Intelligent Automation & Soft Computing

Tech Science Press

DOI: 10.32604/iasc.2023.038444
Article

Fuzzy Rule-Based Model to Train Videos in Video Surveillance System

A. Manju[1], A. Revathi[2], M. Arivukarasi[1], S. Hariharan[3], V. Umarani[4], Shih-Yu Chen[5,*] and Jin Wang[6]

[1]Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India
[2]Department of Computational Intelligence, SRM Institute of Science and Technology, Kattankulathur, Chennai, India
[3]Department of Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad, India
[4]Department of Computer Science and Engineering, Saveetha Engineering College, Chennai, Tamilnadu, India
[5]Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Taiwan
[6]School of Computer & Communication Engineering, Changsha University of Science & Technology, Changsha, 410004, China
*Corresponding Author: Shih-Yu Chen. Email: sychen@gemail.yuntech.edu.tw
Received: 13 December 2022; Accepted: 20 February 2023

Abstract: With the proliferation of the internet, big data continues to grow exponentially, and video has become the largest source. Video big data introduces many technological challenges, including compression, storage, transmission, analysis, and recognition. The increase in the number of multimedia resources has brought an urgent need to develop intelligent methods to organize and process them. The integration between Semantic link Networks and multimedia resources provides a new prospect for organizing them with their semantics. The tags and surrounding texts of multimedia resources are used to measure their semantic association. Two evaluation methods including clustering and retrieval are performed to measure the semantic relatedness between images accurately and robustly. A Fuzzy Rule-Based Model for Semantic Content Extraction is designed which performs classification with fuzzy rules. The features extracted are trained with the neural network where each network contains several layers among them each layer of neurons is dedicated to measuring the weight towards different semantic events. Each neuron measures its weight according to different features like shape, size, direction, speed, and other features. The object is identified by subtracting the background features and trained to detect based on the features like size, shape, and direction. The weight measurement is performed according to the fuzzy rules and based on the weight measures. These frameworks enhance the video analytics feature and help in video surveillance systems with better accuracy and precision.

Keywords: Video analytics; video semantic substance model; fuzzy rule; image processing

1 Introduction

The growing population of human society has introduced several challenges in different corners of the world. Human society accesses various services through their mobile phones to access different data. Multimedia data is the one in form of big data being used for different analyses. Multimedia data has been used in different issues and they are captured, indexed, and mined to retrieve such big data [1]. The multimedia data include text, image, audio, and video data. The recent development in information technology has represented the data of huge volume and size as big data. The increasing volume and size of data have been represented in terms of big data which has been used to perform several activities and been used to produce content on different social networks like Facebook, Twitter, YouTube, and so on. Even educational organizations maintain their course materials in form of webinars which is a multimedia data to provide efficient education to the students.

The challenge here in ViolationDetection of Live Video Based on Deep Learningis, the detection and retrieval of any specific video from the huge volume of multimedia data as the number of videos and volume of data increases every day [2,3]. There are millions of videos being uploaded each day which challenges the data retrieval process. Mobile users spend a lot of time on their mobile phones and involve in searching for different videos and content related to their interests. As the volume of data is higher and to identify the required video for the user query, there is a huge requirement for strategic approaches [4]. The conventional methods in AnoVid: A Deep Neural Network-based Tool for Video Annotationare not suitable for the retrieval and tracking of specific information from huge big data. On the other side, apart from retrieval accuracy, the time complexity makes big difference. The time complexity should be reduced which supports the mining and retrieval process highly. The method should be capable of mining required content from the big data set in the most effective way to produce meaningful results for the user [5,6].

Video surveillance is the process of monitoring and tracking the object or person located in different videos. A video is a sequence of frames combined to produce a moving picture. From the frames of the video, the object present should be identified and classified to perform object tracking. Video surveillance is used for several purposes. For example, consider a missing case where a person is missed. To track and identify the person, video surveillance can be used. To perform this, the object or the person's face feature is extracted. Further, with the extracted feature from a single frame, the presence of the same feature or face in the different videos can be identified. This process of tracking the face has been named video surveillance. So, by monitoring and tracking the face object in the videotapes, human tracking can be performed. It is not necessary that the tracking of a person should be performed using just the face feature but can be performed with the help of several features. Video surveillance has several functional components which include frame generation, background elimination, object detection, tracking, and several others [7,8].

2 Related Work

Support Vector Machines are used in several scientific problems and the same can be adapted to the problem of video annotation. The method would extract the features from the video frames and estimate support measures to perform video annotation and classification. The end-to-end NCNN-based model is designed to recognize human actions from still images [9]. The model trains the features using Nonsequential Convolutional Neural Network (NCNN) which reduces the time complexity. It overcomes the issues with a small dataset. The Deep Ensemble Learning Based on Voting Strategy (DELVS) model combines different classifiers to produce a better prediction. Only human actions were predicted from still images in this model. The background was not concentrated and motion images

were not analyzed. Image retrieval techniques based on Content-Based Image Retrieval (CBIR) and Text-Based Image Retrieval (TBIR) bridge the issue of the semantic gap between low-level and high-level image features [10].

The modern educational society maintains several video lectures which have been accessed by different students. To produce an optimal result for the student an SVM-based video classification algorithm has been presented [11]. The method classifies the video into various classes like a blackboard, talk, presentation, and mixed. Similarly, a semi-automatic video analysis scheme is presented that uses active learning to perform clustering [12]. A lazy learning-based approach is presented using the parent window method to build a pair-wise relationship, and based on that the classification is performed [13]. Similarly, an end-to-end Deep Ensemble Based Weighted Optimization (DELWO) developed extracts deep information and generates ensembles according to the weighted coefficients toward video classification [14]. Similarly, Deep Spatio-temporal predictive work is quite interesting and investigated by the research community with the involvement of network traffic analysis [15–17].

