International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 13, Issue 4 [April. 2024] PP: 120-127

Visual Vortex: Multimodal Video Classification Technique Based on CNN Architecture and Tensor flow

DR.M.DEEPA (ASP/IT)
VARUN.D, VIGNESH.S, RAVIKUMAR.J BACHELOR OF TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY
SRI SHAKTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY
(AUTONOMOUS) COIMBATORE - 641062

Abstract
Video classification, a vital aspect of computer vision, automates the labeling of videos based on content. This involves analyzing spatial and temporal features for pattern recognition. Key steps include data collection, feature extraction, model training, evaluation, and deployment. Deep learning architectures like 3D CNNs and RNNs are utilized, with datasets such as UCF101 and frameworks like TensorFlow. Ongoing advancements in spatiotemporal modeling continually enhance video classification performance.
Keywords: *Alex Net; Convolutional Neural Networks; Deep Learning; Image Classification; Image Net; Machine Learning.*

Date of Submission: 04-04-2024 Date of acceptance: 15-04-2024

I. Introduction

Because of the enormous evaluations of different games all over the planet, sports telecasters have created a lot of video content. As per measurements, the greater part of the total populace (around 3.6 billion) watched the 2018 Men's Football World Cup, and the worldwide viewership of the 2020 Olympic Games in Japan arrived at around 4 billion. What's more, a comparable expansion in viewership has likewise happened in different games all over the planet, which has carried extraordinary difficulties to physically examining and processing such a lot of video content. In this manner, there is a dire need to create effective strategies to consequently process and examine countless games recordings show uping in the internet. Among them, programmed video classification gives significant specialized instruments to a progression of utilizations, for example, video ordering, perusing, explanation, and recovery, as well as working on the proficiency and effectiveness of its admittance to sports video documents.

Since video is made out of many single picture outlines, video handling can be computationally difficult [1, 2]. In this manner, one technique for video grouping is to just look at the casings as a solitary picture and attempt to characterize them and afterward consolidate the outcomes into a solitary result classification of the whole video. Albeit this thought is moderately intuitive, individuals can utilize a solitary casing to recognize shots and shots. In any case, more often than not, the data encoded in the video is disposed of. To tackle this issue, there are now numerous specialists grouping recordings in view of video highlights, sound elements, and different recordings. Writing [3] utilized two different brain network techniques and surface element strategies and consolidated them to think about the consequences of each of the three strategies. In writing [4], Gade et al. utilized warm imaging innovation to create heat guides and afterward utilized head part investigation (PCA) and Fischer straight discriminator to project the intensity maps into a low-layered space and afterward characterize them. As announced, this technique accomplished great outcomes in five classifications. On this premise, Gade et al. [5] consolidated the Mel recurrence cepstral coefficient (MFCC) highlights of sound and visual movement elements to order movement recordings and acquired generally excellent arrangement results. The writing [6] combined the three sorts of information of video, sound and sensor in the movement video, and utilized the multiclass support vector machine (SVM) strategy for characterization. Writing [7] utilized a secret Markov model (Well) strategy to recognize occasions in recordings and arrange them, yet the outcomes just favorable to vided computational time execution without referencing the exactness.

In the beyond a decade, profound learning hypothesis has been broadly utilized in PC vision fields, for example, picture and video handling [8], and extraordinary leap forwards have been made. Dissimilar to pictures, video contains a great deal of time information, so an enormous number of scientists have completed research on the best way to address the time data in the video. Ji et al. [9] proposed a 3DCNN across spatial and temporal aspects to separate data about the movement that happened between outlines. Simultaneously, writing

[10] utilized the time pooling technique and long transient memory (LSTM) strategy to address time information. Wang et2al. [11] proposed an all the more generally material activity acknowledgment strategy, which gone past the layer type and model engineering. The strategy partitioned every video into various blocks, and afterward ordered each block into miniature recordings, and afterward collected the consequences of all predictions to create a last expectation for each total video. Writing [12] concentrated on video grouping depending on combination of time data from different outlines, utilizing three different procedures: early combination, late combination, and slow combination. It found that sluggish combination accomplished the best outcome in these models.

