electronics

Article

A Player-Specific Framework for Cricket Highlights Generation Using Deep Convolutional Neural Networks

Rabbia Mahum [1], Aun Irtaza [1], Saeed Ur Rehman [2,*], Talha Meraj [2] and Hafiz Tayyab Rauf [3,*]

[1] Department of Computer Science, University of Engineering and Technology Taxila, Taxila 47080, Pakistan
[2] Department of Computer Science, COMSATS University Islamabad, Wah Campus, Islamabad 47040, Pakistan
[3] Centre for Smart Systems, AI and Cybersecurity, Staffordshire University, Stoke-on-Trent ST4 2DE, UK
* Correspondence: srehman@ciitwah.edu.pk (S.U.R.); hafiztayyabrauf093@gmail.com (H.T.R.)

Abstract: Automatic ways to generate video summarization is a key technique to manage huge video content nowadays. The aim of video summaries is to provide important information in less time to viewers. There exist some techniques for video summarization in the cricket domain, however, to the best of our knowledge our proposed model is the first one to deal with specific player summaries in cricket videos successfully. In this study, we provide a novel framework and a valuable technique for cricket video summarization and classification. For video summary specific to the player, the proposed technique exploits the fact i.e., presence of Score Caption (SC) in frames. In the first stage, optical character recognition (OCR) is applied to extract text summary from SC to find all frames of the specific player such as the Start Frame (SF) to the Last Frame (LF). In the second stage, various frames of cricket videos are used in the supervised AlexNet classifier for training along with class labels such as positive and negative for binary classification. A pre-trained network is trained for binary classification of those frames which are attained from the first phase exhibiting the performance of a specific player along with some additional scenes. In the third phase, the person identification technique is employed to recognize frames containing the specific player. Then, frames are cropped and SIFT features are extracted from identified person to further cluster these frames using the fuzzy c-means clustering method. The reason behind the third phase is to further optimize the video summaries as the frames attained in the second stage included the partner player's frame as well. The proposed framework successfully utilizes the cricket videoo dataset. Additionally, the technique is very efficient and useful in broadcasting cricket video highlights of a specific player. The experimental results signify that our proposed method surpasses the previously stated results, improving the overall accuracy of up to 95%.

Keywords: deep learning; feature extraction; classification; fuzzy c-means; clustering

check for updates

Citation: Mahum, R.; Irtaza, A.; Rehman, S.U.; Meraj, T.; Rauf, H.T. A Player-Specific Framework for Cricket Highlights Generation Using Deep Convolutional Neural Networks. *Electronics* **2023**, *12*, 65. https://doi.org/10.3390/electronics12010065

Academic Editor: Yue Wu

Received: 1 December 2022
Revised: 16 December 2022
Accepted: 20 December 2022
Published: 24 December 2022

1. Introduction

Digital videos have become pervasive, therefore the proficient approach of pulling out the information from the video becomes gradually more important [1]. Videos contain an enormous amount of data and intricacy that makes the analysis more challenging. In the past few years, various techniques have been presented for video summarization to solve the problem such as extracting information from lengthy videos. Video summarization has pragmatic in numerous domains inclusive of sports [2,3], medical [4,5], web [6], and surveillance system [7,8]. In recent years there have been more efforts in automated video summarization of sports highlight generation [9,10].

Manual highlight generation specific to player is tedious work for long-duration videos for example cricket videos. Moreover, a massive collection of cricket videos are broadcasted, including many additional events that lack viewers' interest. Therefore, automatic cricket video summarization specific to player can be performed via machine learning algorithms, which aims to automatically assign a label to a specific shot to determine whether that segment is specific to player and interesting to viewers or not. Most video summarization techniques are based only on visual features such as crowd, audience noise, and pitch view and do not focus on individual player's performance.

Various works have been proposed by researchers focusing on cricket video summarization, however, if any research has provided player-specific summary then it is based on a textual log only [11]. Therefore, the results are not considerable on unseen videos due to avoidance of visual features. In [12], original cricket event video is divided into shots and then shots having significant events are included in the summary. In [13], cricket highlights generation has been done based on low-rank visual features such as color histogram (CH). The replay detection technique is used to generate highlights in sports videos [14], which take advantage of the gradual transition effect at the beginning and end of the re-play segment and lack of score caption in the re-play shot. Video summaries are generated in [15], by applying text processing on the shots of the speech through speech recognition but completely avoiding visual features. The rank-based algorithm is implemented to generate a video summary that utilizes quality, user attention, representativeness, temporal coherence, and uniformity to select key frames [16]. Additionally, audio-based techniques have also been exploited to generate highlight summaries. However, these are unable to use important visual events. Various existing techniques have been suggested for cricket video summarization, however the generated summaries still include unnecessary frames and they are generalized.

Therefore, to overcome the above-mentioned challenges we propose a hybrid model for the cricket highlights generation specific to player. In first stage, we used OCR which focuses on log information to find player specific frames. Then, we trained AlexNet, a pre-trained model for the classification of frames to discard unnecessary ones. In third step, we optimize the summarized video employing a person detection algorithm to further cluster the frames using fuzzy c-means clustering based on SIFT descriptor. The proposed system is a successful execution for player specific summary generation for cricket videos.

The main contributions of this study are as below:

- A novel technique to generate player-specific summaries of cricket videos is introduced using the Deep Neural Network along with person detection and other machine learning operations.
- Dataset for training and testing is used from 4 famous broadcasters such as Sky Sports, Fox Sports, ESPN, and Ten Sports. The benefit of using a Deep Neural Network is that a small dataset is required for training purposes. Therefore, we utilized minimal frames while attaining the significant results such as accuracy of 95%.
- The technique can be generalized to generate video summaries for other sports games as well i.e., soccer, football, tennis, etc.
- To the best of our knowledge, our proposed model is first one to consider the generation of video summaries specific to player in cricket videos.

This work presented in this paper is organized as: Section 2 describes the proposed technique, while results and discussion are presented in Section 3. At last, the conclusion is given in Section 4.

2. Related Work

Various video summarization techniques have been researched in the literature [17,18]. Generally, most of the methods are using visual features at the initial phase and then some threshold criteria to include frames in the final output summarized video. keyframes-based techniques and object tracing give better information, are proposed in [19–21], in which the subset of frames are selected for summarized video, while Story-board is used in [22]. Skimmed video methods divide the original video into short clips and then specific shorts are selected based on some criteria for video summarization [23,24]. Most of the above techniques are based on unsupervised algorithms. Some research-ers have used supervised algorithms such as to create regions through supervised learned metrics [25], or trained algorithms to generate canonical viewpoints [26]. In [27], convolutional neural network-based summarization technique is applied to surveillance videos. Shot segmentation is used on deep features, hence the highest image entropy and memorability are used as keyframe detectors to include in the video summary. In [28], unsupervised object-level summarizations are created through the auto-encoder technique based on segmented video clips.