The neural network has been adapted to the classification of different categorical videos. In these approaches, the videos are generated or grabbed for video frames and the features from the frames are extracted and trained in several layers according to the number of categories. A text localization-based video classification approach is presented which uses a neural network to train the features like texture, size, location, and alignment [18]. According to the features extracted, the method classifies the data according to the textual features of the neural network. Similarly, a background subtraction algorithm is presented in video classification for unseen videos, which eliminates the background according to the background features obtained from the captured frame at different time frames to generate segmentation maps [19]. A human activity recognition scheme that considers the shape of objects towards classification. The method recognizes multi-agent events according to the Bayesian classification with fuzzy rules [20]. The method considers the temporal features and applies them to the classification.

The video contained in the database would represent any concept. So, the videos can be grouped according to the concept which supports the classification and retrieval. The problem of crowd behavior analysis towards video surveillance identifies the anomalies present in the video and annotates the video according to semantic ontology to provide a concept-based search framework [21] is studied by the researchers [22]. A multi-video summarization model [23,24] presented a framework that extracts keyframes from related videos according to the user query and inference from ontology. An ontology video genre identification scheme according to the video specific to gender.

A coherent picture-based video recognition algorithm uses five different coherent pictures and with the frames extracted from the video, the method extracts a set of features and classify using LSTM. The method has been evaluated using a shipping data set [25]. Deep-learning Semantic-based Scene-segmentation model (called DeepSSS) that considers image captioning to segment a video into scenes semantically. First, the DeepSSS performs shot boundary detection by comparing color histograms and then employs maximum-entropy-applied keyframe extraction. Second, for semantic analysis, using image captioning that benefits from deep learning generates a semantic text description of the keyframes. Finally, by comparing and analyzing the generated texts, it assembles the keyframes into a scene grouped under a semantic narrative [26].

A detailed survey on video analytics explores various methods related to the problem. A collaborative scheme of video annotation around semantics that uses semantic ontology and consensus-based social network analysis (SNA). The method enables the sharing of content in social networks between

friends and enables ontology-orient search which describes the content semantically. A video structure-based two-layer annotation scheme extracts the structure of different objects and context details from video clips [27]. According to the details, the method combines the title and text information in video retrieval and ranking the video clips. A tensor flow-based video structure analysis scheme is presented which analyzes the structure to identify the shot boundaries according to the image duplication technique. The method identifies a set of objects and extracts their features to measure the semantic weight to perform analysis and retrieval [28].

The initial stage of video surveillance is performing frame generation and preprocessing the frame to extract the required feature. Once the objects of different frames are extracted, then object tracking can be performed after identifying the background features. By tracking the objects, different problems can be handled effectively. The problem of video surveillance has been approached with several techniques like template matching, color matching, shape matching, and so on. However, surveillance can be performed by including several features towards improving the quality of the system.

The growth of information technology has been used in several applications. The crime investigation is the area that applies such technology support at the maximum. The changing social environment introduces different threats to society. Several crimes occur every day in different parts of the country. However, the investigating departments and private sectors have launched their monitoring systems which record the videos through CCTV cameras. Such videos being tapped are monitored by different security professionals. In any application, there will be human error and the human would miss a set of events or crimes in which it is difficult to identify the accused.

Video analytics is the process of analyzing the video for the presence of a specific object or event. The video would contain several objects and events; however, by analyzing the video frames obtained from the specific video, the presence of the object can be tracked and identified. In some situations, the crime would be captured in the videotapes and would help to solve the case. But in many situations, the accused cannot be identified at the first instance and the video being recorded would not be clear. The quality of the videotape is depending on the lighting conditions, and the angle of the person or object that appears in the video. If the object or person does not capture exactly, it cannot be identified manually and needs to explore several consequent videos obtained from neighboring locations' CCTV. This also increases the human effort required and increases the time complexity.

To reduce such human effort and time complexity, automated systems can be used in this problem. The automated system would train the objects present in images and based on that the tracking of the objects can be performed. There are several approaches available in object tracking some of the methods use shape, texture, size, and so on. However, they suffer to achieve higher performance in tracking the objects. By automating the tracking problem, the person involved or the object present at the crime scene can be tracked for its presence in different video files and frames available in the data set.

The growing population and traffic challenge the administration of any city or country. Several crimes are happening in the cities, to solve the cases and to track such crimes and the persons involved, it is necessary to monitor the visuals captured from different locations. In terms of investigation and criminology, the model and the way the crime is carried out is varying between criminals. They differ in the way they perform the crime. By analyzing the videotapes of different crimes, the investigating officer would get enough knowledge about the crime.

On the other side, video big data has been used in several problems. For example, it has been used in several applications like sports analysis, education improvement, and so on. The sportspeople get more information from past videos obtained from the database. For example, a cricket learner who

looking to learn about hitting a six over a square leg can be done by viewing a set of videos. In this case, the videos related to hitting six in the square leg side can be retrieved only when such videos are grouped tentatively. Otherwise, identifying such videos with exact information cannot be retrieved.