This paper utilizes the worldly difference network (TDN) to accomplish video characterization. The proposed strategy can catch multiscale time data and further understand the timing portrayals of different recordings. The center of TDN is to plan an effective time module by unequivocally utilizing the time difference administrator and methodicallly assess its effect on present moment and long haul movement demonstrating. To completely catch the time data of the whole video, TDN laid out a two-level difference displaying standard adigm. For neighborhood movement demonstrating, the time difference on successive casings is utilized to give a fine movement example to CNN. For worldwide movement displaying, the time difference across portions is joined to catch the long haul structure for movement include excitation. Probes two bar lic informational collections show that the order execution of the proposed strategy surpasses the cutting edge.

1. Algorithm Principle

The TDN-based technique proposed in this paper really utilizes the whole video data to become familiar with the video activity model. Because of the restriction of GPU memory, following the fleeting fragment organizations (TSN) system [13], a scanty and in general testing methodology is proposed for every video. Different from the TSN strategy, the TDN proposed in this paper predominantly utilizes the time difference administrator in the organization plan to obviously catch present moment and long haul movement data. To work on the proficiency of the calculation, this paper integrated the lingering association into the principal organization to finish the movement supplement in the neighborhood window and the movement upgrade across different sections. Specifically, every video is first separated into sections of equivalent span and nonoverlapping. Then, a casing is haphazardly examined from each portion, and a sum of T outline $I = [I1, I2, \cdots, IT] \in \mathbb{R}T\times C\times H\times W$ are gotten. These edges are contribution to CNN to separate casing level elements $F = [F1, F2, \cdots, FT]$, in which $F \in \mathbb{R}T\times C'\times H'\times W'$ addresses the element portrayal in the secret layer. The motivation behind the short- time difference module is to give neighborhood movement data to these casing by-outline portrayals of the past layers to further develop their portrayal abilities:

$$Fi = Fi + H(Ii), \quad (1)$$

where Fi addresses the improved portrayal of the timing differential module and H addresses the brief time frame differ-ence module, which can extricate the nearby movement of Ii in the encompassing neighboring layers. The long haul difference module is utilized to improve the casing level element portrayal by utilizing the range time structure:

$$Fi = Fi + Fi \odot G(Fi, Fi+1). \quad (2)$$

where G addresses the long haul difference module. As just nearby portion level data is thought of, the drawn out demonstrating can be acted in each long haul difference module. By stacking different long haul difference modules, the drawn out time construction can be caught.

1.1. Short-Time Difference Module. Nearby approaches in a video are practically the same in a neighborhood time window, and it is inefficient to stack various casings for resulting supportive of cessing straightforwardly. Then again, examining a solitary edge from every window can remove appearance data, yet can-not catch neighborhood movement data. Subsequently, the brief time frame difference module decides to furnish a solitary RGB outline with a period difference to deliver an effective video portrayal to encode appearance and movement data all the while.

In particular, the brief time frame difference module performs low-level element extraction in the initial not many layers of the neural network and empowers a solitary edge of RGB to catch neighborhood movement by melding time difference data. For each example outline Ii, the halfway window removes the timing RGB difference and afterward aggregates it into the channel dimension $D(Ii) = [D-2, D-1, D1, D2]$. As needs be, the effective type of TDM can be communicated as

$$H(Ii) = \text{Upsample}(\text{CNN}(\text{Downsample}(D(Ii)))), \quad (3)$$

where D(Ii) addresses the RGB difference around and CNN is a particular brain network at different stages. To keep up with effectiveness, this technique plans a lightweight CNN module to manage stacked RGB differences

D(Ii). It typically follows a low-goal handling system: (1) The normal pooling is utilized to down sample the RGB difference to half. (2) CNN is utilized to separate movement highlights. (3) The movement highlights are up sampled to match RGB highlights.
This is on the grounds that the RGB difference shows a tiny worth in many regions and contains high reaction just in regions where movement is critical. Consequently, it is adequate to utilize

a low-goal engineering for this meager sign without losing an excess of precision. The data of the brief time frame difference module is converged with a solitary RGB outline, so the first edge level portrayal grasps the movement mode and can all the more likely depict the nearby time window. This combination is accomplished through the even connection, and the combination association of the brief time frame difference module is appended to the edge level portrayal of each beginning phase. By and by, the leftover association is additionally com-pared with other combination methodologies.

