Moment Features based Violence Action Detection using Optical Flow

A F M Saifuddin Saif[1], Zainal Rasyid Mahayuddin[2]
Faculty of Science and Technology, American International University–Bangladesh, Dhaka, Bangladesh
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM, Bangi, Selangor, Malaysia

***Abstract*—Instantaneous detection of violence is still an unsolved research problem although artificial intelligence lives its prosperous years. The severity of injury causes due to violence can be minimized by detecting violence in real time demands for effective violence detection. Various methods were previously proposed for violence detection which could not provide robust results due many challenges, i.e. noise, motion estimation, lack of appropriate feature selection, lack of effective classification approach, complex background and variations in illumination. This research proposes an efficient method for violence detection using moment features to use motion patterns to facilitate detection in each frame and provides smaller area as region of interest. This means probability for extraction of motion intensity is getting lost because of same colored object in the background is reduced and thus minimizes background complexity. After that, proposed method uses optical flow to calculate angles and linear distances in each frame. In this context, if there is any frame loss due to noise or illumination variation, proposed method uses Kalman filter to process that frame by illuminating noise. Finally, decision for violence is determined using random forest classifier from single feature vector by generating a set of probabilities for each class. Proposed research performed extensive experimentation where accuracy rate of 99.12% was achieved using frame rate of 35 fps which is higher comparing with previous research results. Experimental results reveal the effectiveness of the proposed methodology.**

***Keywords*—Violence detection; feature extraction; classification; optical flow**

I. Introduction

Surveillance applications have been used to monitor public and private areas where intelligent violence detection is still an unsolved research problem. Surveillance system with violence detection capability can be used to monitor various places, i.e. airports [1], football matches and protests [2], parks [3], stadium [4], internet video filtration [5], markets [6], Government offices [7]. Previous researchers proposed various methods to provide intelligent violence detection which could not provide satisfactory results due to many challenges, i.e. lack of efficient feature extractions [8, 9], lack of appropriate segmentation method [10], motion estimation [11, 12], variation of brightness or illumination [13], cluttered backgrounds [14], scene complexities [15], low level image [16], lack of efficient noise reduction method [17], occlusions [18, 19]. Proposed method by this research used motion intensity characteristics by extracting moment features to facilitate violence detection in each frame effectively.

Till now many researchers have done their researches regarding violence detection. Research in [20] used convolutional neural network (C3D) and CNN with long short-term memory (CNN-LSTM) through shallow neural network to learn high level spatial-temporal information from raw image data for violence detection. However, their proposed method was suitable for still images. Research in [21] used set of selectively distributed frames and spatio-temporal features using both space and time dimensions provided to fully connected neural framework to classify violence or non-violence action which required further validation in terms with complexity. Research in [22] distinguished physical violence by designing DT–SVM (Decision Tree–SVM) two-layer classifier. However, their overall methodology requires further improvement towards more complex scenes with both nearby and distant objects. In in [23] modeled crowd dynamics using temporal summaries of grey level co-occurrence matrix (GLCM) features. However, improvement towards adaptive selection of optimal parameters based on given data was required in their proposed methodology.

This research performs moment features extraction to facilitate violence detection as the basis of motion pattern. In this context, linear distances and angles are calculated using optical flow. Besides, if there is any frame loss due to noise or illumination variation, proposed method uses Kalman filter to consider that frame for further processing by illuminating noise. Overall contributions by this research are stated below:

- Proposed method uses moment features by implicating weighted average of pictorial intensities to use motion patterns for facilitating detection in each frame in lieu with reducing background complexity and provides smaller area as region of interest to reduce overall computation time per frame.
- After calculating linear distances and angles as the basis of optical flow, proposed method uses Kalman filter to rectify frame loss due to nose or illumination variation which plays significant role to optimally estimate distance and angles for higher accuracy.
- As part of overall proposed methodology, this research uses random forest classifier to classify single feature type in order avoid complication like using multiple types of features causes lower processing time per frame comparing with previous research methods.

Rest of this paper is organized as follows. Section 2 demonstrates comprehensive and critical reviews in the

existing research, Section 3 illustrates proposed methodology for violence detection, Section 4 depicts extensive experimental validation for the proposed method and finally Section 5 presents concluding remarks.

II. Background Study

Various methods aiming to solve violence detection problem has been proposed mentioned in Fig. 1.