To perform video retrieval and classify a video into several categories, there are several approaches available. Some of the methods use shape features of objects and perform object tracking to identify the class of video, whereas some other methods use texture and other features in identifying the class of the video. But most methods suffer to achieve higher performance video classification and retrieval.

The issues identified in this study are summarized as 1. The semantic link network model works efficiently for images only. 2. Current video surveillance system is subject to human error and it processes video files of small size only. 3. Need to concentrate on the reduction of computational time as well as the increase in recognition for various irregular shape objects from videos. 4. The accuracy level towards recognizing with limited trained objects as reference need to be improved.

3 Video Semantic Substance Extraction Framework

The growth of engineering and technology has been adapted to several purposes in the real world. As the trust among humans is getting reduced, it is necessary for organizations in monitoring whatever happens within their premises. Such activity is named surveillance which is earlier performed by engaging humans and is very much scalable. To handle this issue, video surveillance has emerged where a dedicated system has been deployed with a set of video cameras, and the cameras deployed are capturing the video and display it in a particular control section. The administrator can monitor the activities of different employees located and working within the premises. Not just that, surveillance has been applied to several problems like traffic monitoring.

On the other way, searching for an event from a group of videos is a challenging task. Manual searching is time-consuming and also it is error-prone. For example, event detection and behavior monitoring has been handled with the support of video surveillance. In general, video surveillance has been performed by extracting the objects from the images or frames of video. The video contains several frames and the number of frames varies according to the quality of the video. Also, the frame quality differs according to the video quality. By extracting features like shapes, objects, and colors from the video frames, the problem of searching for an event for video surveillance can be handled.

To improve the performance of video classification and retrieval, it is necessary to consider a variety of features other than low-level features. For example, the features like shape, speed, size, the direction can be used in the classification of video. Similarly, by considering the semantic features, concepts, and events, video classification can be performed effectively. This chapter discusses such an approach toward video classification and retrieval.

3.1 Fuzzy Systems in Video Classification

Big data contains several videos that belong to different aspects and events and concepts. However, the properties of objects in the videos are different and vary according to the event and concepts. To classify the video and to identify a set of related videos, the classification approaches would consider several features like shape, texture, size, speed, and direction. The value of such features would vary according to the event. For example, the size of the object would be increasing when the object is moving towards the camera whereas the size of the object will be small when the object is leaving. Also, for an object which is moving towards the camera, the direction will be positive whereas, for a leaving object, the direction value will be negative. So, the value of different features, the value of

them will differ according to the event present in the video scene. By considering such fuzzy values, the classification of events can be performed.

The events and their fuzzy values on various features are presented in Table 1. For each case of fuzzy features, the values will be varying. For example, in the case of shape, the shape of the object will be within a fuzzy range which is measured according to the coordinate points of the shape of objects.

Table 1: Fuzzy values towards different events

Event	Shape	Size	Directions	Speed
Moving car	Vary	Decreasing	Negative	Increasing
Throwing ball	Same	Decreasing	Negative	Increasing
Catching ball	Same	Increasing	Positive	Increasing
Sweep shot	Vary	Decreasing	Neutral	Neutral

3.2 Fuzzy Rule-Based Video Semantic Analysis Model

The fuzzy rule-based video semantic analysis model classifies the videos according to different semantic features. First, the method generates the video frames and preprocesses the image by applying the adaptive median filtering technique. Using the set of median-filtered images, the method detects the background and eliminates the background from the video frames. Further, the object features are detected and extracted. The features from the detected objects are extracted. Using the features extracted, the method performs classification with fuzzy rules. The features extracted are trained with a neural network where each network contains several layers among them each layer of neurons is dedicated to measuring the weight towards different semantic events. Each neuron measures its weight according to different features like shape, size, direction, speed, and other features. The weight measurement is performed according to the fuzzy rules and based on the weight measures; the classification is performed. The detailed approach is presented in this section.

Fig. 1, shows the architecture of the proposed fuzzy rule-based video semantic analysis model and the functional components involved in the model. Each functional component has been detailed below.

3.3 Video Frame Generation

The video clips given with the data set have been read. Each video clip is considered as containing an event. To identify this, the method generates several video frames according to human visual perception. In general, nearly 18 numbers of video frames are generated from the video. Such video frames generated are used for event classification and semantic content extraction.

3.4 Preprocessing

Preprocessing is the process of eliminating the noisy features from the video frames and improving the quality of the video frame to support classification. To perform this, the method reads the image and initializes the adaptive median filter which finds the boundary values according to the mean value of neighboring pixels. Each pixel has been normalized according to the neighbor pixels and median values.