1.2. Long-Time Difference Module. The edge by-outline rep-resentation of the brief time frame difference module is very effective for catching spatio worldly data in the nearby fragment (window). In any case, this portrayal is restricted with regards to time responsive field, so it is difficult to investigate the drawn out time design of the learned activity model. In this manner, the long haul difference module endeavors to utilize cross-section data to upgrade the first representation through the new two-way and multiscale time difference module. Notwithstanding proficiency, the absence of arrangement of spatial situations between long haul outlines is one more issue that should be addressed. Subsequently, this strategy plans a multiscale design to smooth the difference in the huge responsive field before the difference calculation. The element aspect is first compacted into a proportion and convolved to further develop proficiency, and the arrangement time difference is determined through neighboring video cuts:

$$C(Fi, Fi+1) = F - Conv(Fi+1), \quad (4)$$

where C(Fi, Fi+1) addresses the difference in the arrangement season of the section, which is utilized for spatial smoothing channel convolution, in this manner reducing the lost arrangement
issue. Then, the adjusted time difference is extricated through the multiscale module for long haul movement information:

$$M(Fi, Fi+1) = \text{Sigmd Conv} \sum_{j=1}^{N} CNNj(C(Fi, Fi+1)) \quad (5)$$

In Condition (4), different spatial scales are intended to remove movement data from different open fields. Practically speaking, N = 3. For missing arrangement issues, their combination might be more hearty. As far as execution, it includes three branches: (1) short association, (2) 3 × 3 con- volution, and (3) normal pooling. At long last, the two-way range time difference is utilized to improve the edge level highlights, as displayed beneath:

marginally different from the brief time frame difference module. The difference portrayal is utilized as the consideration guide to upgrade the casing level highlights. This part founded on the noticed consideration model is more effective in the later phase of CNN. By and by, the remaining association is additionally com-pared with other combination methodologies.

In synopsis, the TDN-based strategy proposed in this paper depends on the meager examining of TSN, which operates on a progression of casings uniformly circulated across the whole video. TDN gives a two-level movement demonstrating mechanism, zeroing in on catching time data in a neighborhood to worldwide way. Specifically, the momentary timing difference module is embedded in the beginning phase to perform better and low-level movement extraction. Furthermore, the drawn out timing difference module is embedded in the later stage to perform more coarse and high level time extraction structure modeling. Like the lingering network [14], this strategy takes on its principal structure. The initial two phases utilize a brief time frame difference module to separate transient data, and the last three phases are outfitted with a long-term difference module to catch the cross-time frame timing structure. To work on computational productivity, two measures are embraced. For nearby movement demonstrating, a remaining connection is added between the brief time frame difference module in the first and second stages and the fundamental organization. For long haul movement demonstrating, a long haul difference module is added to every lingering block. 3

2. Experiment and Discussion

In this segment, the effectiveness of the proposed technique is checked and contrasted and other existing strategies. In the first place, the two public informational indexes utilized for assessment are momentarily presented. Then, the exploratory subtleties are given. Lastly the trial results are broke down.

Video Data Modalities

When contrasted with pictures, recordings are more difficult to comprehend and arrange because of the complicated idea of the worldly substance. In any case, three unique modalities, i.e., visual data, sound data, and text data, may be accessible to group recordings rather than picture characterization, where just a solitary visual methodology can be used. In view of the accessibility of various modalities in recordings, the errand of characterization can be sorted as a uni-modular video grouping or a multi-modular video characterization, as summed up in Figure 2. The current writing has used both of these models for the video arrangement errand, and it is for the most part accepted that models using multi-modular information perform better compared to the models in view of uni-modular information [20,21]. Besides, the visual depiction [22] of a video works better compared to the text [23] and the sound [24,25] portrayal for the order motivation behind a video.

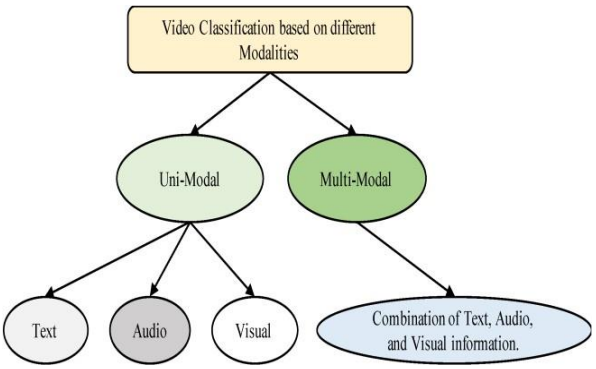


Figure 2. Various modalities utilized for order of recordings.

