Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Recognising Actions with Segmentation and Prediction Techniques in ROI based Deep Learning Framework

Manoj Kumar.K,

Research Scholar,
Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai

L. Sujihelen

Assistant Professor,
Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai

Article Info
Page Number: 4072 - 4090
Publication Issue:
Vol 71 No. 4 (2022)

Abstract

Predicting the action of human beings has been an interesting research domain in the recent decades. Computer Vision is the domain associated with monitoring of video sequences for extracting meaningful patterns and predicting the future action by mapping the features of one or more frames. Motion detection, featured patterns derivations and supposed action prediction are defined by annotations and human action labelling performed through conventional neural network architectures. The known shortcoming of the conventional techniques is the exclusion of temporal features from annotations. Considering the cost associated with human action prediction, many models tend to exclude the temporal features or in certain models, the feature cannot be included into the mutual processing model along with other significant features. The proposed model investigates the challenges and difficulties associated with traditional models, and hence provides a resolution to explore the betterments introduced in the pre-processing stages. Regions of Interests are defined in potential frames of video sequences to highlight the important features and these ROI will be acting as representative factors for deriving better outcomes. The proposed model demands lesser computational costs as the

number of features to be used in computation depends on ROIs, thereby limiting the number of required resources. The ROIs are described using the timestamps (start and end) apart from the location of those regions, enabling them to determine the actions based on action queries. The proposed model is categorized into pre-processing with background separation and HOG, followed by censoring of relevant regions of interests, summarizing the action frames, and finally predicting the future actions. Since the temporal information is considered in the proposed approach, action based queries and prediction of future action is facilitated by proper action based video segmentation. Investigated simulation results prove that the proposed technique predicts the complete actions with greater accuracy and quicker detection, better than conventional techniques.

Article History
Article Received*: 25 March 2022*
Revised*: 30 April 2022*
Accepted*: 15 June 2022*
Publication*: 19 August 2022*

Keywords: Deep Neural Networks, human action prediction, HOG, video segmentation, spatio-temporal information**.**

INTRODUCTION

Action recognition is an interesting domain in the field of Computer Vision, that classifies an action from a sequence of images or frames from a video. A predominant problem in the domain is multi-classification, where the images or frames are retrieved from a video sequence from various sources, making the annotation process challenging [1].Output of prediction models is a label used to define the action that best describes the human action. According to multiple researchers, human action is a complex term to be defined or predicted as the chances of false positives are higher. Prediction of human action is subjected to multiple factors, actually more than machine could derive, as they involve factors like motivation, intention and emotion. These factors play a significant role in the next action and hence ordinary motion and appearance factors cannot be sufficient for any prediction algorithm [2]. Apart from the challenges, the error prone domain gets affected due to composition, image or video quality and stress. Videos recorded everyday might result in multiple frames or images, leading to extensive and exhaustive quantity of volumes. Literally, manual action recognition processes could take years for a successful prediction, demanding an automated technique for action recognition and prediction techniques. The number of applications to be benefitted through human action recognition are limitless, ranging from biochemical analysis for drug detection in sports, crime detection and prevention, behaviour analysis, real time video

monitoring, traffic analysis, healthcare and elderly care, social media video censoring etc. Automatic tagging of videos is best known for masking the videos of vulgarity, objectionable, harmful and sensible information [3].

In the traditional approaches, machine learning has been extensively used for recognizing human actions based on expert intervention and definition of relevant features. The problem with defined human features was the restriction of images or videos owing to environmental features, geometrical, gradient, depth and other metrics commonly found in images and videos. These mathematical and geometrical factors affect the performance of machine learning models, also resulting in excessive computational time. Raw frames collected from input videos are subjected to restricted discriminative features, due to incomplete or inaccurate labelling of features. Over all these challenges, the models attempt to yield outcomes for various purposes such as object detection, positional and posture detection, trajectories for vehicles or pedestrian patterns, or information about different structures [4]. To overcome the restrictions of machine learning models, deep learning approaches were investigated, considering its ability to monitor complex information and hence derive utmost levels of interpretations derived from complex raw videos. Complex information is derived from significant patterns or similarities from previous comparisons. The standard forms of deep learning approaches are Convolutional Neural Networks and Recurrent Neural Networks. Especially for computer vision applications, CNN and RNN are extensively used for action recognition frameworks in various domains. Object detection, posture detection and calculation, text and writing detection, annotations for scenes, trajectory detection and estimation, and saliency detection was easily predicted by convolutional neural networks. The approaches commence with computation of relevance, based on which the features are derived, later grouped for classification of features based on similarities [5]. For a video level prediction mechanism, CNN expected a longer version of videos for classification and prediction. Yet, the common drawback of the videos in captured devices is the limitation of video length that restricts the CNN to predict the sequence of actions. Motion information was inadequate for labelling or annotation of features and hence making the prediction process challenging.