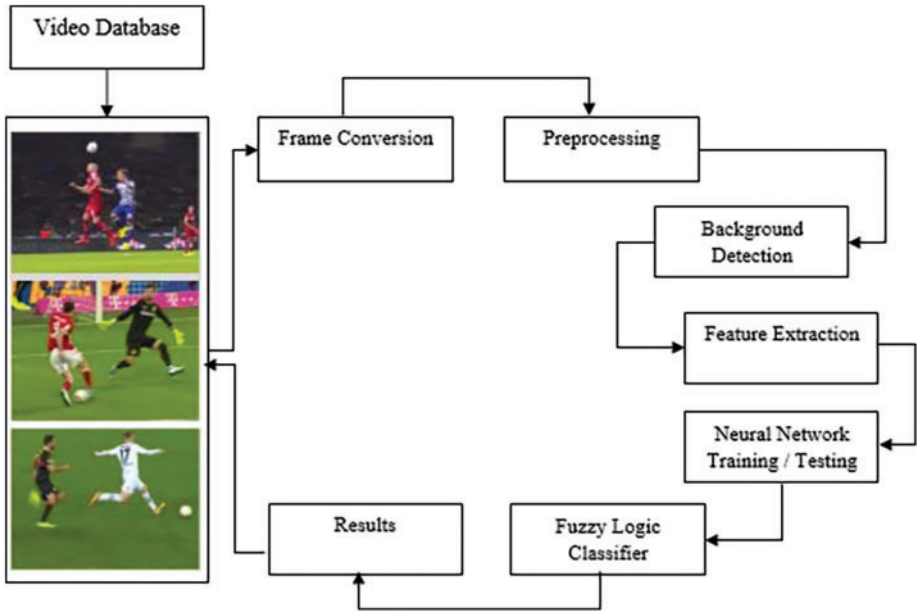

Figure 1: Architecture of fuzzy rule-based video semantic analysis model

Consider the original image is sized as 4×4 as follows:

100 100 100 100
100 100 100 100
100 100 100 100
100 100 100 100

then if the corrupted image is as follows:

100 255 100 100
100 255 100 100
225 100 100 100
100 100 100 100

then to apply the kernel filter with a 3×3 median filter should have 1 column of zero on the left and right edges. Also, the kernel should be padded with 1 row of zeros on the top and bottom edges, which produces the result as follows:

0 0 0 0 0 0
0 100 255 100 100 0
0 100 255 100 100 0
0 255 100 100 100 0
0 100 100 100 100 0
0 0 0 0 0 0

According to the values of the top window, the value of the mean has been measured as: (0, 0, 0, 0, 0, 100, 100, 255, 255) which shows the sorted value of the window, and the median value is obtained as 0 which is replaced on the value of 100 at the first column. Similarly, each spot has been applied with the filter to produce the resultant image as follows:

0 100 100 0
100 100 100 100
0 100 100 0
100 100 100 100

The above procedure has been continuously applied with the original image in different window sizes. This produces tampering on the given image. To overcome this deficiency, the padding is performed by the proposed approach by selecting the median value of the first row and column. This allows the padding to be done as follows:

[100, 255, 100, 100, 100, 255, 100] = 100 which selects the value 100 to be padded with the corrupted image as follows:

100 100 100 100 100 100
100 100 255 100 100 100
100 100 255 100 100 100
100 255 100 100 100 100
100 100 100 100 100 100
100 100 100 100 100 100

By applying this way, the method would produce the result as follows:

100 100 100 100
100 100 100 100
100 100 100 100
100 100 100 100

This shows the proposed filtering technique has removed the distortion and improved the quality of the image with higher performance. The proposed preprocessing technique overrides the existing median filter by eliminating the distortion generated by the approach. Read the video frame V_f and generate Filter coefficient F_c using the below equation

$$\text{Filter coefficients } F_c = \sum Vf\ (i) \in \{row \underset{i=1}{\overset{size(Vf)}{=}} 1,\ \ Col = 1\} \tag{1}$$

Then for each window w, apply adaptive median filtering

$$V_f\ (p_i) = \text{AMF}\ (V_f\ (p_i), F_c) \tag{2}$$

Thus, the adaptive median filtering is applied over the video frames and the quality of the image has been improved to support classification. This adaptive median filtering scheme eliminates the noise from the images and applies the histogram equalization technique.

3.5 Background Detection

In this stage, the method first detects the background. Towards this, a Gaussian model is used, which uses probability density. It is necessary to generate the background model to eliminate the background from different video frames.

For instance, contrast, correlation, homogeneity, energy, shape, area, perimeter, filled area, and eccentricity. For the division fuzzy logic is actualized. To perform this, first, the images are read and a pixel-to-pixel comparison is performed. It has been measured according to the likelihood value of different pixels. The likelihood value of the different pixels is measured as follows:

$$P_r(X_i) = \sum_{i=1}^{K} M_I, X_t.C_i, E_{ri}, H \sum i, t \tag{3}$$

defaults $K = 3$, C is the compactness parameter of the K^{th} Gaussian model, M is a likelihood thickness capacity, E is the energy that indicates the pixel power value, H signifies the homogeneity and $\sum i, t$, is the covariance matrix of the Kth distribution.

According to this, the compactness value is measured as follows:

$$C_i = \frac{A}{p2} \tag{4}$$

where p is the perimeter and A is the area.

3.6 Feature Extraction

The video provided would have several features. However, to annotate the video database to provide effective results to the user. To provide better results to the user or to identify the class of video clip provided, it is necessary to consider different features like low, high and semantic. There are several approaches available to classify the video which uses different features like shape, texture, color, and so on. However, these algorithms consider different features like shape, size, direction, and speed with the rest of the features. To perform this, the method uses the background generated with the Gaussian mixture model. With the background model, the method measures the absolute difference between the input and background image. According to the difference, the method detects a set of moving objects. From the objects being identified, the method extracts the features of objects. Similarly, the features of the moving object are extracted from different video frames. Such features extracted are used to perform classification. The features were extracted using the below equations.

