, but it will be larger if they are not. Because the frame images within a shot are very similar, but the frame images between different shots are very different, consider using mutual information to compare two frames. The experiment is conducted on a data set of 12 video sequences, each of which represents a half-court football match. In this paper, the first six sequences are used to train SVMs, and the last six sequences are used to test the algorithm. All video sequences have a resolution of 754 680 pixels and a frame rate of 29.07 frames per second (frames per second). These videos were shot in various stadiums and edited by various television stations, and they are sufficient to put the algorithm in this paper to the test. The results of the algorithm on the test set are shown in Figure 4.

From the overall results, the middle-level features proposed in this paper obviously improve the performance of the shot transition detection algorithm. At the same time, it can be found that the improvement of the algorithm in gradient detection is more obvious. This algorithm can achieve high precision and recall in all test sequences, so it can be fully applied to automatic sports video analysis. At the same time, the computational complexity of this algorithm is acceptable in practical application.

When there is global motion in the video shot, the histogram will change greatly. However, the disadvantage of the histogram-based method is that sometimes the detection of video shot changes will be missed, because it is possible that two video images have completely different structures, but their histograms are very close so that the missing detection will occur. However, the block-based method is proposed to solve the defect that histogram-based method is easy to miss detection, but the block-based method is very sensitive to the local motion in the video shot. Figure 5 shows the feature results obtained by analyzing abrupt and gradual shots.

In general, the existence of cumulative frames will increase the difference between gradient frames. The accumulation process will continue to increase as an object moves, and the judging link will primarily occur when the threshold reaches the appropriate level. The gradual boundary is more complicated than the abrupt boundary. To begin with, the length of gradual change can range from a few frames to hundreds of frames, necessitating the detection algorithm's ability to examine the entire gradual change process. Second, the transition is relatively gentle; the difference between adjacent frames is small, and there is no discernible difference from the normal frame difference. This necessitates the detection algorithm to be tolerant of a small number of frames that change gradually. Single static features and dynamic features are chosen for comparison testing in order to assess the benefits of multifeatures. The modeling methods have all been improved by AMNN, and

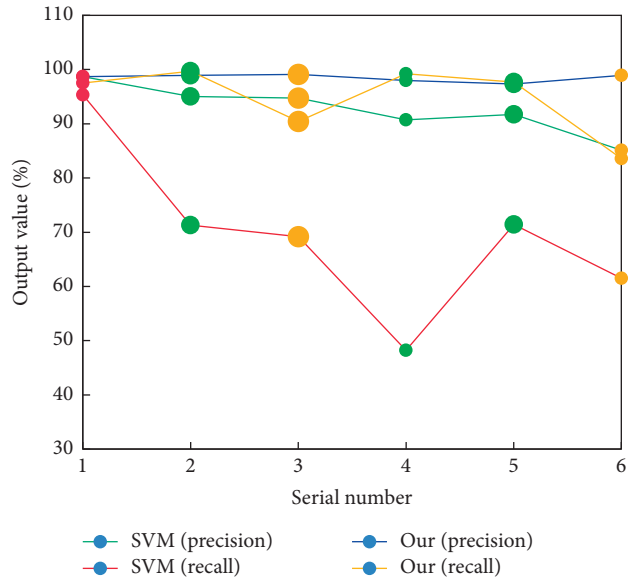

Figure 4: Experimental result.

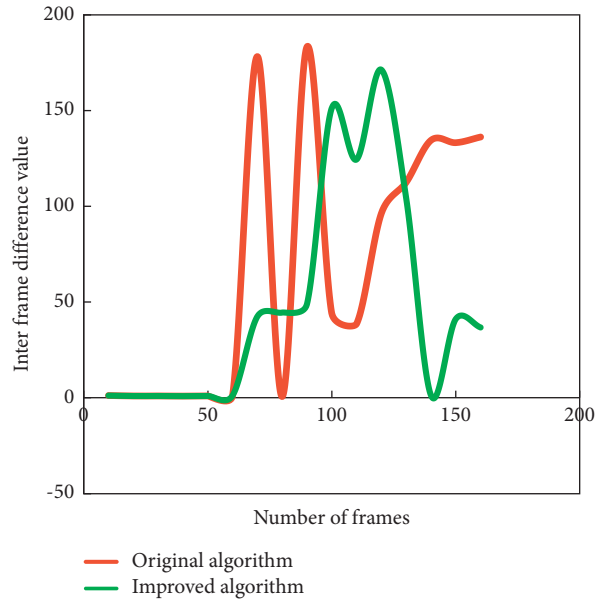

Figure 5: Comparison of test simulation between the original algorithm and the improved algorithm.