Recurrent Neural Networks on the other hand, included a Long Short Term Memory framework for estimating the sequential actions from different frames of the input video. The model has considered the modes of measuring temporal features and modelling based on temporal factors. Individual frames were carefully analysed using the LSTM models and better dynamics of individual video

frames were yielded with better accuracy. Yet, the drawbacks of RNN and LSTM were prevalent when the redundant information is processed. Sequential information in consecutive frames of a video will eventually fail to derive significant features and thus classification became challenging [6]. Derived high level features from redundant video frames were unable to predict or recognize the action of human beings. The models were expected to identify numerous individual parameters from independent frames or shots of different video sources. All these individual parameters eventually led to the unique action. With frequent and dynamic changes in the video frames, added with redundancy, forced the approach for the look out of better video frames or key video frames. From the surveyed articles, it is evident that deep learning has been a well-established framework for monitoring complex inputs such as videos for extracting important patterns. This feature has advocated the implementation of deep learning frameworks in multiple research works. The highly beneficial deep learning frameworks were dependent on increased training time from the very basics, training on all probable causes and thus highly expensive hardware and software resources. The consumption of time can be reduced if a network can be trained in advance, thereby implementing a pre-trained network and to imply the learnings onto a new framework through transfer learning models [7-8]. The lower levels of the deep learning architecture are frozen in order to prevent additional computations, and the upper levels are trained over the fresh set of input videos. Every new video will be training the models for better perceptions and hence the model gets improvised.

The proposed model in the article contemplates a model for processing relevant key frames which possess the significant regions of interest from input videos, based on HSV colour variations from different histograms [9]. Various frames containing the regions of interest will be delivered from a deep learning model that is composed with LSTM and CNN. The model is trained on defined actions, their vocabulary, which are later used for predicting the future actions. Presented article contributes to the domain through the following summarization.

1) The actions are predetermined, and their vocabulary are defined, from various datasets retrieved from real time sources.
2) Long Short Term Memory is included into the architecture for maintaining the stages of memory of regions of interest from previous key frames.
3) Detection algorithm for identifying the reasons of interest is presented after the pre-processing steps to remove the redundancy found in different sources of video.

4) Depiction of lexical analysis for all actions present in the regions of interest and hence use them for prediction the future actions.

5) Segmentation of the video depending on the identified regions of interest and predicted actions and validating the model with different test cases.

This research article is sectioned as follows. Section 2 has carried out an extensive review and documented the existing techniques for action recognition and prediction. The proposed method illustrated and presented in the section 3. Section 4 compare the performance of proposed method over other state of art techniques and finally Section 5 concludes with future directions.

RELATED WORKS

The previous techniques presented for action recognition can be classified into two major divisions, namely, models which used human defined features for identifying actions, and the models which implemented a deep learning framework for action recognition and prediction. In the models which required human intervention, apriori model was implemented for understanding and perceiving the features defined by human experts [10]. There was no dependency on any mathematical approaches when deep learning frameworks are implemented, as they are capable of learning intricate details during the training duration on their own.The underlying features for every input video can be detected easily as the model has been already trained on similar videos. In the conventional approaches for action recognition, the model should already include descriptions of the actions, action and its corresponding feature descriptors, detectors and all this will eventually lead to the prediction of reaction by deriving the probable trajectories. With every new type of video, new descriptors and detectors are predefined for including all possible actions into the training model. The common types of destructors are Scale Invariant Feature Transform (SIFT), histogram of oriented gradients (HOG), speed up robust features (SURF), motion boundary histogram, local binary pattern, local ternary pattern, shapes-based distributors, location-based descriptors, size-based descriptors [12]. All these types of descriptors have been extended to meet the requirements of specific industries and videos. Similarly, there are models which were capable of detecting human action from a significant distance and recognizing them for necessary counter-actions. In this model, a bounding box was in place to identify the action and tracking the bounding box enabled the model to predict the next set of actions. The level of displacement of the bounding box assisted the model to predict and recognize human action. Another technique implemented, calculated the differences

between space and time using SVM classifier but the test was limited to 25 people and four different scenarios [13].