$$\text{Shape feature } S_{hf} = \frac{\sum_{i=1}^{size(Obs)} \frac{\sum_{j=1}^{size(Obs(i))} Dist(Obs(i(j)), \forall Obs(i))}{size(Obs(i))}}{size\,(Obs)} \tag{5}$$

$$\text{Size feature } S_{if} = \frac{\sum_{i=1}^{size(Obs)} \sum pixel \in (Obs\,(i))}{size\,(Obs)} \tag{6}$$

$$\text{Speed feature } S_{pf} = \frac{\frac{\sum_{i=1}^{size(Obs)} \sum dist(loc(obs(i) \to vf\,(i-1)), loc(obs(i) \to vf\,(i)))}{size(Obs)}}{Time} \tag{7}$$

$$\text{Direction feature } D_f = \frac{\sum_{i=1}^{size(Obs)} \sum dist\,(loc\,(obs\,(i) \to vf\,(i)), loc\,(refobject))}{size\,(Obs)} \times S_{pf} \tag{8}$$

Feature extraction is used to analyze the motion feature of video frames for activity recognition. It first detects the feature from the frame and then extracts it from it. It was presented using different image algebraic operations. *Feature extraction* starts from processing, in which algorithms are used to detect and isolate various desired portions or shapes of a digitized image or *video* stream. The method extracts shape, size, direction, and speed features that are used to perform event classification and video classification.

Table 2 presented above shows the fuzzy value of any event being identified from the videotapes. The rule is generated according to several features like shape, size, speed, and direction. According to the fuzzy rule generated for different semantic classes, the neurons estimate the semantic weight for the given input feature towards the semantic class. The final layer produces several weights and according to the waits; the mean value of the semantic weight is generated towards various semantic

classes. Based on the semantic weight, the class with the maximum semantic weight is selected as a result of classification.

Table 2: Fuzzy value of any event

	Shape	Size	Speed	Direction
Minimum	X	M	U	P
Maximum	Y	N	V	Q

3.7 Fuzzy Rule-Based Classification

The classification of video clips given to the framework is performed in this stage. The data set given has been read and the features of each clip are extracted. Using the features extracted, the method trains the network where each semantic concept has been organized with a set of neurons where each neuron has been initialized with the specific feature set. Similarly, there will be several layers according to the semantic concept available. The neurons are capable of computing the semantic weight measure according to the fuzzy rule available. The neurons of the layer generate the fuzzy value according to the feature set of the concept or semantic class. The fuzzy value is generated according to the features of each video clip available and extracted. The method computes the minimum and maximum values on each class feature to generate the fuzzy value.

Pseudocode of fuzzy rule-based classification

Given: video data set V_{ds}, Testing Video Clip V_c
Obtain: Semantic Class
Start
    Read V_c and V_{ds}.
    For each video clip
        Frameset = Frame generation (V_c)
        For each frame f
            Pi = Preprocessing (f)
        End
        Object set O_s = Feature extraction (frame sets)
    End
    Initialize neural network NN.
    Train neural network.
    For each semantic class sc
        Compute semantic weight $= \frac{\sum_{i=1}^{size(Rule)} Obs\,(i) \;<\; Rule\,(i)\,.\min \&\&\; Rule\,(i)\,max\,>}{size\,(Rule)}$ (9)
    End
    Class C = Choose the semantic class with maximum semantic weight.
    Stop

The above-discussed pseudo-code shows how video classification is performed towards annotation and retrieval of video from big data. The method trains the network with different features and estimates the semantic weight according to the fuzzy rule. Based on the semantic weight, the methods perform the classification of the video.

4 Results and Discussion

The proposed multi-component tracking mechanism is actualized in MATLAB 7.11.0 (R2010b) with an i5 processor and 4 GB RAM. The datasets utilized for testing the tracking system are football, auto, and barbellcolor video arrangements with a minimum 2 GB size. The datasets are exceptionally challenging due to the overwhelming inter-person objects and poor picture differentiation amongst components and foundations. The algorithm was accessed on its tracking execution and it was noticed that the detection execution contrasted our outcomes and existing strategy. Multi-component tracking for the most part confronts three difficulties: component switch among overlapping, new component introduction, and re-acknowledgment of re-entering objects. The accompanying part quickly presents two videos and after that discusses the outcomes as far as previously mentioned challenges. The tracking outcome of football, car, barbell has been illustrated in Fig. 2.