Research in [20] applied two deep neural networks (DNNs), i.e. 3D-based convolutional neural network (C3D) and CNN with long short-term memory (CNN-LSTM) for learning high level spatial-temporal information from raw image data. They combined features map achieved from C3D and CNN-LSTM through designing a shallow neural network. However, combination of features map from C3D and CNN-LSTM is suited for objet detection from still images causes high complexity in their overall research. In this context, estimation of computation time was not considered in their research during validation. Research in [21] illustrated an end to end deep neural network for violence detection using surveillance cameras. They extracted set of selectively distributed frames from video in lieu with passing spatio-temporal features to a fully connected neural network in order to classify violence or non violence action. Although, they created spatio-temporal features by performing features extraction using both space and time dimensions through a custom build convolutional neural network and long short term memory LSTM recurrent neural network, validation against computation time or processing time per frame was ignored in their research. Research in [22] selected some prior features to distinguish physical violence from daily-life activities. They designed DT–SVM (Decision Tree–SVM) two-layer classifier, i.e. first layer was acted as decision tree for using benefits of previously selected features and second layer was SVM classifier which used features for classification. However, for nearby and distant objects under complex scenarios, their research could not provide satisfactory results. In addition, their proposed approach provided significant misclassification for the frames with significant changes in light and shadow. Research in [23] proposed real time descriptor to model crowd dynamics by encoding variations in crowd texture by implicating temporal summaries of grey level co-occurrence matrix (GLCM) features. They measured inter-frame uniformity and illustrated that violent behavior varies in a less uniform manner. In addition, they performed discrimination between abnormal and normal scenes by generating scene description. However, adaptive selection of optimal parameters based on given data requires further improvement in their research. Research in [24] extracted key features, i.e. speed, direction, centroid and dimensions where they used Linear SVM to classify input video as violent or non-violent. Their proposed method considered two feature vectors, i.e. Local Binary Pattern (LBP) and Violent Flows (ViF). As Local Binary Pattern (LBP) or Violent Flows (ViF) takes less time for calculation separately than applying these feature vectors together, for this reason combination of Local Binary Pattern (LBP) and Violent Flows (ViF) in their research did not provide significant direction for future improvement. Research in [25] tested Bag-of-Words framework for detecting fight or violence by constructing a versatile and accurate fight detector using local descriptors. Although, they achieved encouraging accuracy rate, computation cost for extracting local descriptor is prohibitive for practical applications, particularly in surveillance and media rating systems. Research in [26] proposed a method using extreme acceleration pattern estimated by Radon transform to the power spectrum of consecutive frames. Their method assumed that kinematic cues that represent violent motion and strokes can be used to detect fights. In addition, they hypothesized if motion is considered as sufficient characteristics for recognition, in that case their overall methodology requires significant additional computation in lieu with confusing the detector. However, global motion estimation in their research did not seem to improve results significantly. Their proposed method required further perfection by approximating the Radon transform, which was the most time-consuming stage. Research in [27] proposed Oriented VIolent Flows (OViF) for feature extraction to take optimum benefits of motion magnitude variations in statistical motion orientations. In addition, they implicated feature combination and multiclassifier combination strategies. However, their proposed method could detect violence only in crowded scenes. Research in [28] proposed a method where corner joints of pictures are detected using Shi-Tomasi corner detection algorithm. They used optical flow parameter which was calculated using Lucas-Kanade pyramid optical flow algorithm for violence detection. However, for discontinuous and fast motion their proposed method did not provide robust performance.

Proposed method by this research calculated distance and angles as the basis of optical flow. In addition, proposed method used Kalman filter to handle illumination variation in case of any frame loss due to noise. Besides, motion pattern based on moment features extraction is used to reduce background complexity in case of similar colored objects by implicating weighted average of pictorial intensities.

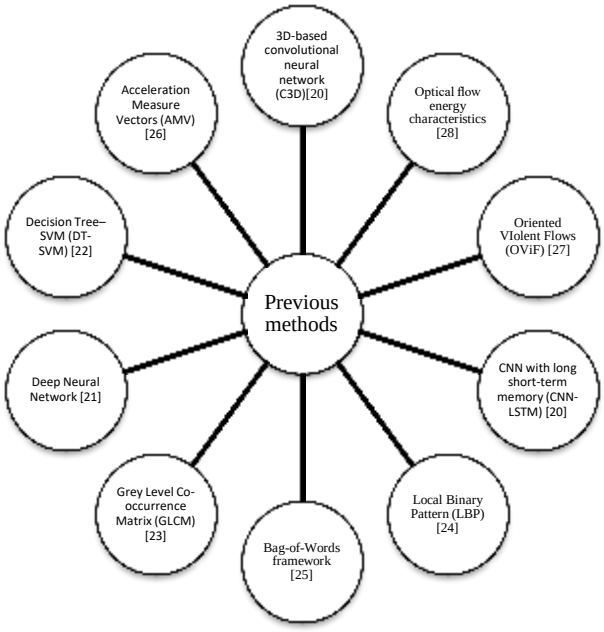

Fig. 1. Existing Methods in Previous Research for Violence Detection.

III. Proposed Methodology

This research follows nontracking based action recognition due to achieve computationally sound and effective violence action. Proposed research uses object motion to identify the event whether violent or non violent. This research uses moment features in lieu with optical flow to calculate linear distances and angles which works in pixel intensities of objects. Overall proposed method is depicted in Fig. 2.

A. Input Image and Preprocessing

Preprocessing is essential part to extract features efficiently to classify violence action. This step is the crucial as any inconsistency arising from it may lead to a misclassification of violence action. Proposed research uses monocular video camera using 35 fps frame rate for input video frames collection. Proposed research uses median filter [29, 30] to remove noise from the collected video frames. In this context, morphological processing such as resizing of the frame into 300x250 dimension, erosion and dilation are applied to ensure noise free frames to the next subsequent frames. In addition, this research also uses two frame differential approaches to find the difference between frames for finding the initial change between consecutive frames.