Firstly, all moving objects in videos are detected through the Markov Decision Process (MDP), then the super-frame segmentation technique is used to cut video clips based on motion logs. In [29], video summaries are generated deep attentive mechanism with the application of GoogleNet to get visual features. Then, Bi-directional Long Short Term Memory (BiLSTM) is applied as an encoder in which input flows in both directions and the proposed attention mechanism is used. Later Long short term memory (LSTM) is used for decoding to predict the scores of frames. In the end regression and distribution [29], losses are used as final objective functions to add the keyframes in the output video.

In [11], textual commentary is used based on n-gram, to detect the specific events for cricket videos. The algorithm is tested against the commentary lines which give recall and precision approximately 90%. But the drawback of this technique is that it doesn't depend on visual features and relies only on a textual log. But in our proposed algorithm we are generating highlights based on text and visual features both and attain a result of 95% accuracy. Ali et al. [14] introduced the local octal pattern features for the representation of video shots. They trained the model using extreme learning to classify them into replay and non-replay classes. They produced the significant summaries, however, the proposed model was not verified for player specific summaries. Mahum et al. [6] developed a video summarization system for abnormal activities that may be utilized for videos using combined form of Zernike moments and R-transform. Moreover, the authors employed KNN and AlexNet for further classification of frames. The system attained valuable results representing key-events of original videos. In [3], the authors utilized transfer learning technique for the generation of summarized videos and verified results for lectures and surveillance datasets. Various deep learning algorithms have been proposed for different domains such as medical [30,31], agriculture [32], and surveillance systems, however, AlexNet has been proved one of the simplest deep learning algorithm. The purpose of [33] study was to extract both visual and audio features. Speech-to-text framework and audio energy were used to detect the excitement frames. The template learning process was used to locate the score box with the use of SIFT features. A deep learning model was used for the extraction of text from the score box.

To handle the limitations of existing techniques such as avoiding high-level visual features and text summary, camera variations, low network bandwidth, time constraints, computational complexity, etc. a computationally expert method is presented for automatic video summarization specific to the player. The proposed technique utilizes text summary from cricket video with the help of optical character recognition (OCR) to extract frames of a specific player and a pre-trained deep learning algorithm which is exploited to classify which frames should be included in the summary. Furthermore, fuzzy C-means (FCM) clustering aided with SIFT features and histogram is applied to remove extra frames. The classified frames which will be included in the summarized video represent the important events related to the specific player such as sixes, fours, out, etc. and hence reduce the redundant data. The presented technique is dynamic to camera variations, velocity, logo designs, etc. The pro-posed method is applied to publicly available datasets having three diverse categories from cricket videos. The results produced after experiments specify that the system attains the summarization accuracy approximately 94% using videos.

3. Methodology

The proposed technique is comprised of three main stages, (1) Frames identification of specific player through OCR, (2) Classification through deep learning model, (3) Fuzzy c-means clustering on SIFT features. The block diagram of the proposed technique is shown in Figure 1.

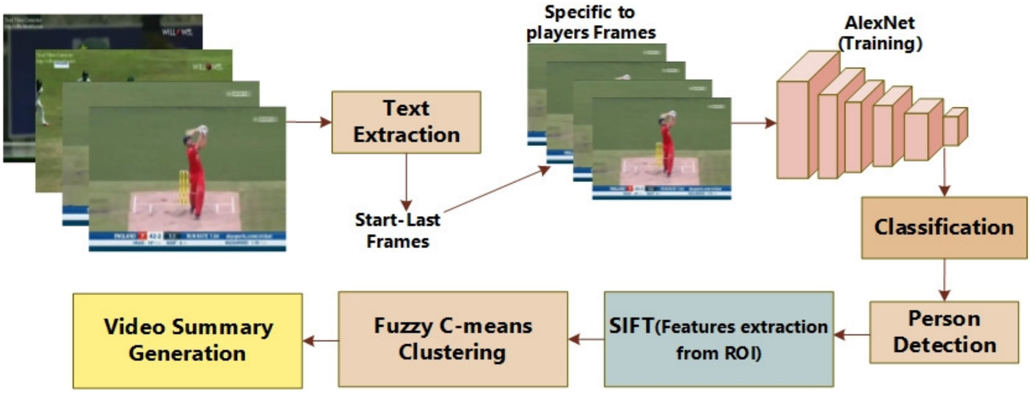


Figure 1. Block Diagram of the proposed system.

F_{in} is a collection of frames that are extracted sequentially from the original video. Initially, these F_{in} frames are used to find F_{sel} frames that are contending with the specific player by finding SFs and LFs through OCR. F_{sel} is selected from the initial collection of frames F_{in} to reduce the complexity. False Negative (FN) are frames of the negative class, while False Positive (FP) represents frames of the positive class. FP and FN are used for training of the deeply learned model. F_{cp} denotes classified frames of positive class which should be included in a video. In a cricket game, there are two possible players available for batting, so if F_{cp} has a person's view, therefore it is needed to identify that whether that person is a specific player. For this cause, F_{clus} frames are passed through the SIFT feature extraction algorithm. Furthermore, N clusters are made and all those frames which do not be-long to the relevant cluster are subtracted. At last, those frames which belong to F_{cp} and are stored in F_n are also added sequentially to F_{out} to make a complete summarized video.

3.1. Frames Identification of Specific Player through OCR

At the first stage we need all frames specific to the player name. Therefore, the Optical Character recognition (OCR) mechanism is used to extract text from score captions (SC) detail to identify the frames which include details i.e., name and status of a specific player as shown in Figure 2. This stage includes the following steps:

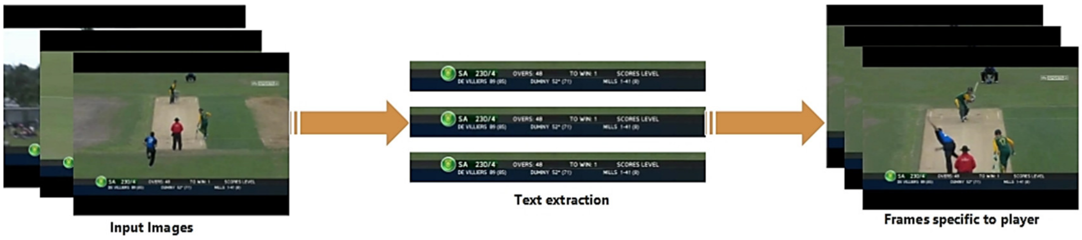


Figure 2. Extraction of Frames based OCR.

As a first step, the frames are preprocessed to prepare them for later use. As there are 24 frames per second in a video, which makes it is difficult and complex to process. To minimize this complexity, we reduce the number of frames and preprocess frames to enhance the visibility of frames with step size 10. For enhancement of frames, we improved the contrast of images. Moreover, identification of a region of interest (ROI) is required to summarize the video, however, in our technique ROI is SC in cricket videos. We resized the frames, though text detail is shown in the lower part of the frame. In the end, the OCR mechanism is used to extract text information from SC. Now we have those frames on which we will apply our technique of video summarization.