Alongside the improvement of all the more remarkable profound learning models in the new years, the pattern for the video grouping task has followed a shift from customary hand tailored ways to deal with the completely computerized profound learning draws near. Among the exceptionally normal profound learning designs utilized for video grouping is a 3D-CNN model. An illustration of 3D- CNN engineering utilized for video order is given in Figure 3 [43]. In this design, 3D blocks are used to catch the video data important to group the video content. Another extremely normal engineering is a multi-stream design, where the spatial and fleeting data is independently handled, and the elements removed from various streams are then combined to pursue a choice. To handle the worldly data, various techniques are utilized, and the two most normal strategies depend on (I) RNN (predominantly LSTM) and (ii) optical stream. An illustration of a multi- stream network model [44], where the transient stream is handled utilizing optical stream, is displayed in Figure 4. An undeniable level outline of the video order process is displayed in Figure 5, where the phases of element extraction and expectation are displayed with the most well-known sort of methodologies utilized in the writing. In the forthcoming segments, the forward leaps in video grouping and review connected with order of recordings, explicitly utilizing profound learning structures, are summed up, portraying the achievement pace of using profound learning models and the related constraints.

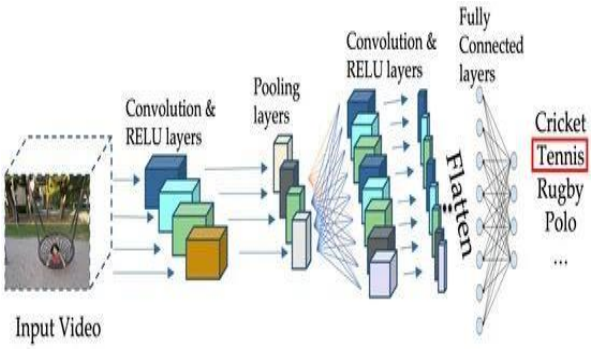


Figure 3. An example of 3D-CNN architecture to classify videos.

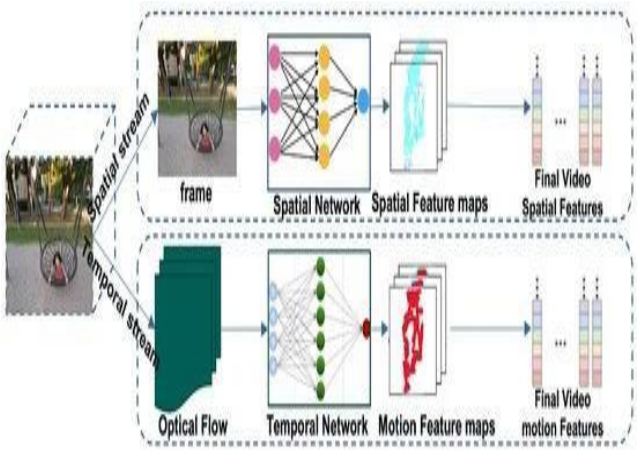


Figure 4. An example of two-stream architecture with optical flow.

2.1. Experimental Information. Two informational indexes are utilized for tests, and their fundamental portrayals are given in Table 1. The UCF Sports Activity informational index [15] is made out of a bunch of activities gathered from different games. These games are generally communicated on radio and TV slots, and the video groupings are gotten from different material sites. The informational index contains the accompanying 10 games: jumping (14 recordings), golf swing (18 recordings), kicking (20 recordings), weightlifting (6 recordings), horse riding (12 recordings), running (13 Video), skateboard (12 recordings), swing stool (20 recordings), swing side (13 recordings), strolling (22 recordings), and test outlines are displayed in Figure 1. The informational index incorporates a sum of 150 groupings with a goal of 720 × 480 and 10 fps. The accessible comments are the jumping box for activity situating and the class mark for movement acknowledgment. Furthermore, the informational collection likewise gives comments from the crowd.