B. Moment Features Extraction

Proposed method extracted moment features from median filtered image. If m and n are the co-ordinates of median filtered frame $Z_f\ (m,n)$, raw moments of $Z_f\ (m,n)$ for order $(i+j)$ is defined as in (1).

$$R_{ij} = \sum\sum m^i y^j Z_f(m,n) \quad (1)$$

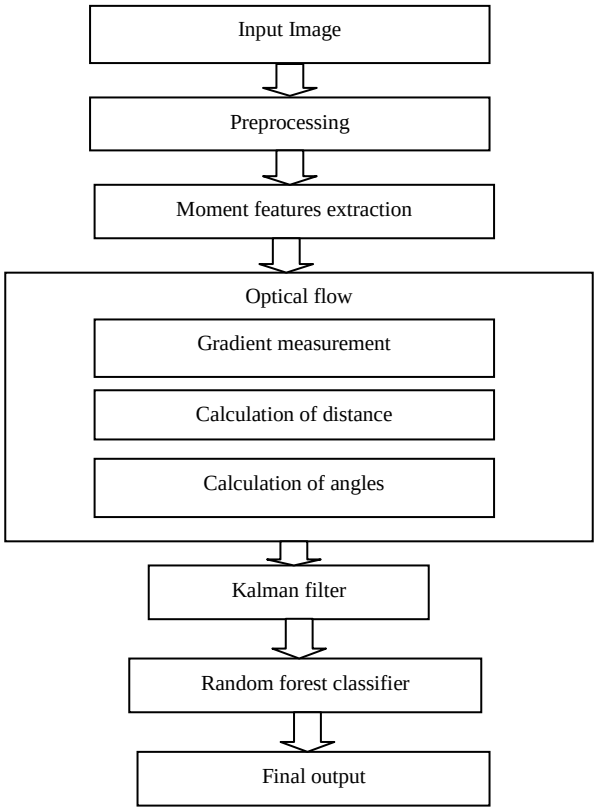

Fig. 2. Proposed Method for Violence Detection.

When considering $I_f\ (x,y)$ as 2D continuous function, (1) can be expressed as (2).

$$R_{ij} = \int\int m^i n^j Z_f(m,n) \quad (2)$$

Here, R_{ij} is denoted as raw moments to be used for calculating distances and angles as the basis of optical flow for further processing.

C. Calculation of Distance and Angle

Proposed method determines motion pattern among consecutive frames from the extracted moment features using two frame differential approach where angles and linear distances are calculated for each frame of input video. Parameters like pyramid scale, levels, window size, iteration are used to calculate motion. Pyramid scale of 0.5 is used as classical pyramid where each next layer is twice smaller than previous one. Value of levels 5 is used as mean number of pyramid layers including the initial frame. Proposed method uses larger window size of 20 to improve efficiency of the classification in terms with detecting motion. This research uses determined window around foreground object which limits the scope of motion segmentation to a smaller area. This means that the probability of the extraction of motion is getting lost because a similar colored object in the background is reduced in lieu with improving processing speed.

Proposed method calculates combination of distance and angles as the basis of optical flow for each frame of input video to classify whether action is violent or not. Distances are categorized for every pixel in the consecutive frame by 60° interval followed by summation of all distances. Proposed method calculates image gradients in horizontal and vertical direction in lieu with gradient along time. Proposed method uses optical flow to calculate angles and linear distance for each frame of input video. From the calculated angles, proposed method categorizes linear distances which generate oriented histogram of 6 bins and equally divided by 60 degree. Distance and angles are calculated as in (3) and (4).

$$K=\sqrt{dm^2 + dn^2} \quad (3)$$

$$L=\tan^{-1}\left(\frac{dn}{dm}\right) \quad (4)$$

Here, K is denoted as distance, angle is denoted by L, dm denotes distance change in horizontal direction and dn denotes distance change in vertical direction. Distance and angle is achieved for certain pixel position denoted as (m_1, n_1) of (i-1)th frame and same pixel position denoted as (m_2, ny_2) of ith frame. Pythagoras theorem [31] is used to calculate angles and linear distances by the proposed method where by angle L changes of degree is defined from (i-1)th frame to ith frame in every pixel. Linear distances of pixels between (i-1)th and frame nth frame are defined by K. Finally, proposed method uses Random forest classifier to classify whether the video scene contains violence or not.

D. Kalman Filter

Kalman filter is used to provide the best estimate of states in the presence of noise. Proposed method uses Kalman filter to optimally estimate distances and angles for higher accuracy rate. During distances and angles calculation, if there is any

frame loss due to noise or illumination variation then Kalman filter is used by the proposed method to process that frame by illuminating noise. Although, median filter was applied during preprocessing step, some frames can be often still noise may cause deviation in performance.

E. Classification

Finally, decision for violence is determined using random forest classifier from single feature vector by generating a set of probabilities for each class. In this context, probabilities are estimated using mean predicted class probabilities of the trees in the forest where class probability of a single tree is the fraction of samples of the same class in the tree. Class with highest probability is the one that is assigned to the frame as the "decision". In this context, ratio of the highest probability to the second highest probability is referred to as "confidence" of the decision. In this regard, proposed method by this research used adaptive threshold using (5) [32, 33]. Any decision with confidence more than threshold is considered as "violence" and others are "non violence".

$$T = V - \frac{V \times (log_2(I)+1)}{100} \quad (5)$$

Here, total number of frames is denoted as I, mean value of pictorial intensities is denoted as V in a video frame. Threshold value is denoted as T.