3.2. Classification through Deep Learning Model

Since Convolutional Neural Networks (CNNs) have multiple layers which take sequential weeks for the training of large-scale datasets. Although GPUs can accelerate the processing speed but still the computational time is a critical challenge to deal with. The solution to this can be to reduce network layers but it is found [34] that this continuously decreases the performance. Instead of reducing the layers, we use the transfer learning technique (pre-trained network: AlexNet) for the classification of frames. The AlexNet [35] contains 5 convolutional network layers and 3 layers that are fully connected. A softmax function is applied to the final fully connected layer to measure the class scores for classification. This function implements the following equation:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{l=1}^{k} e^{z_l}} \; for \; i = 1, \ldots, K, \tag{1}$$

where K is the dimensionality of vector z and e^z is a vector range (0,1) which adds up to 1. Figure 3 shows some example frames which are used for training.

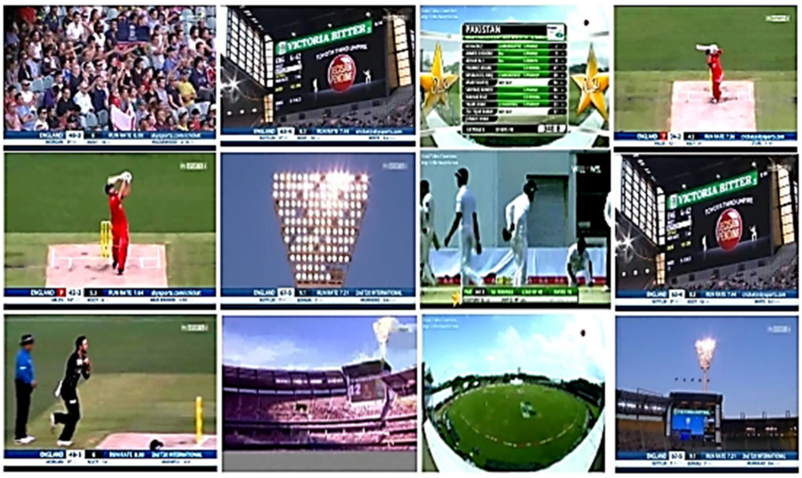


Figure 3. Some sample frames from positive and negative class.

3.2.1. Conversion to Gray Scale Image

After extracting frames from the cricket videos the first step is to convert all the frames into the grayscale so that they can be used in later stages as shown in Figure 4.

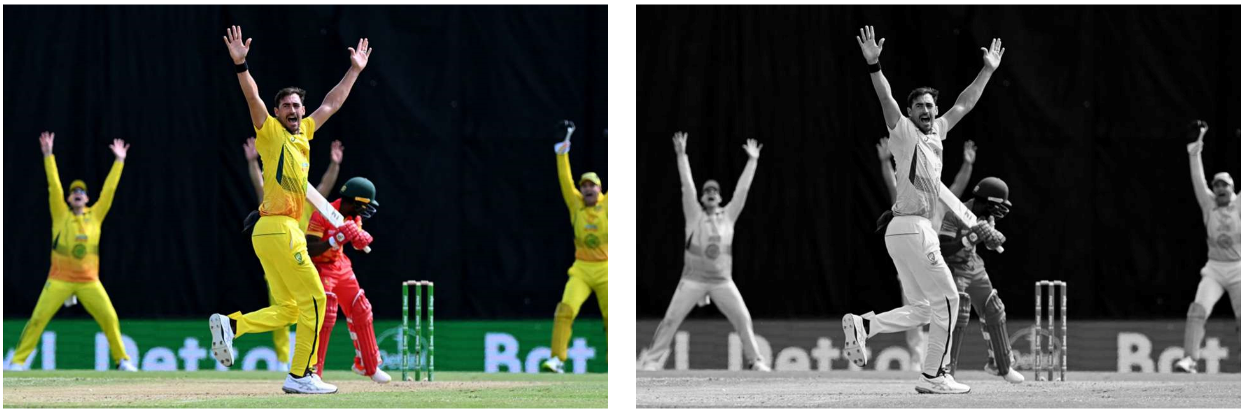


Figure 4. Conversion of color frame to gray scale.

3.2.2. Resizing

In [35], it is defined that training of architecture can be done with sub-images of size 227 × 227. Therefore, grayscale images from the above step are resized up to 227 × 227 to be used in our experiment.

3.2.3. Training

The transfer learning models have been already trained on millions of images [35]. Here we have used 40 different images include 20 images for the positive class (which should be included in the summary) and 20 images for the negative class (which should not be included in the summary). The parameters of the AlexNet are reported in Table 1.

Table 1. The parameters for AlexNet.

Parameters	**Value**
Dropout rate	0.5
Learning policy	Step-down
Epochs	40
Training batch size	128
Momentum	0.9
Weight decay	0.005

3.2.4. Testing

Testing is performed on thousands of images from cricket videos, which include scenes of positive class (player, ball view, bowler, catch the view, etc.) and negative class (crowd, pitch view, advertisement, etc.) and model accurately classifies them as positive and negative class.

3.3. Clustering

After classification, the need is to remove frames (partner player) that should not be included in highlights of the viewer's choice player. For this purpose fuzzy c-means clustering (FCM) [36], is used along with scale-invariant feature transform (SIFT) [37]. Following steps are implemented to remove extra frames

3.4. Person Detection

To classify frames according to the player, such frames are needed which include that specific player in them. For this purpose aggregated channel feature [38] algorithm is applied which gives a bounding box for each person in the frame. These bounding boxes are used in later steps for classification purposes. Figure 5 shows a bounding box having a person detected. Person re- identification can be performed as in [39], however, it costs more complexity for the proposed method.

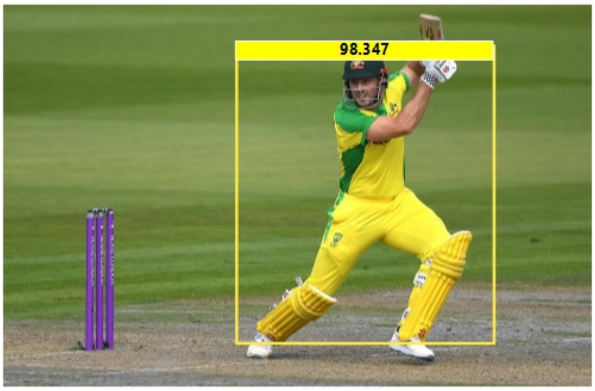


Figure 5. Bounding box showing detected person after application of aggregated channel features algorithm.

3.5. Cropping

In this step, those frames are croped to 137 × 197 dimension that contain the bounding box around the player, therefore, frames can be clustered based on the player in next step. On the other side, at this step, those frames are discared which don't have any bounding box, which consequently reduce the number of frames for next step. Figure 6 shows a cropped form of the original frame.