Vocabulary of actions was defined as spatial-temporal measures and used for action recognition along with techniques such as SURF and HOG, bag of words and by combining various other features. These techniques are proven to work independently as well as combined for better feature detection. The common data sets used for action recognition purposes were Skating, KTH, HOHA and its variants. These models were efficient but facing a limitation of detecting actions of the same genre and hence failed to detect temporal distinguishing factors between one frame and another [14]. On the whole, the model failed to derive the overall purpose and context of the given situation. Owing to difficulties in aligning the timestamps in accordance to real actions performed in real time scenarios, the model delivered an abstract action corresponding to the common actions. The models had huge reliance on mathematical models for discriminating different backgrounds based on environmental conditions. From the input video frames, the models delivered a little context about the actual action performed, which cannot be considered for predicting the future actions [15].Detecting the actions and predicting the future actions, especially with deep learning frameworks, has gained popularity due to its easiness and contemplated factors. Convolutional Neural Networks were the common models implemented irrespective of the input streams, being a single stream or multi-streams, or fusion of different networks together for processing the videos [16]. Three dimensional conditional neural networks and recurrent neural networks along with LSTM have been introduced lately for eliminating the effects of the previous models. Many models advocated the monitoring mechanism of spatial temporal features over a prolonged period of time for correctly recognising the action performed by an individual in a given scenario. The distinguishing factors between two keyframes of the same video, capture at different time stamps, can relatively help out the prediction of future actions.

Image classification is a significant problem in many methods due to various reasons such as improper capturing, distorted images, pixelated images, lack of focus and other pre-processing defects. This drawback was addressed in a model, which was trained for identifying unique features of nearly 1.2 million images of very high resolution. Images belonged to 1000 independent classifications, which was completely different from one another. The deep convolutional neural network was implemented and was rigorously trained using the thousand different classes of images captured in different sceneries [17]. The said approach was later extended to include spatial-

temporal and the deep learning model was extensively trained on two dimensional images retrieved as frames of video sources. From the identified features, the modern still lacked real time motion sensing and it was not capable of detecting in real time situations. These models also faced the difficulty in the learning process for distinguishing the unique features present in different key frames [18]. At this point, models with human defined key features performed better than automated models.

Three-dimensional CNN models concentrated on defining the motion, and differences between two keyframes based on time, for two sequential keyframes found in the videos. These models implemented both spatial and temporal features between different layers of the implemented three-dimensional CNN model [19]. It was tested against the data set retrieved from Sports1M1 video libraries, for estimating the performance of feature detection. The cost associated with the implementation was pricey and hence researchers looked up cheaper versions for automated feature extraction processes. In an extended model of 3D-CNN, a one-directional convolution approach was implemented to achieve asymmetric benefits. The asymmetric benefits of CNN range from lesser cost to reduced dependency over the resources and better efficiency over the traditional three-dimensional models [20]. Many researchers implemented a multi-stream network for improving the accuracy and efficiency and presented a separate model for spatial recognition and a separate model for temporal recognition. All these models have exhibited significant betterments even when a diverse range of data sets were implemented and capable of performing end-to-end video analysis. The drawback of these types of multi-stream networks still lacked the solution for long term temporal analysis.

CNN and RNN were combined in a technique that developed a recurrent network along with LSTM for extracting the features found in sequential order of keyframes in the entire video. These features were extracted based on spatiotemporal measurements found between two significant keyframes in sequential order [21]. Another remarkable improvement was attained when LSTM was backed with temporary features, that enabled the model to work for a longer sequence of videos. Based on the spatial features, along with relevant information about the time differences, the models were able to capture motion successfully and were tested with UCF101 data sets. In order to ensure that LSTM with pooling of temporal features, C2LSTM was also subjected to test the effectiveness of the model and the model delivered promising results [22]. From then, all models included the spatiotemporal attention (STA) which was extended into the two-stream models for better efficiency. After

extracting the relevant key features, a knowledge integration network was implemented for recognizing the different actions present in the keyframes. HMDB512, Kinetics-4003 data sets were also used for evaluating the performance of LSTM with temporal features pool, which exhibited the same operations and performance [23].

The next important stage of action recognition is to provide proper validations for the prediction and segmentation of the videos. A bag of words or a vocabulary of all possible actions were defined in a few models which acted as the key set for further research works. A shot detection algorithm was in place to find out the probable action in a single keyframe, that can be used as a reference for other video analysis tools. The drawbacks of the tagging model are the inability to mark the motions with relevant keywords, as all models will be processing video as their inputs. The SVM classifier [24] was implemented to detect human actions along with CNN models and was tested against the human-defined features ovary UCF sports and KTH input videos. AlexNet has also proven to be a fruitful solution for a Pre-trained model and was implemented in sports data sets [25]. The common videos used for evaluating the models were extracted from cricket and football where the approaches implemented HOGs for the identification of human action, summarization, and recognizing the supposed actions in a free keyframe in the dynamic videos.From the surveyed articles, it is evident that action recognition is a significant process in video analysis and still lacks in terms of accuracy and quality [26]. The commonly identified issues in the proposed domain are irregular motions of a human being or the subject, background noises, vantage point disadvantages, population, occlusion, the complexity of computations, illumination, and much more. Frame by frame analysis is extremely important for differentiating the keyframes from normal frames, the definition of significant features is important for action recognition, spatiotemporal information is required for accurate prediction of future actions [27]. The previous models have been successful to some extent in achieving either of the stated requirements but a comprehensive solution is yet to be defined.