IV. Experimental Results and Discussion

A. Hardware and Software Set Up

C# programming language [34] is used to validate the proposed method by this research. Core i7 processor with 8 Gigabyte RAM is used for experimentation. Various performance metrics are used to validate the proposed methodology, i.e. accuracy rate [22, 35, 36], error [27, 37, 38], computation time [28, 39, 40] and frame rate [20, 41, 42, 43, 44].

B. Datasets

Proposed method is validated using Hockey Fight and Movies datasets [24, 25, 38, 26, 27]. Total 1000 videos are taken from Hockey fight dataset where 500 videos contain violent sequence and rest 500 videos contain non-violent video sequence. Besides, total 200 videos are taken from Movies datasets where 100 videos contain violent sequence and rest 100 videos contain non-violent video sequence. Whole datasets are divided into five sets for five cross validation. For Hockey Fight dataset and Movies datasets, resolution of frames is fixed to 300X250.

C. Experimental Results

Proposed method received accuracy rate of 99.12% for Movies dataset and 94.82% for Hockey Fight dataset shown in Table I. Accuracy rate achieved by the proposed method for Hockey Fight dataset is lower comparing with Movies dataset as there are some ambiguities in Hockey Fight dataset to distinguish that whether two players are playing or fighting with each other. Proposed method received error rate of 0.88% for Movies dataset and 5.18% for Hockey Fight dataset indicating that Hockey Fight dataset causes more error than Movies dataset. Proposed method required computation time of 0.0010 second for Movies dataset and 0.0020 sec second for Hockey Fight datasets indicating that for Hockey Fight dataset proposed method required more computation time due to bigger size of Hockey Fight dataset in lieu with containing more complex scenarios than Movies dataset.

D. Comparison with Previous Research Results

Research in [20] received accuracy rate of 63% using 3D-based convolutional neural network (C3D) and 61% using CNN with long short-term memory (CNN-LSTM) by using frame rate of 25 fps. Due to the additional step such as design of shallow neural network by combining features map obtained from C3D and CNN-LSTM can lead to increase computation complexity for the overall methodology. Research in [21] received accuracy rate of 94.5% using spatio-temporal features and passing them to a fully connected neural framework to classify the video to violence or non-violence action. Although, they performed feature extraction using both space and time dimensions to create spatio-temporal features through a custom build convolutional neural network, estimation of processing time per frame was ignored in their research. Research in [22] received accuracy rate of 97.6% by designing DT-SVM (Decision Tree SVM) using prior determined features to distinguish physical violence from daily-life activities. However, in case of complex scenes with both nearby and distant objects, their research could not provide satisfactory validation results. Research in [23], received accuracy rate of 90.5% using frame rate of 30 fps by implicating crowd dynamics with the use of encoding changes in crowd texture and temporal summaries of grey level co-occurrence matrix (GLCM) features. However, their research requires adaptive method for choosing optimal parameters based on given data. Research in [24] tested their proposed method in non-crowded and crowded scenarios to verify the effectiveness of Local Binary Pattern (LBP) features. They received accuracy rate of 89.1% with error rate of 10.9%. However, they did not validate their method based on computation time and frame rate. Research in [25] received approximate accuracy rate of 90% by using popular bag-of-words approach which can accurately recognize fight sequences. However, computational cost of extracting features in their research is not encouraging for practical applications. Research in [27] received accuracy of 88% using statistical motion orientation information. However, they received error rate of 12% as their proposed Oriented VIolent Flows (OViF) could detect violence in crowded scenarios only which demands more robust validation. Research in [28] received accuracy rate of 72% using histogram of the computed optical flow energy values. However, they received high error rate of 28% due to inability of their method to perform under discontinuous and fast motion. Research in [26] received accuracy of 98.9% using extreme acceleration patterns as the main feature where their required processing time was 0.0419 second. However, further perfection was required in their method by approximating Radon transform, which is the most time-consuming stage. Among all these previous methods stated above, proposed method by this research received higher accuracy rate of 99.12% with lower required computation time per frame of 0.0010 second and low error rate of 0.88% using 35 fps frame rate mentioned in Table II, Table III, Table IV and Table V.