Figure 6 shows the correct rate of shot transition detection in sports video.

By comparing and analyzing the correct rate of shot transition detection in sports video in Figure 6, it can be found that the correct rate of shot transition detection in sports video with a single static feature and dynamic feature is low because the complete shot information of sports video can not be provided, which leads to a larger error of shot transition detection in sports video. Because it can describe sports video shots more completely and obtain more ideal sports video shot transition detection results, multifeature sports video shot transition detection has a high accuracy, which reduces the error of sports video shot transition detection. For each frame of the image, the model generates a value of dynamic attention and a value of static attention and fuses them to form the value of the user's attention. The values of three models of a video frame sequence change with the frame sequence as shown in Figure 7.

From the observation in Figure 7, it can be found that when the dynamic attention model increases, the value of the fused model also increases significantly, and the value of dynamic attention tends to be consistent with that of the fused user's attention. That is, the frame with the highest value of the user's attention after fusion is used as the keyframe of the whole video sequence. In order to enhance the robustness of the algorithm, the following conditions are added to the maximum value detected. First, the average value of the neighborhood sequence is calculated, and then its variance is calculated. Then, the distance difference between the maximum value and each element in the neighborhood sequence is calculated. If the maximum value of the difference is greater than the variance, it is considered that the maximum value satisfies the condition and is assigned +1; otherwise, it is assigned −1. Through the analysis of the experimental results, the characteristics of visual attention are sensitive to the rapid movement of objects. Based on the statistical learning theory, a new pattern recognition method SVM is developed. It can transform practical problems into high-dimensional feature space by nonlinear transformation, construct linear functions in high-dimensional space to realize nonlinear discriminant functions in the original space, and skillfully solve the dimension problem. Its algorithm complexity has nothing to do with the dimension of samples. Figure 8 shows the comparison of recognition performance of various algorithms after adding different levels of salt and pepper noise.

It can be seen that, even for human beings, the recognition rate of the AMNN model for five different levels of salt and pepper noise (from 5% to 25%) is almost unaffected and stable at around 88%, while the performance of the other two algorithms decreases significantly with the increase of noise intensity. By extracting the dynamic features and static features of sports video shots, the defect that the current single dynamic features or static features have simple information and cannot describe the sports video shot transition is well solved, which is conducive to improving the detection accuracy of sports video shot transition. That is to say, each peak value is a maximum value in a specific

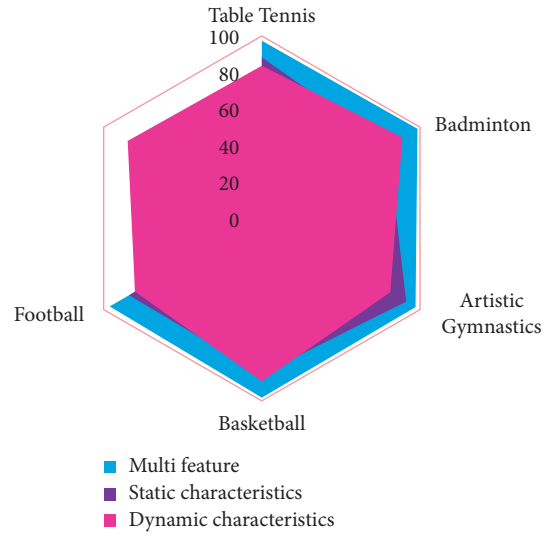

Figure 6: Comparison of the accuracy rate of single feature sports video shot transition detection.