PROPOSED METHODOLOGY

Datasets

Whenever the models of deep learning are implemented for analysing human action, the required data sets but usually incomplete or unavailable. Based on the available standard datasets, UCF-101 and UCF11 are taken into account for this research work and the template is formed by the

combination of multiple data sets such as videos from YouTube and KTH. UCF dataset is predominantly used for action recognition [28-30], which comprises 101 categories of unique actions. It comprises nearly 13320 videos identified based on the camera angle, motion, posture, object detection, scales, viewpoint, background separation, cluttering, illuminations, and lighting conditions etc.UCF11 is another standard dataset that consists of 1600 videos with the frame rate of 29.97 per second. The video classes are organized in 25 groups and contain a minimum of 4 videos per category. Every video may contain similar backgrounds and human beings, sharing the same features across the data set.Similarly, KTH dataset is a standard dataset for activity recognition and it consists of clapping, jogging, waving, walking, running, and boxing as the standard action classes [31]. For every action there are approximately 25 different individuals performing the same action on a different background. The conditions are varied based on indoor, outdoor and with different clothes on the outdoors. This data set comprises of nearly 600 videos with a resolution of 160X120.

Pre-processing Stage

The overall model is differentiated into three stages, pre-processing, identification of regions of interest, and classification. The overall quality of the model is determined by efficiently processing and feature selection techniques. The accuracy and efficiency of the deep learning model rely on quality pre-processing techniques. The input videos, taken from the data sets, are implied onto the pre-processing where the entire video is split into sequential frames indicating every individual action [32]. The background is separated using a local SVD binary pattern algorithm which is best known for identifying unique features using descriptors and to address the variations resulting due to shadows, noise, and illumination. The purpose of implementing the background separation technique is to reduce the computational time and to narrow it down to the specific regions of interest in every keyframe. The number of frames is based on the number of rows and columns of every individual pixel on the given keyframe. In this regard, the foreground and background are carefully analysed in every keyframe and ground truth for the background is defined by BG(x,y). Once the background information is defined for the entire video, the foreground aspects will commence the lookout for individuals on the given keyframe. In order to remove the additional computational cost, the proposed pre-processing technique implemented an approach to reduce the number of redundant images or frames [33]. Processing one keyframe for every 6 frames has yielded the same accuracy and improved the speed of pre-processing with reduced cost.The following equation 1 expresses the function of background separation from the input videos.

$$BG(x,y) = \{BG_1(x,y),...,BG_{index}(x,y),...,BG_n(x,y)\} \quad (1)$$

The action recognition model usually commences with background separation in order to identify the moving object. There are other statistical methods, temporal methods for differentiating, and optical flow techniques that perform the same operations for identifying the background and foreground objects [34]. Especially in a video sequence, and when objects are moving, background subtraction techniques are mandatory. A single background removal model is required for accomplishing the background subtraction from the keyframes. This is applicable for all indoor and outdoor videos and by comparing the number of consecutive frames, the static background can be determined. The subtraction of backgrounds ensures faster detection and identification of moving objects.Once the background has been determined using equation 1, the frames are identified by the following equation 2.

$$f_k = \begin{bmatrix} f_{11} & f_{12} & \cdot & \cdot & f_{1m} \\ f_{21} & f_{22} & \cdot & \cdot & f_{2m} \\ \cdot & \cdot & f_{ij} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ f_{n1} & f_{n2} & \cdot & \cdot & f_{nm}, \end{bmatrix} \quad (2)$$

Potential Factors for Action Determination

From the obtained data sets, the actions are defined as potential factors based on which the classification will be done by the deep learning model. In this approach, we have defined a standard set of potential factors which correspond to individual actions. Also known as action vocabulary in previous techniques, the proposed model also highlights the potential factors based on the background information obtained from the keyframes [35]. Depending on the classification of background information, event-based potential factors are defined in the proposed technique. The number of potential factors can increase from time to time when transfer learning is applied over the new model. The common set of actions or potential factors are determined as the following table 1.