TABLE I. ACCURACY, ERROR RATE AND COMPUTATION TIME FOR THE PROPOSED METHOD

Datasets	Accuracy	Error Rate	Computation Time	Frame Rate
Movies Dataset [25, 38, 26]	99.12%	0.88%	~ 0.0010 sec	35fps
Hockey Fight Dataset [24,25, 38, 26, 27]	94.82%	5.18%	~ 0.0020 sec	

TABLE II. COMPARISON WITH PREVIOUS METHODS BASED ON ACCURACY

Previous Methods	Accuracy
Proposed method	99.12%
C3D [20]	63%
CNN-LSTM [20]	61%
DNN [21]	94.5%
DT-SVM [22]	97.6%
GLCM [23]	90.5%
Local Binary Pattern (LBP) [24]	89.1%
Bag-of-Words framework [25]	90%
Oriented VIolent Flows (OViF) [27]	88%
Optical flow energy characteristics [28]	72%
Acceleration Measure Vectors (AMV) [26]	98.9%

TABLE III. COMPARISON WITH PREVIOUS METHODS BASED ON ERROR RATE

Methods	Error Rate
Proposed method	0.88%
Local Binary Pattern (LBP) [24]	10.9%
Oriented VIolent Flows (OViF) [27]	12%
Optical flow energy characteristics [28]	28%
Acceleration Measure Vectors (AMV) [26]	1.1%

TABLE IV. COMPARISON WITH PREVIOUS METHODS BASED ON COMPUTATION TIME

Previous Methods	Computation time
Proposed method	~ 0.0010 sec
Acceleration Measure Vectors (AMV) [26]	0.0419sec
Optical flow energy characteristics [28]	0.025 sec

TABLE V. COMPARISON WITH PREVIOUS METHODS BASED ON FRAME RATE

Previous Methods	Frame rate
Proposed method	35 fps
C3D [20]	25 fps
CNN-LSTM [20]	25 fps
GLCM [23]	30 fps

E. Analysis and Discussion

Research in [20] utilized 3D-based convolutional neural network (C3D) and CNN with long short-term memory (CNN-LSTM). For C3D and CNN-LSTM they achieved accuracy rate of 63% and 61% respectively using 25 fps frame rate. Although, their accuracy was not promising, usages of two deep neural networks (DNNs) were robust on learning high level spatial-temporal information from raw image data. They combined features maps obtained from C3D and CNN-LSTM networks by designing shallow neural network which acted as third scenario in their research demands for further validation to establish their overall research in terms with computational time. Proposed method by this research estimates approximate computation time of 0.0010 sec per frame in lieu with accuracy rate of 99.12% using moment features instead of combing other feature measurement indicates better validation performance than research in [20]. Research in [21] received accuracy rate of 94.5% by extracting set of selectively distributed frames of the video clip and passing spatio-temporal features to a fully connected neural architecture in order to classify the video as violence or non-violence action. However, due to usage of a custom build convolutional neural network and long short term memory LSTM recurrent neural network to process spatio-temporal features based on space and time dimensions for feature extractions, validation against computation time or processing time was totally ignored in their research. In this context, proposed method by this research uses moments feature to extract motion characteristics for classifying violent characteristics and estimates processing time per frame to validate the overall proposed methodology. Research in [22] received accuracy rate of 97.6% where they designed DT–SVM (Decision Tree–SVM) two-layer classifier, i.e. first layer was a decision tree to take benefits of prior determined features, second layer was SVM classifier to use features for classification. However, frames with significant variations in light and shadow caused misclassification in their research. In addition, their research requires further investigation in case of nearby and distant objects for complex scenarios. Proposed method by this research is validated using sufficient datasets comparing with research in [22] in lieu with that proposed method uses random forest classifier to classify single feature type in order to avoid complication like using multiple types of features causes lower processing time per frame comparing with previous research methods. Research in [23] received accuracy rate of 90.5% using 30 fps frame rate by introducing measurement of inter-frame uniformity in lieu with demonstrating violent behavior changes in a less uniform manner. Although, their proposed method performed discrimination between abnormal and normal scenes, in case of choosing optimal parameters based on given data initiates the need of adaptive method to be integrated with their overall proposed methodology which will surely demand for further validation. Proposed method by this research uses adaptive threshold during classification in lieu with extracting single feature type initially remedies the need of choosing additional parameters to achieve optimal performance like in research [23] causes gaining better performance in terms with accuracy and frame rate. Research in [24] achieved accuracy rate of 89.1% using Linear SVM to classify video as violent or non-

violent. They received error rate of 10.9% which indicates lower performance comparing with the proposed method by this research. In their research, Local Binary Pattern (LBP) or Violent Flows (ViF) takes less time for calculation than applying Local Binary Pattern (LBP) and Violent Flows (ViF) together. However, need of applying Local Binary Pattern (LBP) and Violent Flows (ViF) together instead of applying them separately did not provide any future direction for improvement in terms with performance by their research. In addition, their proposed method was not validated based computation time to indicate efficiency in terms with processing duration per frame. In this regard, proposed method by this research received higher accuracy rate of 99.12% and error rate of 0.88% using random forest classifier from single feature vector by generating a set of probabilities for each class indicates higher efficiency that research in [24]. Research in [25] achieved accuracy rate of 90% by constructing versatile and accurate fight detector using local features descriptors. Although, their accuracy rate was impressive, computational cost to construct local features vectors was impractical particularly in surveillance and media rating system. Proposed method by this research received accuracy rate of 99.12% by using moment features as means of the pixel intensity distribution to limit segmentation tasks in lieu with reducing computational complexity. Oriented VIolent Flows (OViF) by research in [27] received accuracy rate of 88% by taking full advantage of motion magnitude change information in statistical motion orientations. They received accuracy rate of 12% due to adaptation of features combination and multiclassifier strategies. In addition, their method was applicable only for crowded scenarios. In this context, proposed method by this research achieved accuracy rate of 99.12% with low error rate of 0.88% by measuring distance and angles in horizontal and vertical direction as the basis of optical flow strategy indicates better performance than research in [27]. Research in [26] received accuracy rate of 98.9% due to efficient estimation of extreme acceleration patterns as the main features by implicating Radon transform to the power spectrum of consecutive frames causes low error rate of 1.1% with required computation time of 0.0419second. Although, they hypothesized that motion was sufficient for recognition, global motion estimation experiments did not seem to improve results significantly. In addition, their proposed method needs further investigation to estimate relative importance of motion and appearance in formation for the recognition of violence or nonviolence actions. Proposed method by this research considers motion estimation using two frame differential approach and moments feature extraction where accuracy rate of 99.12% was achieved using 35 fps and error rate of 0.88% were received using 0.0010 second computation time per frame indicates better performance than in research [26] shown in Fig. 3, Fig. 4, Fig. 5 and Fig. 6.