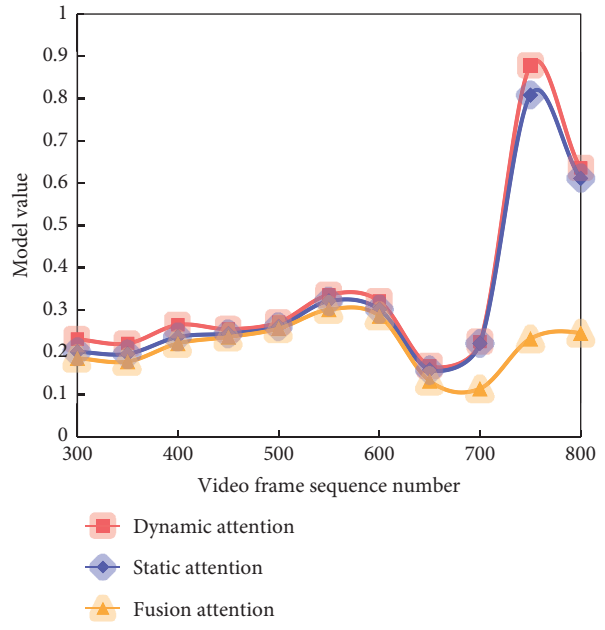

Figure 7: Model value diagram.

neighborhood. The search algorithm for superimposed signal and finding out the maximum value in this neighborhood is equivalent to finding out the peak value. However, this does not guarantee that we are searching for characteristic peaks because if the above conditions can be met for some noise signals, it may cause detection errors.

Figures 9 and 10 all show that the AMNN model is obviously superior to other algorithms in all cases of mosaic occlusion. Moreover, when the upper and lower occlusion rates are lower than 25%, respectively, the recognition rate of the AMNN model is acceptable. Only when the occlusion rate is higher than 35%, the recognition rate drops sharply with the increase of occlusion rate, which intuitively accords with the human recognition mechanism.

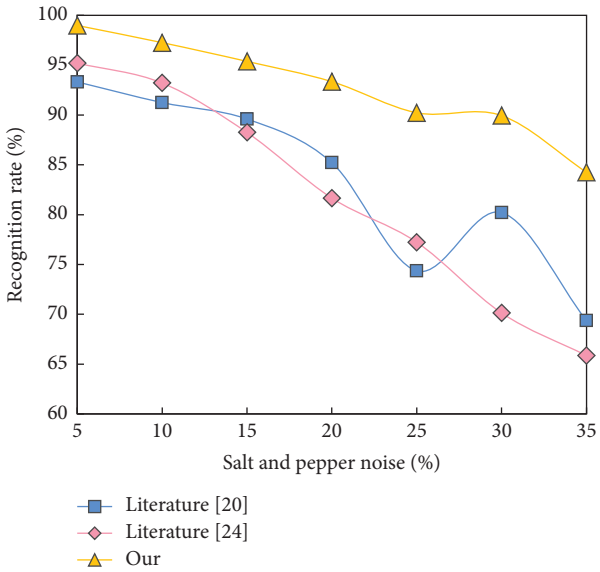

Figure 8: Comparison of recognition rate of three algorithms in the presence of salt and pepper noise.

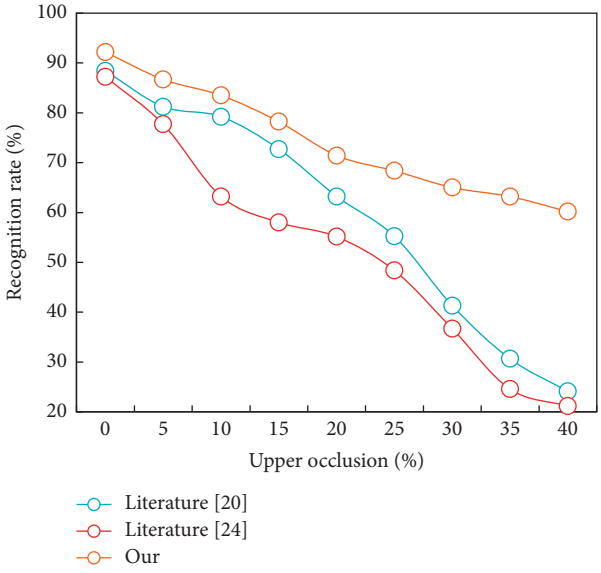

Figure 9: Comparison of the recognition rate of three algorithms in the case of upper occlusion.