Table 1: Potential Factors for Action Recognition

Activity	Background	Probable Action
Writing	Blackboard / Green Board	Teaching in an Institution

Moving Arms / Limbs	Playground / Road / Bars	Fighting
Panicked People based on trajectory	Any Building	Fire / Gun shoot / Accident / Disaster
Pistol / Rifle / Guns	Practice Area / School	Shooting
Temperature Checks	Entry point of Building	Corona Checks

Proposed Model

The shape of a human being is a potential factor for identifying the required regions of interest on any given keyframe. Human action can be recognized based on the extracted blobs after background separation techniques are implemented. Various techniques involved in human figure recognition concentrate on universal features, specific regions or boundaries, and motion parameters [36]. The noise present in the input videos can be removed by applying the background subtraction techniques. The background separation techniques are useful for erosion control and morphological dilation control, thereby preserving the structural elements of any given keyframe. The next phase is to implement boundary mapping for a human being by marking histogram-oriented gradients. Support vector machine classification is used for frame voting and measuring the spatiotemporal factors. HOG is a standard technique used for retrieving potential information about features by using relevant descriptors during the training and testing phases.The boundary of a human being can easily be detected after the background subtraction and by using a technique for measuring in abrupt changes from 1 frame to another based on visual contents.

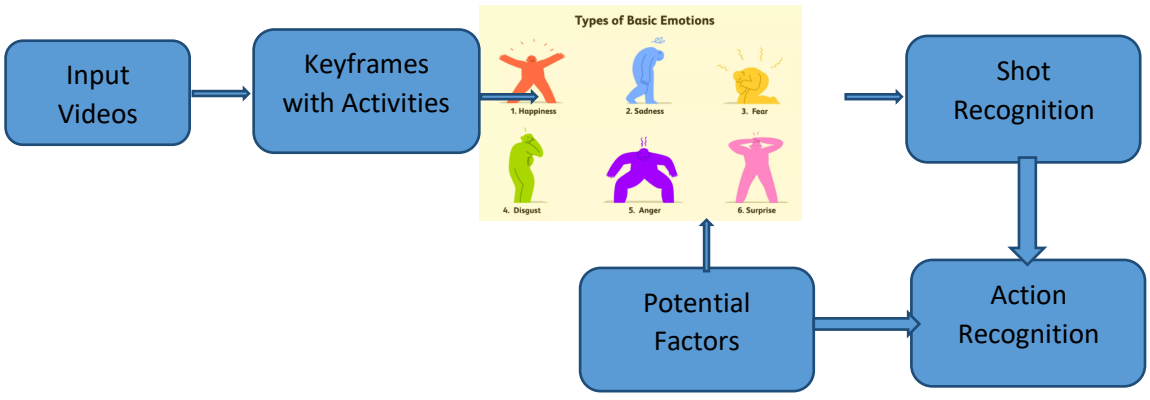

Figure 1: Proposed Architecture

The number of input videos depends on the types of classes taken for action recognition and the proposed model has considered 900 videos of 10 classes, and each class possessed 90 videos each. The duration of the videos ranged from 3 seconds to 30 seconds and the entire data set is split into 80% of training videos and 20% off testing videos. The keyframes are extracted based on frames per second, Boolean per second indicates the parameters of each frame accordingly. Depending on the location of human beings in any given frame, the Boolean value indicates the potential measures for detecting the action for further classification. The number of frames considered for detecting the action depends on the accuracy delivered in each cross-fold validation. Action recognition and prediction are more concentrated on a particular human being in a given frame and the overall computation time for delivering the results.It is understood that every video can be split into numerous frames and the proposed histogram-oriented gradients can yield better feature extractions for delivering the recognized action. Once the background separation process is completed, histogram-oriented gradients will look out for features significantly corresponding to the action of human beings in the given frame. The proposed method extracts the information about every keyframe and measures them against the standard threshold for or remarkable changes in the visual content. Mean and standard deviation are considered to be the common bases of measurement after HOG features are normalized.

The proposed method as illustrated in Figure 1, also includes a model which can convert human figures into skeletal figures. Many models have included deep integration cameras for commercial purposes, which are capable of detecting the skeletal positions automatically. Human skeletal joints can be detected much quicker than other computational models. The proposed methodology includes the whole body and the skeletal position for better accuracy during action recognition and prediction. When a deep learning model is approached with skeletal information illumination effects, lighting effects and noises will not be a disturbing factor.The skeletal positions can be indicated as (X_i, Y_i, Z_i), where X, Y, and Z indicate the coordinates and i indicates the pixel position. For each frame, the value of Skeleton will fall within a range of 0 to 255.The spatial and temporal information is thus represented by measuring the variation between one frame and another. In the case of a colour image/frame, the skeletal metrics are measured in terms of the following equation 3.

$$\mathbf{(X_i => R,\ Y_i => G,\ Z_i => B)\ (3)}$$

Algorithm

```
while True:
    ret, frame = capture.read()
    if frame is None:
        break
fgMask = backSub.apply(frame)
cv.rectangle(frame, (10, 2), (100,20), (255,255,255), -1)
cv.putText(frame, str(capture.get(cv.CAP_PROP_POS_FRAMES)), (15, 15),
cv.FONT_HERSHEY_SIMPLEX, 0.5 , (0,0,0))
cv.imshow('Frame', frame)
cv.imshow('FG Mask', fgMask)
```