Figure 6. The resized frame of 137 × 197 dimension.

3.6. SIFT

Scale Invariant Feature Transform (SIFT) was introduced in 2004 to extract idiosyncratic in- variant features from frames [40]. A classic frame of 500 × 500 dimensions will have almost 2000 stable features but it al-so depends on the content of the frame and different parameters The proposed system also exploited it to ex-tract feature data from all the frames having a player. Figure 7 shows extracted SIFT features of the detected per-son.

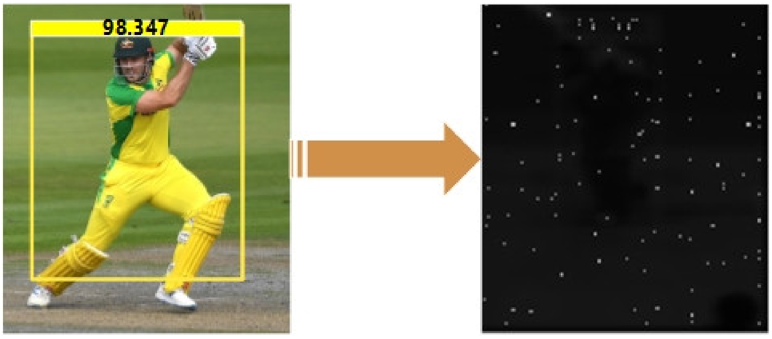


Figure 7. Extracted SIFT features of the detected person.

3.7. FCM

Fuzzy C-means clustering [36] is implemented on above extracted features data with 100 centroids. Let I = (I_1, I_2, . . . , I_n)) represents an image having N pixels to separate into C clusters, while I_i denotes features data. This algorithm decreases the cost function as follows:

$$J = \sum_{j=1}^{N} \sum_{i=1}^{C} \; u_{ij}^{m} ||I_j - V_i||^2, \tag{2}$$

$$v_i {=} \frac{\sum_{j=1}^{N} u_{ij}^{m} \; I_j}{\sum_{j=1}^{N} u_{ij}^{m}} \; , \tag{3}$$

where u_{ij} denotes the membership of pixel I_j in ith cluster, V_i is cluster center I and variable m controls the fuzziness. Cost function is decreased when pixel near to the centroid of its cluster is allocated high membership value while pixel far from its centroid is assigned low membership value. This membership function describes the possibility of pixel belonging to any cluster. Membership functions and centers of clusters are changed through the following:

$$v_i = \frac{\sum_{j=1}^{N} u_{ij}^{m} \; I_j}{\sum_{j=1}^{N} u_{ij}^{m}}, \tag{4}$$

FCM starts with a guess for each cluster center, then converges for v_i which represents the local minima of cost function. Converging point is detected by comparing two consecutive cluster center iterations or changes in member function. The working of the proposed method is given below as Algorithm 1.

Algorithm 1. Algorithm for the proposed system

```
Input: Collection of frames $F_{in}$
Output: Sequential frames having only key events $F_{out}$
Step 1: Take $F_{in}$, extract text to select $F_{sel}$ (specific to player) by finding SFs and LFs
Step 2: Training of model with $F_N$ and $F_P$
Step 3: Classify $F_{in}$ through model to get $F_{cp}$
Step 4: Person detection in $F_{cp}$ to categorize frames according to specific player $p$
            if $F_{cp}$ € $p$
              {$F_{clus} = F_{cp}$; Do step 5;}
            else
              {$F_n = F_{cp}$;      Continue;}
Step 5: SIFT features extraction of $F_{clus}$ from step 4
Step 6: Fuzzy c-means clustering in $N$ clusters
            For $I$ = 1 to $N$
            Check for matching and cluster it in respective cluster $C_i$
            End
        $F_{out} = F_{cp} - \sum_{k=1}^{C_{N-1}} F_{cp}$
$F_{out} = F_{out} + F_n$
End
```

3.8. Video Summary

Following FCM Clustering, the histogram is generated for each image having 100 feature vector lengths. As clusters of frames have been made through FCM, frames have been sequentially classified into positive and negative classes. To generate output video frames of the positive class are combined to generate a summarized video specific to the player.

4. Results and Discussion

For performance evaluation, our method is assessed on a dataset of 20 cricket videos. Measuring metrics such as precision, recall, error rate, and accuracy are applied for assessment.

4.1. Dataset

To evaluate the performance of the proposed method a dataset of 181 videos having a time duration of 8 h is formed. The dataset comprises videos from various sources. Our dataset has videos of varying shots such as long or medium length, close-up, and crowd shots. We included videos having diverse illumin conditions i.e., day light and artificial light. Further, each frame in videos has a resolution of 640 × 480 and a frame rate of 25 fps. Detail explanation of datasets is given in Table 2. The purpose of constructing the dataset from these resources has the following benefits:

- The samples are suitable for classification having various challenges for highlights generation.
- The selected images are from well-known sources that are easily accessible.

Table 2. Description of Datasets.

Dataset	No. of Videos	Video Length (Hours)	No of Frames	Resolution	Format
South Africa vs. New Zealand	50	2	180,000	640 × 480	Avi
Bangladesh vs. Pakistan	48	2	180,000	640 × 480	Avi
Germany vs. Brazil	67	3	270,000	640 × 480	Avi
India vs. Pakistan	16	1	90,000	640 × 480	Avi

In-text extraction from video coordinates are set to [100, 329, 230, 16]. SF and LF are identified by matching the name of that specific player in the extracted text. AlexNet [41] is trained for the classification of 2 classes i.e., firstly for those frames which have meaningful information such as players and fielders, secondly for those which include signboards, runs detail, and audience, etc. Figure 8 shows the basic architecture of AlexNet used for the proposed method.