Research in [28] received accuracy rate of 72% using Shi-Tomasi corner detection and histogram of the computed optical flow energy values. They received error rate of 28% using 0.025 second due to the usage of corner. Although, usage of corner features increases computation time in most of the research in computer vision domain, one of their significant achievement was that their method reduced time cost.

However, in case of discontinuous and fast motion, their method was not robust. Proposed method by this research showed better performance than in research [28] using moment features as the basis of uncertainty measurement of pictorial intensities to use motion pattern and later random forest classifier was used to classy violence and nonviolence behavior.

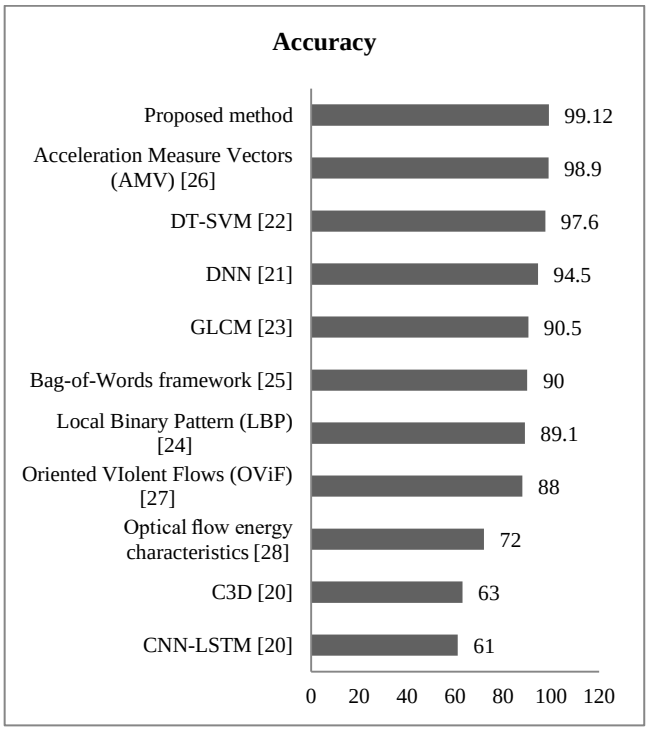

Fig. 3. Comparison among Proposed Method and Previous Research based on Accuracy.

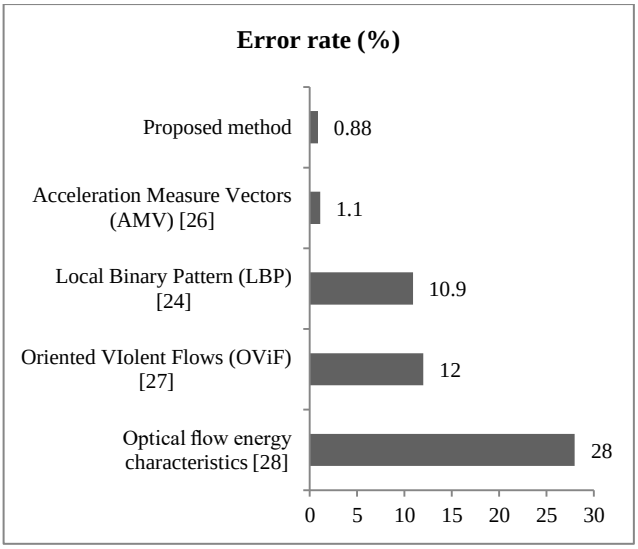

Fig. 4. Comparison among Proposed Method and Previous Research based on Error Rate.

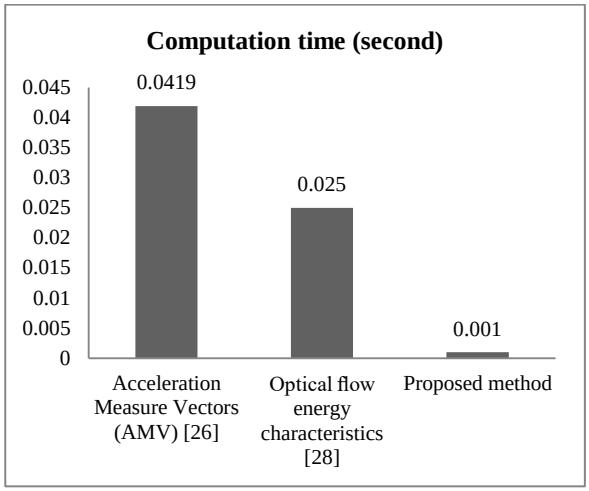

Fig. 5. Comparison among Proposed Method and Previous Research based on Computation Time Per Frame.