RESULTS AND DISCUSSIONS

Estimating the Regions of Interest

The feature extraction process is the most important segment of the action recognition model. Signal features of skeletal information, contour information from abrupt visual changes, motion flow changes in frames are measured through binary pattern variations. Blobs of human figures can be obtained from background separation phase. The datasets will determine the respective activities and actions performed on the frames, owing to the potential factors for action determination. The common drawbacks of applying CNN with RGB models were based on continuous streams of different nature and to overcome this drawback the proposed model investigates classification based upon background separation. In the previous techniques, a voting mechanism was included for performing the classification operation, and the present model considered visual cues, depth, and skeletal parameters are considered for classifications. Another addition to overcoming the drawback considers temporal and spatial dynamics of the skeletal parameters. The proposed CNN architecture understands the unique features of every frame and uses them to predict the skeletal sequences. The layers in the convolutional neural network architecture are commonly on convolutional layers, pooling layers, hidden layers, and fully connected layers. The pooling layers are used for retrieving

the reduced dimensions of input frames and the connected layers are responsible for forming 3-dimensional cubic data.

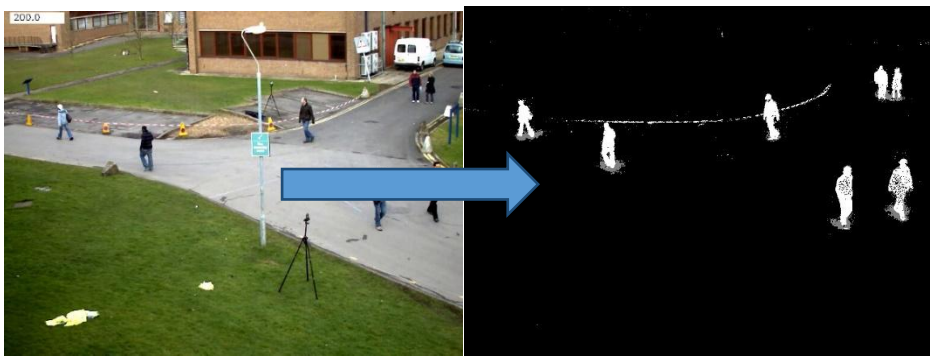

Figure 2: Background Separation and Regions of Interest through Action Paramaters

The proposed CNN architecture is trend according to the data sets and classes adhered to the different potential parameters for activity detection. From the total data set 80% is taken for training the model and the remaining 20% is taken for testing and validation. From the investigative results, it is evident that the proposed architecture has performed better than other state of art techniques in terms of learning the features and recognizing the human action. The training accuracy of the proposed model reaches almost 98% during 30 epochs and validation accuracy is around 88% during 46 epochs. The proposed model has carefully executed the controlling mechanism over the overfitting problem but the same cannot be ensured after 160 epochs. The performance of the CNN architecture along with the LSTM is presented in terms of cross-validation and the confusion matrix depicts the accuracy ratio of feature extraction and action recognition. From the input data set, human beings holding guns were predicted correctly in 39 out of 40 videos. Smoking was successfully detected in all the input video frames, fighting scenes raised false positives for playing a game. The Corona temperature checks were misjudged for holding a gun in two cases. The average prediction accuracy of the proposed system is 96%. The Other standard parameters suggest precision, recall, and F1 score were estimated and tabulated in table 2 and 3.

Table 2: Action Recognition Accuracy Results over UCF101 & KTH

Techniques	Accuracy
ImageNet + SVM Classifier	68.8%
Temporal Stream CNN	87.9%
Two Stream Fusion Network	88.1%

LSTM Composite	83.4%
Convolution Pooling Technique	88.3%
Feature Stacking Technique	89.2%
Proposed Model	**96%**

Table 3: Potential Frame Extraction for Action Recognition

Sample Video	Precision	Recall	F1 Score
Acrobacia	91.12	82.34	86.88
Person Entering a room	92.1	86.42	81.58
Spike	95.45	81.22	82.6
Test Video from Youtube	96.26	92.16	90.14
Test Video from Other sources	96.4	91.17	92.16
Kylie	87.5	100	90.33