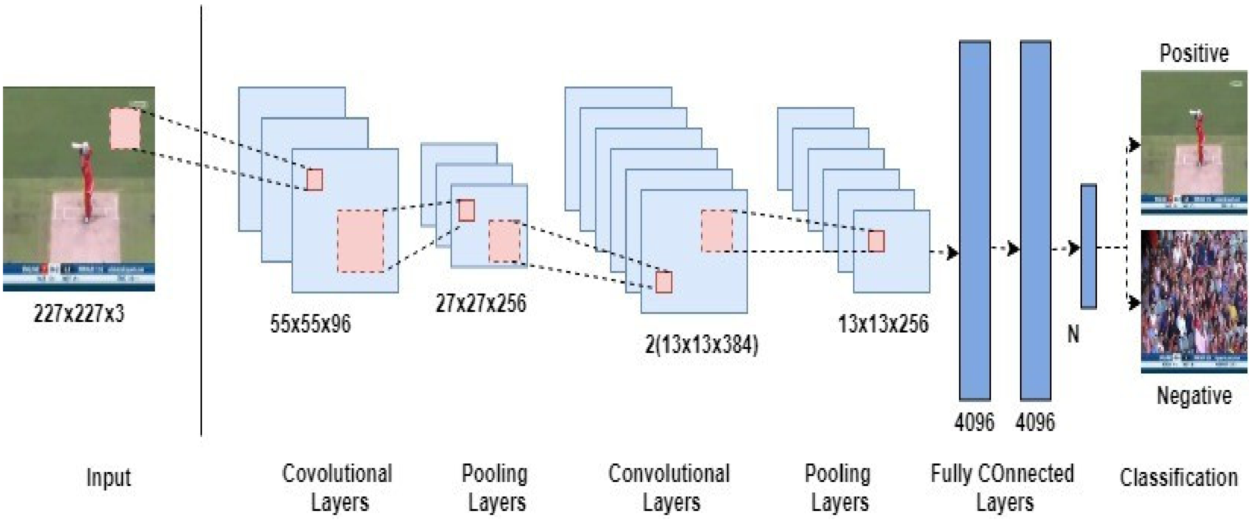


Figure 8. The architecture of AlexNet for the proposed method.

4.2. Performance Evaluation

This section provides detailed results of the experiment conducted on videos along with a comprehensive analysis. The efficacy of our method is assessed by generating highlights of several players of each cricket video in the dataset. SFs and LFs are detected through OCR, classified with the supervised deep learned model, and clustered through FCM to separate frames to generate summarized videos. Results are computed on individual video based on four main parameters such as True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). We utilized four metrics for the performance evaluation of the proposed system such as precision, recall, accuracy, and error rate. As videos are comprised of frames and our proposed system generated the final summarized video consisting of keyfames while discarding non keyframes. Therefore, we considered included frames as TP and discarded frames as TN. The overall performance of the proposed method is computed as the average of individual video performance.

More specifically, TP shows the number of frames correctly classified as a positive class, FP represents the number of frames incorrectly identified as a positive class, TN presents the number of frames correctly classified as a negative class and FN represents the number of frames incorrectly classified as negative class. The recall is the measurement of the ability to correctly classify positive frames. The recall is the measurement of the classifier’s ability to correctly classify Negative Frames. Precision is the measurement of the classifier’s ability to correctly classify positive frames. Accuracy is the measurement of the classifier’s ability to correctly differentiate positive and negative class frames. Error rate refers here to the erroneous percentage of the proposed technique. Equations for precision, recall, accuracy, and error rate are given below:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{5}$$

$$\text{recall} = \frac{\text{frames correctly classified as positive}}{\text{frames correctly classified as positive} + \text{positive frames}}, \tag{6}$$
$$\text{incorrectly classified as negative}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{7}$$

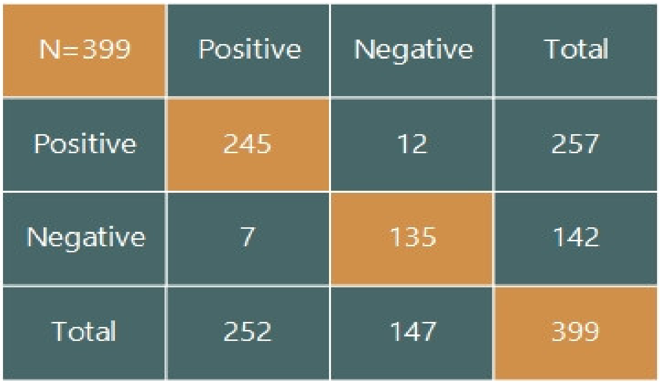
$$\text{precision} = \frac{\text{frames correctly classified as positive}}{\text{frames correctly classified as positive} + \text{negative frames} \atop \text{incorrectly classified as positive}}, \tag{8}$$

$$\text{accuracy} = \frac{\text{TP}}{\text{TP} + \text{TN}}, \tag{9}$$

$$\text{accuracy} = \frac{\text{frames correctly classified as positive}}{\text{frames correctly classified as positive} + \text{frames} \atop \text{correctly classified as negative}}, \tag{10}$$

$$\text{error rate} = 100 - \text{accuracy}, \tag{11}$$

In Figure 9, 399 average frames which have to be classified as positive or negative class. Positive class refers to, frames that should be included in the final summarized video. Among which, 245 frames are those which are correctly classified with our proposed technique in positive class. While 135 frames are those which are correctly classified as negative class. The remaining 19 frames are those which are not classified correctly. Figure 10 shows overall results for each video after the implementation of the proposed method on the dataset.



N=399	Positive	Negative	Total
Positive	245	12	257
Negative	7	135	142
Total	252	147	399

Figure 9. Confusion Matrix for the proposed system.

4.3. Comparison with Existing Techniques

In this experimentation performance of our proposed scheme is compared with the previous state of the art. Precision and Recall are applied to assess performance as shown in Table 3. Figure 11 shows the comparison graph of precision and Recall of proposed and previous techniques based on the results of Table 3. It can be observed from the boldface results in Table 3 that the proposed technique achieves high-performance measures as compared to other existing techniques. [42] proposed the method of soccer video summarization and applies it to a dataset of 17 videos. It can be seen in the graph that our proposed technique gives 5% better precision than [43]. In [44], the proposed method for slow-motion replay detection in basketball videos and applied it on 10 videos of cricket. Whereas, our proposed technique also provides more than 1% better precision than slow-motion replay detection technique as shown in Table 3.

Table 3. Performance comparison of the proposed method with existing techniques.

Techniques	Length (Hours)	Format	Frame	Resolution	No. of Videos	Precision	Recall
[42]	13	MPEG-1	30 fps	352 × 240	17	85.2%	80%
[45]	18	-	-	-	06	61.3%	77%
[46]	3:30	MPEG-7	-	-	05	83%	66%
[44]	2:30	-	-	-	08	61.2%	74.8%
[43]	3	X64	30 fps	320 × 240	04	80.2%	81.1%
[47]	6	-	-	-	10	55.8%	80.7%
[44]	25	MPEG-2	30 fps	480 × 352	10	90%	92.8%
Proposed System	8	AVI	25 fps	640 × 480	181	91.21%	93%