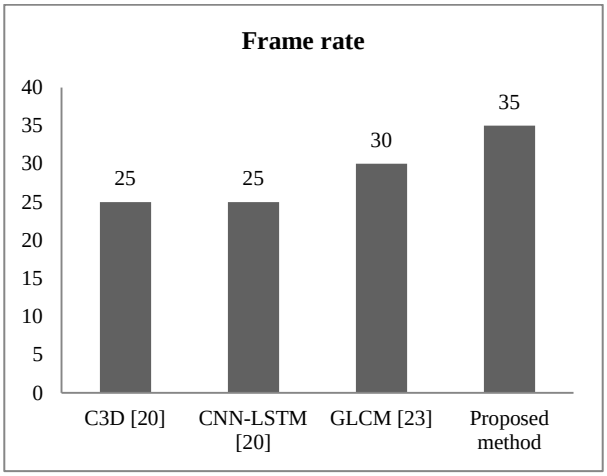

Fig. 6. Comparison among Proposed Method and Previous Research based on Frame Rate.

V. Conclusion

Proposed method used motion patterns by extracting moment features for uncertainty measurement of pictorial intensity distribution based on efficient scene interpretation. In case of similar colored object in the background, moment properties provide certain particular weighted average of pictorial intensities causes attractive interpretation which played vital role to minimize background complexity. After that, linear distances and angles are calculated for each frame as the basis of optical flow followed by Kalman filter to rectify frame loss due to noise or illumination variation which plays significant role to optimally estimate distance and angles for higher accuracy rate. Finally, proposed method used random forest classifier to classify single feature type in order to avoid complication like using multiple types of features causes lower processing time comparing with previous research methods. Experimental results for the proposed method reveal higher efficiency comparing with previous research results in terms with accuracy rate, computation time and frame rate. Proposed method achieved maximum accuracy of 99.12% using frame rate of 35 fps where required computation time per frame was 0.0010 sec. Performance of the proposed method reveals the potentiality to provide significant capability to surveillance applications for monitoring violence efficiently and reduce the impact of violence related injuries. In future, this research intends to be involved more complex activities, i.e. recognition of violence for distant objects and improvement of recognition performance for the misclassified samples.

Acknowledgment

The authors would like to thank Universiti Kebangsaan Malaysia for providing financial support under the Geran Galakan Penyelidikan research grant, GGP-2017-030.

References

[1] Jain and D. K. Vishwakarma, "State-of-the-arts Violence Detection using ConvNets," in 2020 International Conference on Communication and Signal Processing (ICCSP), 2020, pp. 0813-0817.

[2] K. Gkountakos, K. Ioannidis, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, "A Crowd Analysis Framework for Detecting Violence Scenes," in Proceedings of the 2020 International Conference on Multimedia Retrieval, 2020, pp. 276-280.

[3] R. Halder, "Discrete Wavelet Transform for CNN-BiLSTM-based Violence Detection."

[4] E. Fenil, G. Manogaran, G. Vivekananda, T. Thanjaivadivel, S. Jeeva, and A. Ahilan, "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM," Computer Networks, vol. 151, pp. 191-200, 2019.

[5] J. Li, X. Jiang, T. Sun, and K. Xu, "Efficient violence detection using 3d convolutional neural networks," in 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2019, pp. 1-8.

[6] M. Ramzan, A. Abid, H. U. Khan, S. M. Awan, A. Ismail, M. Ahmed, M. Ilyas, and A. Mahmood, "A review on state-of-the-art violence detection techniques," IEEE Access, vol. 7, pp. 107560-107575, 2019.

[7] K. Singh, K. Y. Preethi, K. V. Sai, and C. N. Modi, "Designing an Efficient Framework for Violence Detection in Sensitive Areas using Computer Vision and Machine Learning Techniques," in 2018 Tenth International Conference on Advanced Computing (ICoAC), 2018, pp. 74-79.

[8] M. Sharma and R. Baghel, "Video Surveillance for Violence Detection Using Deep Learning," in Advances in Data Science and Management, ed: Springer, 2020, pp. 411-420.

[9] A. S. Saif, A. S. Prabuwono, and Z. R. Mahayuddin, "Moment feature based fast feature extraction algorithm for moving object detection using aerial images," PloS one, vol. 10, p. e0126212, 2015.

[10] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violence detection in surveillance video using low-level features," PloS one, vol. 13, p. e0203668, 2018.

[11] J. Ha, J. Park, H. Kim, H. Park, and J. Paik, "Violence detection for video surveillance system using irregular motion information," in 2018 International Conference on Electronics, Information, and Communication (ICEIC), 2018, pp. 1-3.

[12] J. Huang, G. Li, N. Li, R. Wang, and W. Wang, "A violence detection approach based on spatio-temporal hypergraph transition," in International Conference on Computer Analysis of Images and Patterns, 2017, pp. 218-229.