The AMNN algorithm achieves a higher recognition rate and shows stronger robustness than the algorithms in literature [19] and literature [23]. Further research includes the following: introducing a small-world network or scale-free network structure to sparse the AMNN model. Block and weighting methods are used to give higher weights to important parts of the face, such as the eyes and mouth, to optimize the connection parameters. From the sequence value, it can be seen that although the sequence can ideally represent the peak number characteristics of the pattern to be identified because AMNN requires the input signal to be binary, finally, binarization is performed. If the quadratic difference value is used to process the parameters, if it is determined that the latter of the two adjacent frame

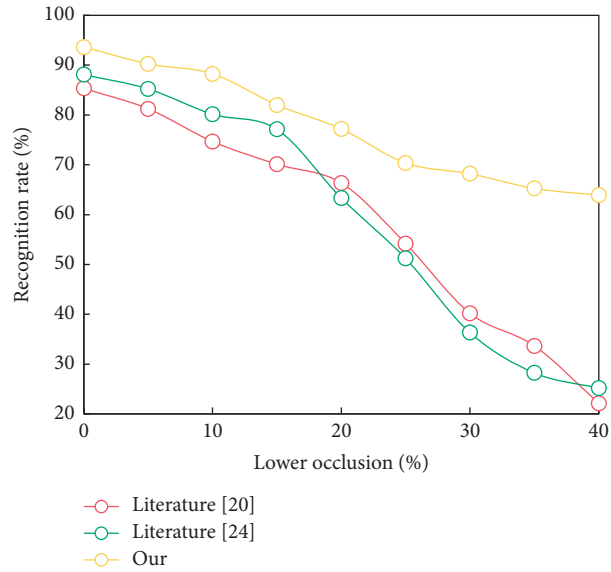

Figure 10: Comparison of recognition rate of three algorithms under occlusion.

differences is negative, the probability of false detection of the former as an abrupt frame is very high. After improving the correlation algorithm, find the position of the quasi-mutation and, at the same time, perform flash detection to reduce the probability of false detection as much as possible. For example, some signal peaks are generally too large, and if the threshold value is set too low, there will be too many peaks to keep the characteristics. If the closed value of some signals is set too high, the peak value of the signals will be eliminated and the features will be obliterated. In fact, if this problem cannot be solved well, it will greatly reduce the ability of network identification.

However, the result of this algorithm is quite satisfactory, because the binary signal sequence with this effect can be completely processed into the ideal peak value by the later AMNN processing. Through our comparative experiments, compared with the previous manual thresholding method, this algorithm has the advantages of strong adaptability and an obvious coding effect. Images with different backgrounds can be automatically encoded without changing the closed value, and the effect is ideal. The significance of this algorithm is that it plays a great role in the practicality of the shape recognition system.

5. Conclusion

This paper examines the detection algorithm of shot transition in digital video live broadcasts based on sporting events. By judging whether there is a shot transition in the accumulated frame difference and improving the algorithm based on the accumulated frame, it is more effective in judging the gradual change and shear shot. In the case of a single training sample, an AMNN model is proposed to solve the problem of shot transition detection. SVM is used to determine the connection weights and thresholds of AMNN, which overcomes the flaw of traditional neural networks'

random determination, improves AMNN's learning performance, establishes a better detection model of sports video shot transition, lowers the detection error rate of sports video shot transition, and can better describe sports video shot. The user attention model-based keyframe extraction algorithm can extract keyframes, represent the entire shot, and effectively summarize video content while reducing data redundancy, which has a wider practical application value.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The author does not have any possible conflicts of interest.

References

[1] K. Calagari, M. Elgharib, P. Didyk, A. Kaspar, W. Matusik, and M. Hefeeda, "Data driven 2-D-to-3-D video conversion for soccer," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 605–619, 2018.

[2] A. Amiri and M. Fathy, "Video shot boundary detection using QR-decomposition and Gaussian transition detection," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, Article ID 509438, 12 pages, 2009.

[3] N. Yaghoobian and R. Mittal, "A computational approach for predicting plant canopy induced wind effects on the trajectory of golf shots," *Sports Engineering*, vol. 21, no. 1, pp. 1–10, 2018.

[4] Y. Lu and S. An, "Research on sports video detection technology motion 3D reconstruction based on hidden Markov model," *Cluster Computing*, vol. 23, no. 3, pp. 1899–1909, 2020.

[5] H. M. Mesbah, "The impact of linear versus nonlinear listening to radio news on recall and comprehension," *Journal of Radio Studies*, vol. 13, no. 2, pp. 187–200, 2006.

[6] R. K. Shen, Y. N. Lin, T. Y. Juang, and V. R. L. Shen, S. Y. Lim, Automatic detection of video shot boundary in social media using a hybrid approach of HLFPN and keypoint matching," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 1, pp. 210–219, 2018.