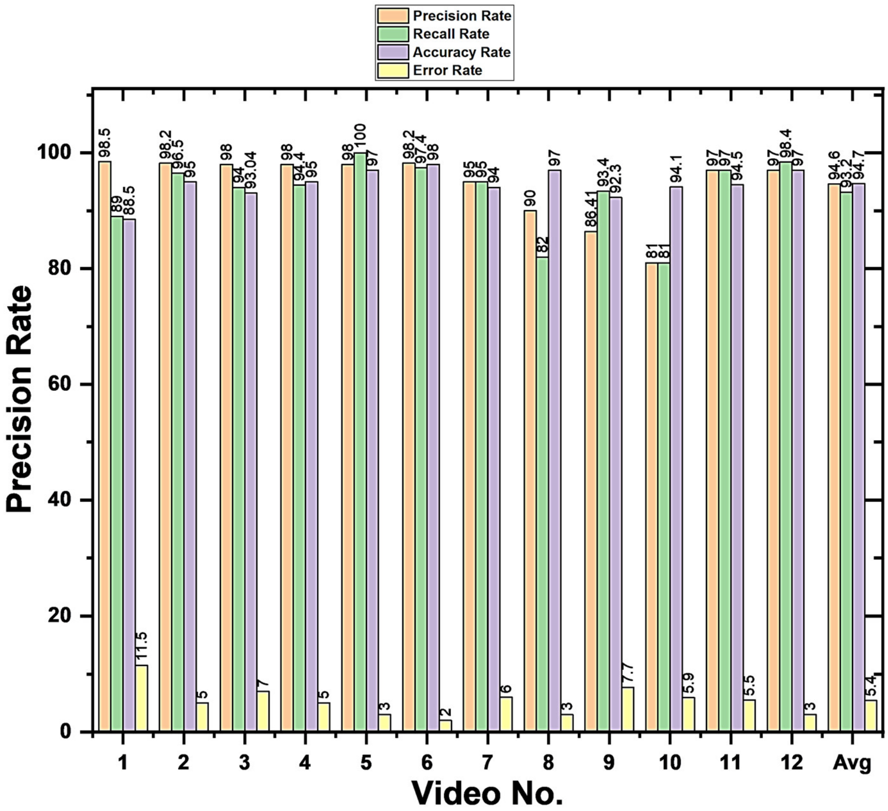


Figure 10. Visual representation of the performance of our proposed model.

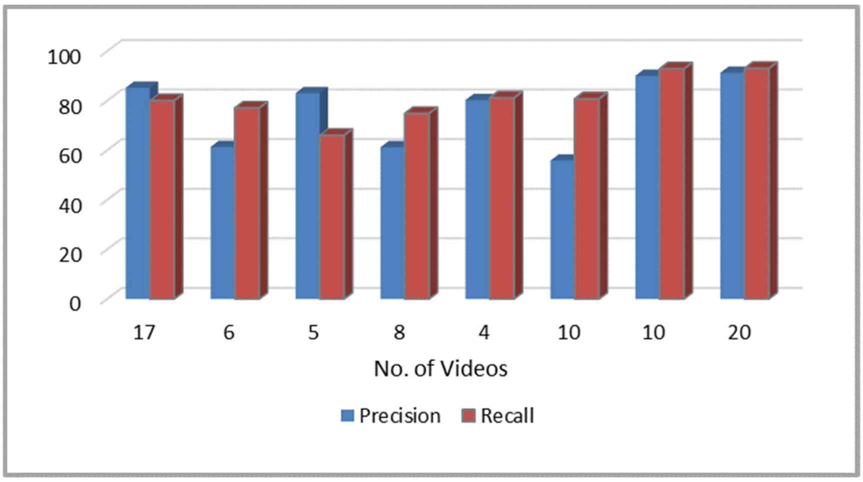


Figure 11. Comparison graph of the proposed method with state of the art.

4.4. Results

Test performance of the proposed system is examined via the receiver operating characteristic (ROC) curve. Figure 12 represents the ROC curve for our system. As result, it can be noticed that the performance of our method is very effective such as the classification of frames and generation of summary.

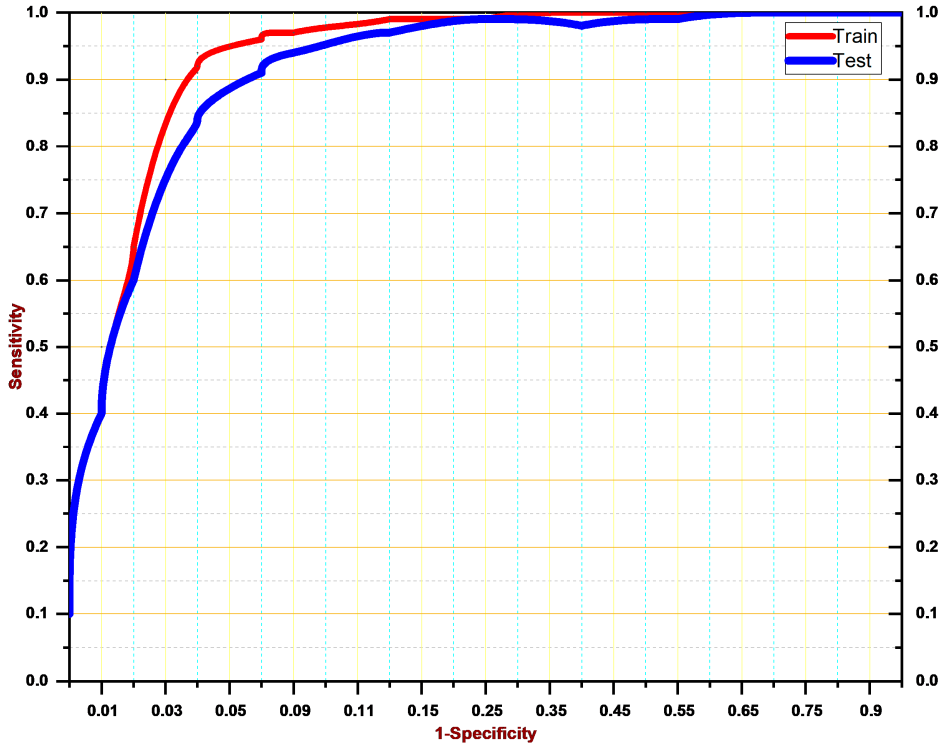


Figure 12. ROC curve for the proposed system.

In Table 4 detail of the videos from dataset is given, that depicts the detail analysis of videos. It is clear from table that the frames in output videos are minimized comprehensively. These are the frames which were classified as positive class and depict the performance of any one specific player of viewer's choice.

Table 4. Detailed analysis of videos.