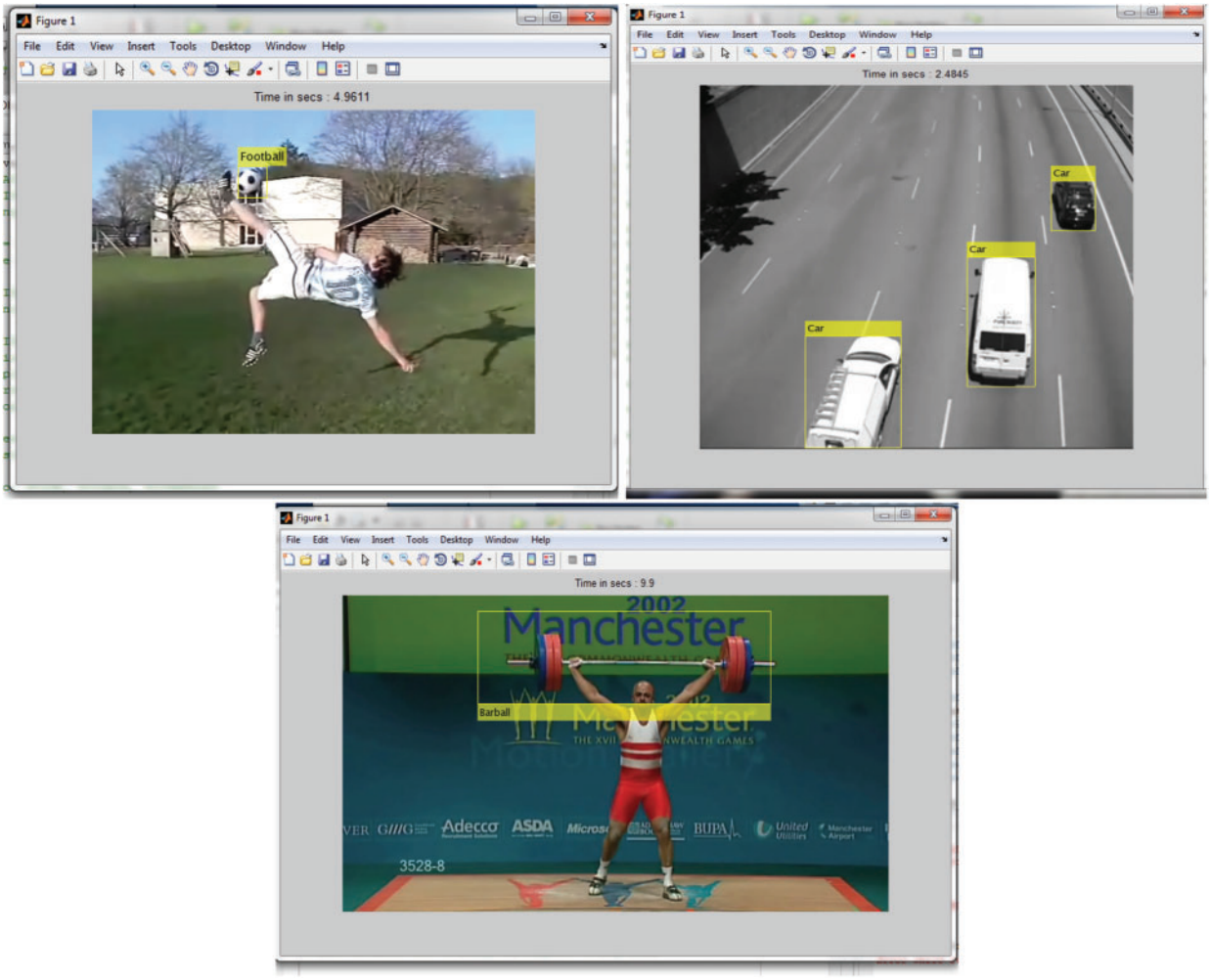

Figure 2: The tracking outcome on football, car, barbell

Give s_n^i be the distance between the assessed outcome and the ground truth for component i at time n and b_n the number of matches discovered, further then, MOTP was represented as:

$$MOTP = \frac{\sum_{i,n} s_n^i}{\sum_n b_n} \tag{10}$$

The distance d_n^i is the covering between the evaluated bounding box and the ground truth. Subsequently, higher estimations of MOTP demonstrate better outcomes. For the MOTA, let w_n be the number of items that exist at moment n. Let likewise k_n, ht_n, and kke_n be the number of misses, false positives, and jumbles, separately. At that point, the metric can be gotten by

$$MOTA = 1 - \frac{\sum_n (k_n + ht_n + kke_n)}{\sum_n w_n} \tag{11}$$

The precision metric correlation using initiated fuzzy and existing neural systems is represented in Table 3. The precision value acquired for the car picture utilizing the existing technique is 0.897237 while initiated has 0.945792 comparatively for Football, Barbell pictures precision value computed utilizing the existing strategy is 0.916324, 0.907507 yet initiated has 0.975404, 0.964889 individually. It can be clear from the above discussion that, the initiated technique performs powerfully and it can deliver outcomes that are near the ground truth, regardless of the identities that the components are swapped.

Table 3: Performance measures of MOTP

	Precision	
Dataset	Neural	Fuzzy
Car	0.897237	0.945792
Football	0.916324	0.975404
Barbell	0.907507	0.964889

The Accuracy metric comparison using initiated fuzzy and existing neural systems is represented in Table 4. The accuracy value got for the car picture utilizing the existing strategy is 0.912699 while initiated has 0.963376 likewise for Football, and Barbell pictures Accuracy value figured utilizing the existing technique is 0.92785, 0.915761 but initiated has 0.954688, 0.970593 respectively. Our initiated technique is better in assigning tracks to the right component, without taking into account how near it is from the right position.

Table 4: Performance measures of MOTA

	Accuracy	
Dataset	Neural	Fuzzy
Car	0.912699	0.963376
Football	0.92785	0.954688
Barbell	0.915761	0.970593

The underneath graph demonstrates the precision comparison of the dataset utilizing initiated fuzzy algorithm with existing neural system procedure for multi-component tracking. From the graph

precision value of the initiated technique is more contrasted with the existing strategy. So, the quantity of significant objects tracked using the initiated technique is better and is presented in Fig. 3.