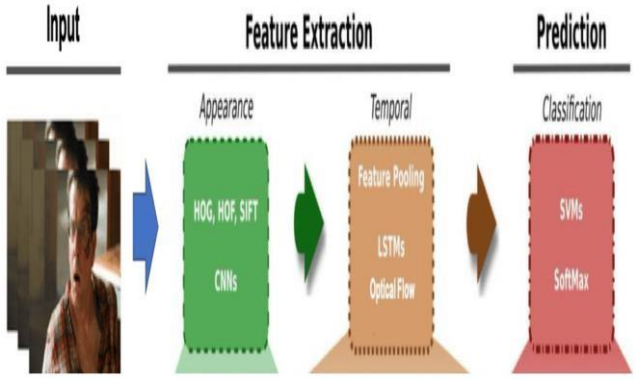


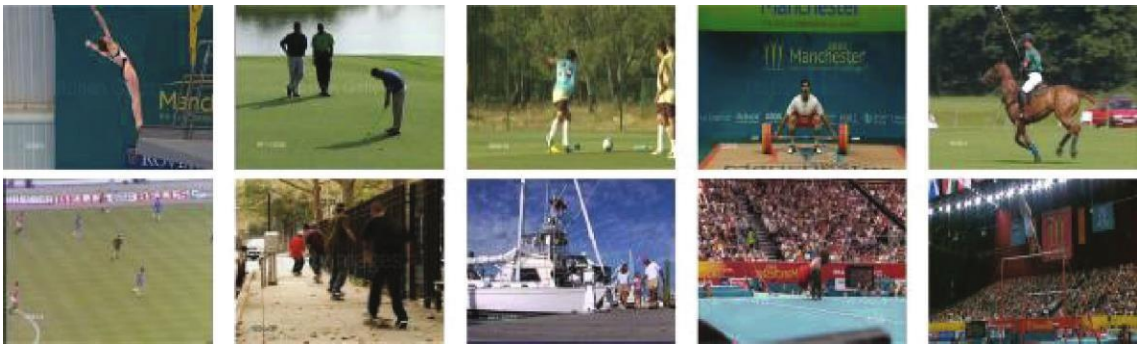
Figure 5. An overview of video classification process.

$$Fi \odot G(Fi, Fi+1) = Fi \ominus_2 [M(Fi,) + M(Fi+1, Fi)], \quad (6)$$

In addition, this paper also uses the SVW field sports data
$Fi+1$
where $\odot$ denotes element-wise multiplication.
The proposed method also combines the original frame- level representation and enhances the epresentation through residual connection. For example, Equation (2) is set for experiments. The data set consists of 4200 videos taken by Coachs Eye smartphones. The Coachs Eye smartphone application is an application for sports training developed by TechSmith. The data set includes 30 types of sports and 44 different actions. Compared with the UCF sports

TABlE 1: Basic descriptions of the two data sets.

Data set	Number of sports	Number of videos	Resolution
UCF	10	150	720 × 480
SVW	30	4200	480 × 272 to 1280 × 720



(a) UCF data set

Figure 1: Examples in the two data sets.

(b) SVW data set

activity informational index, this informational collection is more complicated. The greater part of the recordings taken are beginner sports recordings. Simultaneously, the shooting strategies are not generally so proficient as television recordings. In the first place, the static picture setting has a low level of segregation for characterization. Second, the jumbled foundation and normal climate will likewise carry challenges to unconstrained games video characterization. Then, beginner clients' nonprofessional shooting brings extra difficulties, like outrageous camera vibration, erroneous camera development, being darkened by the crowd, judges and fences because of inappropriate camera position, and uncommon review points. A few instances of the SVW field sports informational index can likewise be seen in Figure 1.

2.2. Experimental Information and Assessment Pointers. In the examination, this strategy utilizes the ResNet50 organization to carry out the timing difference module and tests the time period from every video T = 16. During preparing, every video outline is acclimated to have a more limited edge in and is edited haphazardly 224 × 224. The strategy is pretrained on the ImageNet informational collection. The group size is 128, the underlying master ing rate is 0.001, and the quantity of emphasess is set to 100. At the point when the exhibition of the approval set is soaked, the learning rate will be diminished to 0.0001. For the purpose of testing, the more limited side of every video was acclimated to 256, and afterward, just the 224 × 224 focus yield of a solitary clasp was utilized for assessment. The equipment climate of the entire examination process is Intel Center i7-10700 2.9GHz computer chip, NVIDIA GeForce GTX 2080Ti (11 GB video memory) GPU, and 32GB Slam memory, and the figuring stage is Python 3.7 and Tensorflow2.0.