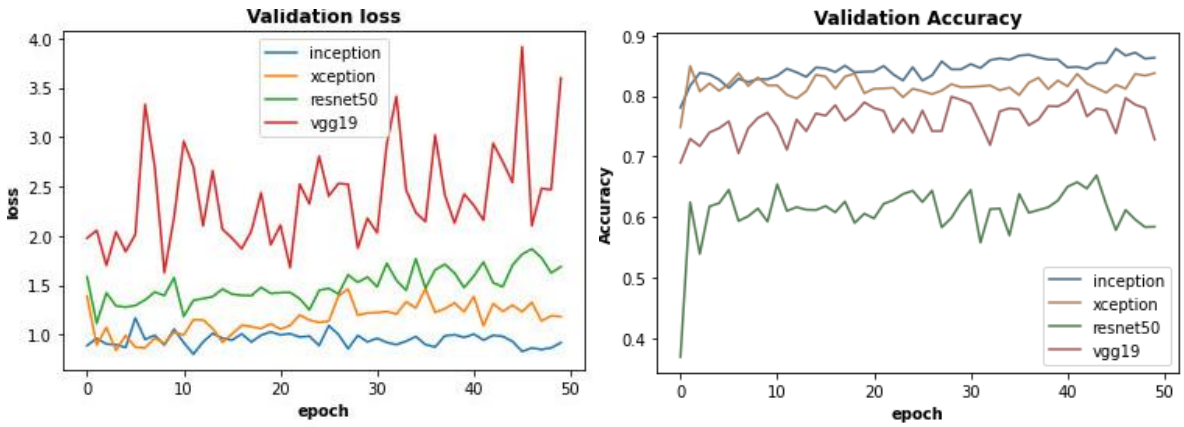

Figure 3a: Validation loss of the proposed model, (3b)Validation accuracy of the proposed model

The number of keyframes extracted from different tests during the cross-fold validation depends on the histogram gradient, the distance between the histogram parameters, and the threshold values. In a video that comprises nearly a thousand frames, 12 to 14 keyframes are extracted for activity prediction and action recognition. The extracted keyframes ensure that there is no similarity between the other consecutive frames, thereby ensuring abrupt changes in the visual content. This mechanism

assures that the proposed technique does not allow redundant information to be processed for action recognition. The approach used in the proposed technique for extracting potential frames with potential factors of activity is tested across the standard videos taken from the data sets along with some test videos taken from YouTube and other sources.

CONCLUSION

The proposed technique contemplated an action recognition technique for predicting the human activity and thereby segmenting the videos according to the respective classes. A convolutional neural network was implemented along with long short-term memory for defining the potential factors for activity prediction and comparing them for recognizing the actions performed in the video. The keyframe extraction technique ensured that there are abrupt visual changes between One Frame and another in order to remove redundant frames. The proposed methodology commenced with a proper pre-processing technique that enabled the key feature to be distinguished from normal frames. This approach has reduced the chances of pregnancy to a great extent and hence promised utmost quality and accuracy. Keyframes acted as a significant measure for detecting the activity leading to proper action generation and recognition. This technique reduced the computational complexity along with enhanced accuracy. Segmentation of videos allowed better-predicted actions for core classification, segmentation, archiving, labels generation, annotations of videos, and summarization of videos. The same methodology can be extended to integrate more Complex actions and alert the officials based on and ethical and crime events. This approach can be a potential lifesaver in alarming situations if the model is trained with intricate details of every keyframes.

REFERENCES

[1] Y. Han, P. Zhang, T. Zhuo, W. Huang, and Y. Zhang, ``Going deeper withtwo-stream ConvNets for action recognition in video surveillance," PatternRecognit. Lett., vol. 107, pp. 83-90, May 2018.

[2] Bajaj P, Pandey M, Tripathi V, Sanserwal V. Efficient motionencoding technique for activity analysis at ATM premises. In progressin advanced computing and intelligent engineering. Berlin:Springer; 2019. p. 393–402.

[3] Carreira, Joao, and Andrew Zisserman. “Quo vadis, action recognition? a new model and the kinetics dataset.” proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[4] Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. ICLR (2019)

[5] Ke, Q., Fritz, M., Schiele, B.: Time-conditioned action anticipation in one shot. In: CVPR (2019)

[6] He D, Zhou Z, Gan C et al (2018) StNet: Local and Global Spatial-Temporal Modeling for Action Recognition. arXiv:1811.01549

[7] Chen BH, Shi LF, Ke X. A robust moving object detection inmulti-scenario big data for video surveillance. IEEE Trans CircuitsSyst Video Technol. 2018;29(4):982–95.

[8] Guo S, Qing L, Miao J, Duan L (2018) Deep Residual Feature Learning for Action Prediction. In: BigMM, pp 1–6

[9] Denton, E., Fergus, R.: Stochastic video generation with a learned prior. arXiv preprint arXiv:1802.07687 (2018)

[10] Zhu, Yi et al. “Hidden Two-Stream Convolutional Networks for Action Recognition.” Lecture Notes in Computer Science (2019): 363–378. Crossref. Web.

[11] Tran, Du, et al. “Convnet architecture search for spatiotemporal feature learning.” arXiv preprint arXiv:1708.05038 (2017).