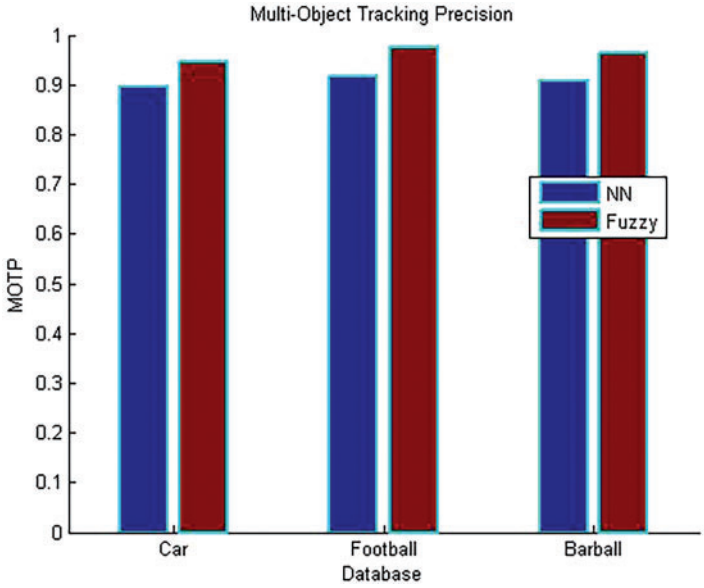

Figure 3: Performance graph of precision

The graph demonstrates the accuracy comparison of the dataset utilizing initiated fuzzy algorithm with the existing neural system method for multi-component tracking. Our initiated strategy has higher values contrasted with the existing technique is presented in Fig. 4.

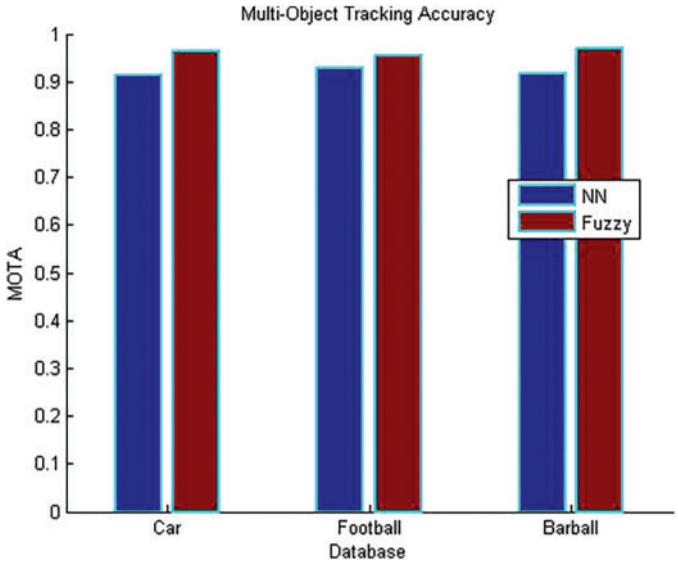

Figure 4: Performance graph of precision

The false ratio introduced by different methods in tracking the object is measured towards various object classes. The proposed method introduced the false ratio in the ratio of 1.8, 2.5, and 1.8 towards

the Car, Football, and Barbell object classes which is less than other neural-based approaches is presented in Table 5.

Table 5: False ratio in object tracking with FRSCE

Dataset	Neural	Fuzzy
Car	8.8	1.8
Football	7.3	2.5
Barbell	8.5	1.8

The proposed method introduced the false ratio in the ratio of 1.8, 2.5, and 1.8 towards the Car, Football, and Barbell object classes which is less than other neural-based approaches due to the lesser number of False positive and False negative samples. The false ratio introduced by the methods towards multiple object tracking is measured towards various object classes and presented in above Fig. 5. The proposed FRSCE has produced less false ratio than neural methods.

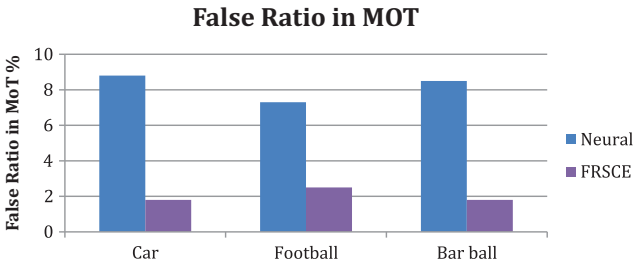

Figure 5: False classification ratio in MOT with fuzzy

5 Conclusion

The Method considers different features of objects towards video annotation and classification. The approach read the input video data set and for each video, the frames are generated. From the frameset, the method applied the Gaussian model to generate the background model and eliminate the background and identifies the moving objects. According to the objects identified, their features are extracted. According to the features of objects that belong to the same event and semantic class, the method generates the fuzzy rule. The features extracted are trained with a neural network and tested with the features of a single clip at the test phase. The neurons estimate semantic weight towards various events and semantic classes based on the fuzzy rule and values of the features extracted from the object. The method selects a single class according to the semantic weight. The proposed method improves the performance of classification and tracking accuracy. In the future, the performance of multiple object tracking and semantic content extraction in video surveillance can be improved by tracking the moving objects, and by computing moderate appearance factors in different video clips, the performance of object tracking can be improved. This would support the movement tracking of different objects or persons in various video clips obtained from various camera devices to support crime investigation.