Video No.	No. of Frames	Starting Frame (SF)	Ending Frame (EF)	Frames in Output	TP	TN	FP	FN
1	2602	560	2602	390	320	25	5	40
2	1728	01	1218	580	545	5	10	20
3	1727	143	307	230	174	40	4	12
4	1728	757	1728	690	420	235	10	25
5	1727	001	1727	515	308	200	7	0
6	1727	686	1197	501	265	224	5	7
7	1727	18	1727	320	180	120	10	10
8	1717	1	1717	245	27	210	3	5
9	1728	1	1728	233	85	130	12	6
10	1727	172	1727	202	25	165	6	6
11	1728	01	1728	440	280	140	10	10
12	1712	184	1712	464	314	135	10	5
Avg					245	135	7	12

In Table 5 performance of our proposed system is presented individually on 12 videos from dataset and Accuracy rates are satisfactory. Our proposed technique is compared with state of the art techniques, and results show that our technique outperforms the existing techniques. Performance of our proposed system is mainly depends on the deep neural network classifier. As through OCR, we get all frames of specific player i.e., when he starts playing to when he gets out, but further selection of key-frames is due to a deep neural network. AlexNet uses visual features while training or testing, therefore we trained algorithm with only two classes. First class is positive, that belongs to frames which have exciting events such as six, fours, catch, out, player running towards wicket, etc. The negative class belongs to those frames, which shows audience, banners or ground etc. Therefore, for summarization of videos, negative class images should not be added in final video. At this stage, proposed algorithm had a drawback that it included all those frames in which partner player of viewer's choice player comes for batting. But we need only specific player's key frames, so to remove this drawback Fuzzy C-means clustering is applied on frames to divide those frames into different clusters according to players. At the end, only those cluster's frames are added in final video which purely belong to specific player. Hence, after application of three stages i.e., OCR, Deep Neural network i.e., ALexNet and Fuzzy C-means clustering, our proposed system generates the summaries of cricket videos specific to the player very well.

Table 5. Results of Video Summarization for Cricket Video dataset.

Video No.	Precision Rate (%)	Recall Rate (%)	Accuracy Rate (%)	Error Rate (%)
1	98.5	89	88.5	11.5
2	98.2	96.5	95	5
3	98	94	93.04	7
4	98	94.4	95	5
5	98	100	97	3
6	98.2	97.4	98	2
7	95	95	94	6
8	90	82	97	3
9	86.41	93.4	92.3	7.7
10	81	81	94.1	5.9
11	97	97	94.5	5.5
12	97	98.4	97	3
Avg	94.6	93.2	94.7	5.4

5. Conclusions

In this study, we suggest a competent procedure for sports highlight the generation of a specific player. Although, there exist various techniques for the videos summary generation, however, our proposed model is first one to tackle the challenge of player specific video summaries generation in cricket games. We utilized machine learning and deep learning techniques to separate the frames of specific players to produce a final video. First, we employed OCR to separate frames of specific player. The proposed scheme reveals that SF and LF do not guarantee to include frames of only one player. Therefore, the proposed technique utilized deep learning algorithm i.e., AlexNet for to discard irrelevant frames performing binary classification. The proposed model does not depend only on AlexNet classification for the player-specific summary which turns it more efficient. Moreover, to optimize the videos we employed person detection technique along with SIFT features descriptor. Then, we applied FCM algorithm to separate only those frames which are related to specific player in form of cluster. The proposed technique is easily adaptable for variations such as camera, score description, broadcaster's limitations, symbol design, and placement. The efficiency of the proposed scheme is analyzed on various real-world cricket videos. Results of our experiments illustrate that we achieve overall classification accuracy greater than 94%. During experiments, it is observed that in videos with high variations, the performance of the proposed technique degrades slightly. Hence, we aim in the future to imrpove our model while redcucing its complexity and analyze the performance of the proposed scheme on another type of sports video in future.

Author Contributions: Conceptualization, R.M. and A.I.; methodology, S.U.R.; software, R.M.; validation, R.M., A.I. and S.U.R.; formal analysis, R.M.; investigation, A.I.; resources, T.M.; data curation, T.M.; writing—original draft preparation, H.T.R.; writing—review and editing, R.M.; visualization, H.T.R.; supervision, H.T.R.; project administration, T.M.; funding acquisition, S.U.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Dataset used in this study may be available on demand.

Acknowledgments: We would like to thank UET Taxila and COMSATS University Islamabad for supporting this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alghoul, A.; Al Ajrami, S.; Al Jarousha, G.; Harb, G.; Abu-Naser, S.S. Email Classification Using Artificial Neural Network. *Int. J. Acad. Eng. Res.* **2018**, *2*, 8–14.
2. Javed, A.; Bajwa, K.B.; Malik, H.; Irtaza, A. An efficient framework for automatic highlights generation from sports videos. *IEEE Signal Process. Lett.* **2016**, *23*, 954–958. [CrossRef]
3. Mahum, R.; Irtaza, A.; Nawaz, M.; Nazir, T.; Masood, M.; Mehmood, A. A generic framework for Generation of Summarized Video Clips using Transfer Learning (SumVClip). In Proceedings of the 2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), Karachi, Pakistan, 15–17 July 2021.
4. Zhu, X.; Aref, W.E.; Fan, J.; Catlin, A.C.; Elmagarmid, A.K. Medical video mining for efficient database indexing, management and access. In Proceedings of the 19th International Conference on Data Engineering (Cat. No. 03CH37405), Bangalore, India, 5–8 March 2003.
5. Mahum, R.; Suliman, A. Skin Lesion Detection Using Hand-Crafted and DL-Based Features Fusion and LSTM. *Diagnostics* **2022**, *12*, 2974. [CrossRef] [PubMed]
6. Wang, M.; Hong, R.; Li, G.; Zha, Z.J.; Yan, S.; Chua, T.S. Event driven web video summarization by tag localization and key-shot identification. *IEEE Trans. Multimed.* **2012**, *14*, 975–985. [CrossRef]
7. Mahum, R.; Irtaza, A.; Nawaz, M.; Nazir, T.; Masood, M.; Shaikh, S.; Nasr, E.A. A robust framework to generate surveillance video summaries using combination of zernike moments and r-transform and deep neural network. *Multimed. Tools Appl.* **2022**, 1–25. [CrossRef]
8. Akhtar, M.J.; Mahum, R.; Butt, F.S.; Amin, R.; El-Sherbeeny, A.M.; Lee, S.M.; Shaikh, S. A Robust Framework for Object Detection in a Traffic Surveillance System. *Electronics* **2022**, *11*, 3425. [CrossRef]
9. Nepal, S.; Srinivasan, U.; Reynolds, G. Automatic detection of 'Goal' segments in basketball videos. In Proceedings of the Ninth ACM International Conference on Multimedia, Ottawa, ON, Canada, 30 October 2001.
10. Li, B.; Sezan, I. Semantic sports video analysis: Approaches and new applications. In Proceedings of the 2003 International Conference on Image Processing (Cat. No. 03CH37429), Barcelona, Spain, 14–17 September 2003.
11. Sandesh, B.; Srinivasa, G. Text-mining based localisation of player-specific events from a game-log of cricket. *Int. J. Comput. Appl. Technol.* **2017**, *55*, 213–221. [CrossRef]
12. Kolekar, M.H.; Sengupta, S. Event-importance based customized and automatic cricket highlight generation. In Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, Toronto, ON, Canada, 9–12 July 2006.
13. Tang, H.; Kwatra, V.; Sargin, M.E.; Gargi, U. Detecting highlights in sports videos: Cricket as a test case. In Proceedings of the 2011 IEEE International Conference on Multimedia and Expo, Barcelona, Spain, 11–15 July 2011.
14. Javed, A.; Ali Khan, A. Shot classification and replay detection for sports video summarization. *Front. Inf. Technol. Electron. Eng.* **2022**, *23*, 790–800. [CrossRef]
15. Taskiran, C.M.; Pizlo, Z.; Amir, A.; Ponceleon, D.; Delp, E.J. Automated video program summarization using speech transcripts. *IEEE Trans. Multimed.* **2006**, *8*, 775–791. [CrossRef]
16. Srinivas, M.; Pai, M.M.; Pai, R.M. An improved algorithm for video summarization—A rank based approach. *Procedia Comput. Sci.* **2016**, *89*, 812–819. [CrossRef]
17. Truong, B.T.; Venkatesh, S. Video abstraction: A systematic review and classification. *ACM Trans. Multimed. Comput. Commun. Appl. TOMM* **2007**, *3*, 3-es. [CrossRef]
18. Money, A.G.; Agius, H. Video summarisation: A conceptual framework and survey of the state of the art. *J. Vis. Commun. Image Represent.* **2008**, *19*, 121–143. [CrossRef]
19. Zhang, H.J.; Wu, J.; Zhong, D.; Smoliar, S.W. An integrated system for content-based video retrieval and browsing. *Pattern Recognit.* **1997**, *30*, 643–658. [CrossRef]
20. Liu, D.; Hua, G.; Chen, T. A hierarchical visual model for video object summarization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 2178–2190. [CrossRef] [PubMed]
21. Wolf, W. Key frame selection by motion analysis. In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA, 9 May 1996.
22. Goldman, D.B.; Curless, B.; Salesin, D.; Seitz, S.M. Schematic storyboarding for video visualization and editing. *ACM Trans. Graph. TOG* **2006**, *25*, 862–871. [CrossRef]
23. Nam, J.; Tewfik, A.H. Event-driven video abstraction and visualization. *Multimed. Tools Appl.* **2002**, *16*, 55–77. [CrossRef]
24. Ngo, C.-W.; Ma, Y.-F.; Zhang, H.-J. Automatic video summarization by graph modeling. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003.