[7] A. Tejero-De-Pablos, Y. Nakashima, T. Sato, N. Yokoya, M. Linna, and E. Rahtu, "Summarization of user-generated sports video by using deep action recognition features," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2000–2011, 2018.

[8] X.-B. Jin, W.-Z. Zheng, J.-L. Kong et al., "Deep-learning temporal predictor via bidirectional self-attentive encoder-decoder framework for IOT-based environmental sensing in intelligent greenhouse," *Agriculture*, vol. 11, no. 8, p. 802, 2021.

[9] L. Yu, "Sports activity detection, organization and evaluation in online to offline sports community," *Cognitive Systems Research*, vol. 52, pp. 785–792], 2018.

[10] F. Yang, H. Chen, J. Li, F. Li, L. Wang, and X. Yan, "Single shot multibox detector with kalman filter for online pedestrian detection in video," *IEEE Access*, vol. 7, pp. 15478–15488, 2019.

[11] Y. Li, Y. Pang, J. Cao, J. Shen, and L. Shao, "Improving single shot object detection with feature scale unmixing," *IEEE Transactions on Image Processing*, vol. 30, 2021.

[12] X. Miao, X. Liu, J. Chen, S. Zhuang, J. Fan, and H. Jiang, "Insulator Detection in Aerial Images for Transmission Line Inspection Using Single Shot Multibox Detector," *IEEE Access*, vol. 7, 2019.

[13] A. A. Cabrera-Ponce, L. O. Rojas-Perez, J. A. Carrasco-Ochoa, J. F. Martinez-Trinidad, and J. Martinez-Carranza, "Gate detection for micro aerial vehicles using a single shot detector," *IEEE Latin America Transactions*, vol. 17, no. 12, pp. 2045–2052, 2019.

[14] Y. Sheng, Z. Zeng, and T. Huang, "Global stability of bidirectional associative memory neural networks with multiple time-varying delays," *IEEE Transactions on Cybernetics*, pp. 1–10, 2020.

[15] C.-J. Xu, P.-L. Li, and Y.-C. Pang, "Finite-time stability for fractional-order bidirectional associative memory neural networks with time delays," *Communications in Theoretical Physics*, vol. 67, no. 2, p. 137, 2017.

[16] T. Kumamoto, M. Suzuki, and H. Matsueda, "Singular-value-Decomposition analysis of associative memory in a neural network," *Journal of the Physical Society of Japan*, vol. 86, no. 2, 2016.

[17] F. Wang, Y. Chen, and M. Liu, "Pth moment exponential stability of stochastic memristor-based bidirectional associative memory (BAM) neural networks with time delays," *Neural Networks:the Official Journal of the International Neural Networks Society*, vol. 98, pp. 192–202, 2018.

[18] S. Recanatesi, M. Katkov, and M. Tsodyks, "Memory states and transitions between them in attractor neural networks," *Neural Computation*, vol. 29, no. 10, pp. 2684–2711, 2017.

[19] Y. Shen, S. Zhu, X. Liu, and S. Wen, "Multistability and associative memory of neural networks with Morita-like activation functions," *Neural Networks: the Official Journal of the International Neural Networks Society*, vol. 142, pp. 162–170, 2021.

[20] A.-H. Tan, B. Subagdja, D. Wang, and L. Meng, "Self-organizing neural networks for universal learning and multimodal memory encoding," *Neural Networks*, vol. 120, pp. 58–73, 2019.

[21] H. Schuetze, E. Barth, and T. Martinetz, "Learning efficient data representations with orthogonal sparse coding," *IEEE Transactions on Computational Imaging*, vol. 2, no. 3, pp. 177–189, 2016.

[22] S. Gupta, M. Imani, H. Kaur, and T. S. Rosing, "NNPIM: a processing in-memory architecture for neural network acceleration," *IEEE Transactions on Computers*, vol. 68, 2019.

[23] J. Wang and Y. Li, "Multi-step ahead wind speed prediction based on optimal feature extraction, long short term memory neural network and error correction strategy," *Applied Energy*, vol. 230, pp. 429–443, 2018.

[24] A. Agrawal, A. Ankit, and K. Roy, "SPARE: spiking neural network acceleration using ROM-embedded RAMs as in-memory-computation primitives," *IEEE Transactions on Computers*, vol. 68, no. 8, pp. 1190–1200, 2019.