Funding Statement: This work was supported in part by the Higher Education Sprout Project from the Ministry of Education (MOE) and National Science and Technology Council, Taiwan, [109-2628-E-224-001-MY3], [111-2622-E-224-009] and in part by Isuzu Optics Corporation.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Wang, Y. Yang, T. Wang, R. S. Sherratt and J. Zhang, "Big data service architecture: A survey," *Internet Technology*, vol. 21, no. 2, pp. 393–405, 2020.
[2] C. Yuan and J. Zhang, "Violation detection of live video based on deep learning," *Scientific Programming*, vol. 2020, Article ID 1895341, 2020.
[3] P. E. I. Xiao Gen, "The key frame extraction algorithm based on the indigenous disturbance variation difference video," *Procedia Computer Science*, vol. 183, pp. 533–544, 2021.
[4] X. Zhaopeng, L. Jiarui, L. Wei, X. Bozhi, Z. Xianfeng *et al.,* "Detecting facial manipulated videos based on set convolutional neural networks," *Visual Communication and Image Representation*, vol. 77, Article ID 103119, 2021.
[5] J. Hwang and I. Kim, "AnoVid: A deep neural network-based tool for video annotation," *Korea Multimedia Society*, vol. 23, no. 8, pp. 986–1005, 2020.
[6] T. Meraj, A. Hassan, S. Zahoor, S. Rauf, H. T. Lali *et al.,* "Lungs Nodule detection using semantic segmentation and classification with optimal features," Preprints 2019, 2019.
[7] X. Yu, Z. Zhang, L. Wu, W. Pang, H. Chen *et al.,* "Deep ensemble learning for human action recognition in still images," *Complexity*, vol. 2020, Article ID 9428612, pp. 23, 2020.
[8] A. Manju and P. Valarmathie,"Video analytics for semantic substance extraction using OpenCV in python," *Ambient Intell Human Comput*., vol. 12, pp. 4057–4066, 2021.
[9] S. L. Stephens, M. Manuguerra and M. W. Bulbert, "Seeing is relieving: Effects of serious storytelling with images on interview performance anxiety," *Multimed Tools Appl.*, vol. 81, no. 16, pp. 23399–23420, 2022.
[10] Z. Shou, D. Wang and S. F. Chang, "Temporal Action Localization in untrimed videos via multistage CNNs," in *Proc. of IEEE Conf. on Computer vision and Pattern Recognition*, pp. 1049–1058, 2016.
[11] D. Rohidin, A. Noor, A. Samsudin and M. Mat Deris, "Association rules of fuzzy soft set-based classification for text classification problem," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, pp. 801–812, 2020.
[12] B. Chen, T. Jiang and L. Chen, "Weblog fuzzy clustering algorithm based on convolutional neural network," *Microprocessors and Microsystems*, vol. 103420, 2020.
[13] K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data," *Journal of Big Data*, vol. 6, Article ID 91, 2019.
[14] B. Ko, J. Hong and J. Y. Nam "Human action recognition in still images using action poselets and a two-layer classification model," *Journal of Visual Languages & Computing*, vol. 28, pp. 163–175, 2015.
[15] J. Wang, H. Han, H. Li, S. He, P. K. Sharma *et al.,*"Multiple strategies differential privacy on sparse tensor factorization for network traffic analysis in 5 G," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 1939–1948, 2022.
[16] J. Wang, C. Jin, Q. Tang, N. Xiong and G. Srivastava, "Intelligent ubiquitous network accessibility for wireless-powered MEC in UAV-assisted B5G," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 4, pp. 2801–2813, 2021.
[17] D. Cao, K. Zeng, J. Wang, P. Kumar Sharma, X. Ma *et al.,* "BERT-based deep spatial-temporal network for taxi demand prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, pp. 9442–9454, 2021.
[18] Z. Xu, J. Liu, W. Lu, B. Xu, X. Zhao *et al.,* "Detecting facial manipulated videos based on set convolutional neural networks," *Visual Communication and Image Representation*, vol. 77, pp. 103119, 2021.

[19] M. O. Tezcan, P. Ishwar and J. Konrad, "BSUV-Net: A fully-convolutional neural network for background subtraction of unseen videos," in *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, Snowmass Village, CO, USA, pp. 2763–2772, 2020.
[20] W. Xu, Z. Miao, J. Yu and Q. Ji, "Action recognition and localization with spatial and temporal contexts," *Neurocomputing*, vol. 333, pp. 351–363, 2019.
[21] E. Hatirnaz, M. Sah and C. Direkoglu, "A novel framework and concept-based semantic search interface for abnormal crowd behavior analysis in surveillance videos," *Multimedia Tools Appl*., vol. 79, pp. 17579–17617, 2020.
[22] H. Ji, D. Hooshyar, K. Kim and H. Lim, "A semantic-based video scene segmentation using a deep neural network," *Information Science*, vol. 45, no. 6, pp. 833–844, 2019.
[23] S. Mu, "Object driven semantic multi-video summarization based on ontology," in *Proc. of the Second Int. Conf. on Data Science, E-Learning and Information Systems (DATA '19)*, Association for Computing Machinery, New York, NY, USA, Article No. 6, pp. 1–6, 2019.
[24] G. Yao, T. Lei and J. Zhong, "A review of convolutional neural network based action recognition," *Pattern Recognition Letters*, vol. 118, pp. 14–22, 2019.
[25] H. Zhang, "Surveillance videos classification based on multilayer long short-term memory networks," *Multimedia Tools and Applications*, vol. 79, no. 17–18, pp. 12125–12137, 2020.
[26] M. Bouchakwa, Y. Ayadi and I. Amous, "A review on visual content-based and users tag-based image annotation: Methods and techniques," *Multimedia Tools and applications*, vol. 79, pp. 21679–21741, 2020.
[27] S. S. Aote and A. Potnurwar, "An automatic video annotation framework based on two level keyframe extraction mechanism," *Multimedia Tools and Applications*, vol. 78, pp. 14465–14484, 2019.
[28] Z. Huang, B. Sui, J. Wen and G. Jiang, "An intelligent ship image/video detection and classification method with improved regressive deep convolutional neural network," *Complexity*, vol. 2020, Article ID 1520872, pp. 11, 2020.