[12] S. Ilyas and H. U. Rehman, ``A deep learning based approach for precisevideo tagging," in Proc. 15th Int. Conf. Emerg. Technol. (ICET), Dec. 2019, pp. 1-6.

[13] O. Makansi, E. Ilg, O. Cicek, and T. Brox, “Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction,” in CVPR, 2019.

[14] Tang, Y., Ma, L., Liu, W., Zheng, W.S.: Long-term human motion prediction by modeling motion context and enhancing motion dynamic. In: IJCAI (2018)

[15] F. Ali, S. El-Sappagh, S. M. R. Islam, D. Kwak, A. Ali, M. Imran, andK.-S. Kwak, ``A smart healthcare monitoring system for heart disease predictionbased on ensemble deep learning and feature fusion," Inf. Fusion, vol. 63, pp. 208-222, Nov. 2020.

[16] C. Sahin, G. Garcia-Hernando, J. Sock, and T. Kim, "A review on object pose recovery: from 3d bounding box detectors to full 6d pose estimators," arXiv:2001.10609, 2020.

[17] Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: Highresolution image synthesis and semantic manipulation with conditional gans. In: CVPR (2018)

[18] M. Majd and R. Safabakhsh, "Correlational convolutional LSTM for human action recognition," Neurocomputing, 2019.

[19] J.-J. Hwang, T.-W. Ke, J. Shi, and S. X. Yu, "Adversarial structure matching for structured prediction tasks," in CVPR, 2019.

[20] S. Shah, K. Khatri, P. Mhasakar, R. Nagar, and S. Raman, "Unsupervised GIST based clustering for object localization," in Proceedings of the 2019 National Conference on Communications (NCC), pp. 1–6, IEEE, Atlanta, GA, USA, February 2019

[21] J. Zhang, Y. Wang, M. Long, W. Jianmin, and P. S. Yu, "Z-order recurrent neural networks for video prediction," in ICME, July 2019.

[22] Z. Hu and J. Wang, "A novel adversarial inference framework for video prediction with action control," in ICCV Workshops, Oct 2019.

[23] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, "Action recognition with dynamic image networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 12, pp. 2799–2813, 2018.

[24] Yogatama, D., Miao, Y., Melis, G., Ling, W., Kuncoro, A., Dyer, C., Blunsom, P.: Memory architectures in recurrent neural network language models. ICLR (2018)

[25] D.-A. Huang, V. Ramanathan, D. Mahajan, L. Torresani, M. Paluri, L. Fei-Fei, and J. Carlos Niebles, "What makes a video a video: Analyzing temporal information in video understanding models and datasets," in CVPR, June 2018.

[26] M. Zerkouk and B. Chikhaoui, ``Spatio-temporal abnormal behavior predictionin elderly persons using deep learning models," Sensors, vol. 20,no. 8, p. 2359, Apr. 2020.

[27] A. Hu, F. Cotter, N. Mohan, C. Gurau, and A. Kendall, “Probabilistic future prediction for video scene understanding,” arXiv:2003.06409, 2020.

[28] A.Bhattacharyya, M. Fritz, and B. Schiele, “Long-Term On-Board Prediction of People in Traffic Scenes Under Uncertainty,” in CVPR, 2018.

[29] K. Singh, S. Rajora, D. K. Vishwakarma, G. Tripathi, S. Kumar, andG. S. Walia, ``Crowd anomaly detection using aggregation of ensembles of Fine-tuned ConvNets," Neurocomputing, vol. 371, pp. 188-198, Jan. 2020.

[30] W. Sultani, C. Chen, and M. Shah, ``Real-world anomaly detection insurveillance videos," in Proc. IEEE/CVF Conf. Comput. Vis. PatternRecognit., Jun. 2018, pp. 6479-6488.

[31] T. Mahmud, M. Billah, M. Hasan and A. K. Roy-Chowdhury, "Pre-diction and description of near-future activities in video", Arxiv, vol. 1908.00943v3, 2020.

[32] W. Yu, Y. Lu, S. Easterbrook, and S. Fidler, “Efficient and information-preserving future frame prediction and beyond,” in ICLR, 2020.

[33] Majd M, Safabakhsh R. Correlational convolutional LSTM forhuman action recognition. Neurocomputing. 2020;396:224–9.

[34] O. Shouno, “Photo-realistic video prediction on natural videos of largely changing frames,” arXiv:2003.08635, 2020.

[35] B. Jin, Y. Hu, Q. Tang, J. Niu, Z. Shi, Y. Han, and X. Li, “Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction,” arXiv:2002.09905, 2020.

[36] D. Epstein, B. Chen and C. Vondrick, "Oops! predicting unintentional action in video", CVPR, pp. 919-929, 2020.