2.3. Experimental Outcomes. For the UCF sports activity informational collection, 75% of the video cuts are utilized for preparing, and the leftover 25% are utilized for testing. To demonstrate the effectiveness of the proposed strategy in video arrangement, this paper contrasts it and the a few existing techniques. These

techniques incorporate manual element strategies and profound learning strategies, essentially including Wang et al. [16], Le et al. [17], Kovashka et al. [18], Thick directions [19], Weinzaepfel et al. [20], SGSH [21], bits [22], and two stream LSTM [23]. Among them, the last two are profound learning techniques, and the others are manual component strategies.

Table 2 gives the grouping precision of every sort of sports in the UCF informational index. Table 3 shows the arrangement consequences of different techniques. It tends to be seen that the

TABlE 2: The accuracy each class of UCF Sports data set.

Class	Accuracy
Diving	100%
Golf	89.7%
Skate boarding	93.1%
Swing bench	100%
Kicking	100%
Lifting	100%
Riding horse	100%
Running	100%
Swing side	100%
Walking	91.1%

TABlE 3: Comparison with the state of the art methods (UCF data set).

Method	Mean Acc.
Wang et al. [16]	85.6%
Le et al. [17]	86.5%
Weinzaepfel et al. [20]	90.5%
SGSH [21]	90.9%
Snippets [22]	97.8%
Two stream LSTM [23]	98.9%
Proposed	99.3%

TABlE 4: Comparison with the state of the art methods.

Method	Test I	Test II	Test III	Mean Acc.
Motion-assisted	\	\	\	39.1%
Context-based [24]	\	\	\	37.8%
Combined CNN [25]	81.9%	82.1%	83.4%	82.5%
RWRS [26]	84.5%	84.3%	85.3%	84.4%
Proposed	87.5%	88.5%	86.3%	86.8%

execution of the profound learning strategy is superior to the manual component technique, and the strategy proposed in this paper accomplishes the best characterization consequence of 99.3%, which surpasses the two stream LSTM [23] technique 0.5%. This is chiefly in light of the fact that the TDN utilized in this paper can more readily portray the timing structure in sports recordings. It tends to be seen from the singular games grouping brings about Table 3 that the vast majority of the games order exactness rates in this informational index have reached 100 percent, with the exception of golf, skating, and strolling sports. The justification for the disarray of these sorts of sports recordings is that they all have a similar activity, that is to say, the activity of strolling.

For the SVW informational index, the examination embraces a similar preparation/testing design depicted in [24], in which there are 3 different preparing/testing set division techniques. The trial examination can be displayed in Table 4. From Table 4, it very well may be seen that the strategy proposed in this paper is superior to the movement based [24] (movement based highlight, Hoard include, and SVM classifier) technique given by the first informational index. The exhibition is 27% higher than that of the CNN technique [25, 26]. In particular, it very well may be seen from Table 4 that the exactness of the "running" occasion class is absolutely

horrible. The vast majority of the mistakes are that the classifier misclassifies the "running" occasion classification picture as the "long leap" occasion classification. These issues emerge in light of the fact that there are a great deal of similar activities between the long leap and running, particularly when the long jumper runs up.

II. Conclusion

This paper proposes a video characterization technique in view of TDN, which is utilized to gain sports activity models from the whole video. The center of the timing difference network is to sum up the time difference administrator into an effective universally useful time module with a particular plan, which is utilized to catch present moment and long haul time information in the video. As the trial results on two public informational indexes show, the presentation of the extricated time series difference calculation is superior to other past techniques. In the following stage, the timing differential organization will be improved to supplant the 3DCNN ordinarily utilized in video demonstrating for video arrangement.

- **Information Accessibility**

The informational indexes utilized in this paper are can be gotten to upon demand.

- **Irreconcilable situations**

The creators announce that there are no irreconcilable situations in regards to the distribution of this paper.

References

[1]. F. Dong, Y. Zhang, and X. Nie, "Dual discriminator generative adversarial network for video anomaly detection," IEEE Access, vol. 8, pp. 88170–88176, 2020.