25. Lee, Y.J.; Ghosh, J.; Grauman, K. Discovering important people and objects for egocentric video summarization. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012.
26. Khosla, A.; Hamid, R.; Lin, C.J.; Sundaresan, N. Large-scale video summarization using web-image priors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.
27. Muhammad, K.; Hussain, T.; Baik, S.W. Efficient CNN based summarization of surveillance videos for resource-constrained devices. *Pattern Recognit. Lett.* **2020**, *130*, 370–375. [CrossRef]
28. Zhang, Y.; Liang, X.; Zhang, D.; Tan, M.; Zing, E.P. Unsupervised object-level video summarization with online motion auto-encoder. *Pattern Recognit. Lett.* **2020**, *130*, 376–385. [CrossRef]
29. Ji, Z.; Zhao, Y.; Pang, Y.; Li, X.; Han, J. Deep attentive video summarization with distribution consistency learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 1765–1775. [CrossRef]
30. Mahum, R.; Ur Rehman, S.; Okon, O.D.; Alabrah, A.; Meraj, T.; Rauf, H.T. A novel hybrid approach based on deep CNN to detect glaucoma using fundus imaging. *Electronics* **2021**, *11*, 26. [CrossRef]
31. Mahum, R.; Ur Rehman, S.; Meraj, T.; Rauf, H.T.; Irtaza, A.; El-Sherbeeny, A.M.; El-Meligy, M.A. A novel hybrid approach based on deep cnn features to detect knee osteoarthritis. *Sensors* **2021**, *21*, 6189. [CrossRef] [PubMed]
32. Mahum, R.; Munir, H.; Mughal, Z.-U.-N.; Awais, M.; Khan, F.S.; Saqlain, M.; Mahamad, S.; Tlili, I. A novel framework for potato leaf disease detection using an efficient deep learning model. *Hum. Ecol. Risk Assess. Int. J.* **2022**, 1–24. [CrossRef]
33. Khan, A.A.; Shao, J.; Tumrani, S. Content-Aware summarization of broadcast sports Videos: An Audio–Visual feature extraction approach. *Neural Process. Lett.* **2020**, *52*, 1945–1968. [CrossRef]
34. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014.
35. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
36. Chuang, K.-S.; Tzeng, H.L.; Chen, S.; Wu, J.; Chen, T.J. Fuzzy c-means clustering with spatial information for image segmentation. *Comput. Med. Imaging Graph.* **2006**, *30*, 9–15. [CrossRef]
37. Abdel-Hakim, A.E.; Farag, A.A. CSIFT: A SIFT descriptor with color invariant characteristics. In Proceedings of the 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), New York, NY, USA, 17–22 June 2006.
38. Brehar, R.; Vancea, C.; Nedevschi, S. Pedestrian detection in infrared images using aggregated channel features. In Proceedings of the 2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania, 4–6 September 2014; IEEE: Piscataway, NJ, USA, 2014.
39. Rehman, S.-U.; Chen, Z.; Raza, M.; Wang, P.; Zhang, Q. Person re-identification post-rank optimization via hypergraph-based learning. *Neurocomputing* **2018**, *287*, 143–153. [CrossRef]
40. Witten, I.H.; Frank, E.; Hall, M.A. Practical machine learning tools and techniques. In *Data Mining*; Elsevier: Amsterdam, The Netherlands, 2005.
41. Ballester, P.; Araujo, R.M. On the performance of GoogLeNet and AlexNet applied to sketches. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Pheonix, AZ, USA, 12–17 February 2016.
42. Ekin, A.; Tekalp, A.M.; Mehrotra, R. Automatic soccer video analysis and summarization. *IEEE Trans. Image Process.* **2003**, *12*, 796–807. [CrossRef]
43. Xu, W.; Yi, Y. A robust replay detection algorithm for soccer video. *IEEE Signal Process. Lett.* **2011**, *18*, 509–512. [CrossRef]
44. Wang, L.; Liu, X.; Lin, S.; Xu, G.; Shum, H.Y. Generic slow-motion replay detection in sports video. In Proceedings of the 2004 International Conference on Image Processing, 2004. ICIP'04, Singapore, 24–27 October 2004.
45. Chang, P.; Han, M.; Gong, Y. Extract highlights from baseball game video with hidden Markov models. In Proceedings of the International Conference on Image Processing, Rochester, NY, USA, 22–25 September 2002; IEEE: Piscataway, NJ, USA, 2002.
46. Takahashi, Y.; Nitta, N.; Babaguchi, N. Video summarization for large sports video archives. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6 July 2005; IEEE: Piscataway, NJ, USA, 2005.
47. Eldib, M.Y.; Abou Zaid, B.S.; Zawbaa, H.M.; El-Zahar, M.; El-Saban, M. Soccer video summarization using enhanced logo detection. In Proceedings of the 2009 16th IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, 7–10 November 2009.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.