[2]. K. Doshi and Y. Yilmaz, "Any-shot sequential anomaly detection in surveillance videos," in Proceedings of CVPRW, pp. 934-935, 2020.

[3]. K. Messer, W. Christmas, and J. Kittler, "Automatic sports classification," in 2002 International Conference on Pattern Recognition, pp. 1005–1008, 2002.

[4]. R. Gade and T. B. Moeslund, "Sports type classification using signature heatmaps," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 999–1004, 2013.

[5]. R. Gade, M. Abou Zleikha, M. G. Christensen, and T. B. Moeslund, "Audio-visual classification of sports types," in 2015 IEEE International Conference on Computer Vision Workshop, pp. 768–773, 2015.

[6]. F. Cricri, M. J. Roininen, J. Leppänen et al., "Sport type classification of mobile videos," IEEE Transactions on Multimedia, vol. 16, no. 4, pp. 917–932, 2014.

[7]. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[8]. J. Shuiwang, X. Wei, Y. Ming, and Y. Kai, "3D convolutional neural networks for human action recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221–231, 2018.

[9]. J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, Eds., "Beyond short snippets: deep networks for video classification," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

[10]. W. Limin, X. Yuanjun, Y. Zhe Wang, L. D. Qiao, T. Xiaoou, and G. L. Van, Eds., "Temporal segment networks: towards good practices for deep action recognition," in European Conference on Computer Vision, 2019.

[11]. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, Eds., "Large-scale video classification with convolutional neural networks," in proceedings of the IEEE conference on computer vision and pattern recognition, 2014.

[12]. L. Wang, Y. Xiong, Z. Wang et al., "Temporal segment net- works: towards good practices for deep action recognition," in European Conference on Computer Vision, pp. 20–36, 2016.

[13]. H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, "Deep residual learning for image recognition," in IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

[14]. K. Soomro and A. R. Zamir, Action Recognition in Realistic Sports Videos, Computer Vision in Sports, Springer International Publishing, 2014.

[15]. H. Wang, U. M. Muneeb, K. Alexander, L. Ivan, and S. Cordelia, "Evaluation of local spatiotemporal features for action recognition," in The British Machine Vision Conference, p. 124, 2009.

[16]. Q. V. Le, W. Y. Zou, S. Y. Yeung, and Y. N. Andrew, "Learning hierarchical invariant spatiotemporal features for action cognition with independent subspace analysis," in IEEE Conference on Computer Vision and Pattern Recognition, pp. 3361– 3368, 2014.

[17]. K. Adriana and G. Kristen, "Learning a hierarchy of discriminative pace-time neighborhood features for human action recognition," in IEEE Conference on Computer Vision and Pattern Recognition, pp. 2046–2053, 2015.

[18]. W. Heng, K. Alexander, S. Cordelia, Liu, and Cheng-Lin, "Action recognition by dense trajectories," in IEEE Conference on Computer Vision and Pattern Recognition, pp. 3169–3176, 2013.

[19]. W. Philippe, H. Zaid, and S. Cordelia, "Learning to track for spatiotemporal action localization," in IEEE International Conference on Computer Vision, pp. 1–9, 2015.

[20]. A. Ashwan, L. Yu-Kun, and S. Xianfang, "Saliency guided local and global descriptors for effective action recognition," Computational Visual Media, vol. 2, no. 1, pp. 97–106, 2016.

[21]. M. Ravanbakhsh, H. Mousavi, M. Rastegari, V. Murino, and L. S. Davis, "Action recognition with image based CNN fea- tures," 2015, http://arxiv.org/abs/1512.03980.

[22]. H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two stream LSTM: a deep fusion framework for human action recognition," in IEEE Winter Conference on Applications of Computer Vision, pp. 1–s, 2015.

[23]. S. S. Morteza, L. Xiaoming, U. Lalita, A. Brooks, W. John, and Craven, "Proc Dean. Sports videos in the wild (SVW): a video dataset for sports analysis," in International Conference on Automatic Face and Gesture Recognition, pp. 1–8, 2015.

[24]. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Image net classication with deep convolutional neural networks," in Proceeding of Advances in Neural Information Processing, pp. 1–9, 2012.