[13] A. Saif, A. S. Prabuwono, and Z. R. Mahayuddin, "Moving object detection using dynamic motion modelling from UAV aerial images," The Scientific World Journal, vol. 2014, 2014.

[14] S. Roshan, G. Srivathsan, K. Deepak, and S. Chandrakala, "Violence Detection in Automated Video Surveillance: Recent Trends and Comparative Studies," in The Cognitive Approach in Cloud Computing and Internet of Things Technologies for Surveillance Tracking Systems, ed: Elsevier, 2020, pp. 157-171.

[15] A. Traoré and M. A. Akhloufi, "2D Bidirectional Gated Recurrent Unit Convolutional Neural Networks for End-to-End Violence Detection in

Videos," in International Conference on Image Analysis and Recognition, 2020, pp. 152-160.

[16] T. Zhang, W. Jia, X. He, and J. Yang, "Discriminative dictionary learning with motion weber local descriptor for violence detection," IEEE transactions on circuits and systems for video technology, vol. 27, pp. 696-709, 2016.

[17] Q. Zhou, C. Wu, J. Xing, J. Li, Z. Yang, and Q. Yang, "Wi-Dog: monitoring school violence with commodity WiFi devices," in International Conference on Wireless Algorithms, Systems, and Applications, 2017, pp. 47-59.

[18] K. Deepak, L. Vignesh, and S. Chandrakala, "Autocorrelation of gradients based violence detection in surveillance videos," ICT Express, vol. 6, pp. 155-159, 2020.

[19] T. Zhang, W. Jia, B. Yang, J. Yang, X. He, and Z. Zheng, "MoWLD: a robust motion image descriptor for violence detection," Multimedia Tools and Applications, vol. 76, pp. 1419-1438, 2017.

[20] B. Peixoto, B. Lavi, J. P. P. Martin, S. Avila, Z. Dias, and A. Rocha, "Toward subjective violence detection in videos," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 8276-8280.

[21] M. M. Moaaz and E. H. Mohamed, "Violence Detection In Surveillance Videos Using Deep Learning," الحاسبات فى المعلوماتية الذشرة والمعلومات, vol. 2, pp. 1-6, 2020.

[22] L. Ye, L. Wang, H. Ferdinando, T. Seppänen, and E. Alasaarela, "A Video-Based DT–SVM School Violence Detecting Algorithm," Sensors, vol. 20, p. 2018, 2020.

[23] K. Lloyd, P. L. Rosin, D. Marshall, and S. C. Moore, "Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (GLCM)-based texture measures," Machine Vision and Applications, vol. 28, pp. 361-371, 2017.

[24] P. Vashistha, C. Bhatnagar, and M. A. Khan, "An architecture to identify violence in video surveillance system using ViF and LBP," in 2018 4th International Conference on Recent Advances in Information Technology (RAIT), 2018, pp. 1-6.

[25] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in International conference on Computer analysis of images and patterns, 2011, pp. 332-339.

[26] O. Deniz, I. Serrano, G. Bueno, and T.-K. Kim, "Fast violence detection in video," in 2014 international conference on computer vision theory and applications (VISAPP), 2014, pp. 478-485.

[27] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented violent flows," Image and vision computing, vol. 48, pp. 37-41, 2016.

[28] Z. Guo, F. Wu, H. Chen, J. Yuan, and C. Cai, "Pedestrian violence detection based on optical flow energy characteristics," in 2017 4th International Conference on Systems and Informatics (ICSAI), 2017, pp. 1261-1265.

[29] Z. R. Mahayuddin and A. F. M. S. Saif, "A Comprehensive Review Towards Segmentation and Detection of Cancer Cell and Tumor for Dynamic 3D Reconstruction," Asia-Pacific Journal of Information Technology and Multimedia, vol. 9, pp. 28-39, 2020.

[30] Z. R. Mahayuddin and A. S. Saif, "A Comprehensive Review Towards Appropriate Feature Selection for Moving Object Detection Using Aerial Images," in International Visual Informatics Conference, 2019, pp. 227-236.

[31] J. D. Sally, Roots to research: a vertical development of mathematical problems: American Mathematical Soc., 2007.

[32] A. S. Saif, A. S. Prabuwono, Z. R. Mahayuddin, and T. Mantoro, "Vision-based human face recognition using extended principal component analysis," International Journal of Mobile Computing and Multimedia Communications (IJMCMC), vol. 5, pp. 82-94, 2013.

[33] Z. R. Mahayuddin and A. S. Saif, "A Comparative Study Of Three Corner Feature Based Moving Object Detection Using Aerial Images," Malaysian Journal of Computer Science, pp. 25-33, 2019.

[34] A. S. Saif, A. S. Prabuwono, Z. R. Mahayuddin, and H. T. Himawan, "A review of machine vision based on moving objects: object detection from UAV aerial images," International Journal of Advancements in Computing Technology, vol. 5, p. 57, 2013.

[35] Z. R. Mahayuddin and A. F. M. S. Saif, "Efficient Hand Gesture Recognition Using Modified Extrusion Method based on Augmented Reality," TEST Engineering and Management, vol. 83, pp. 4020-4027, 2020.

[36] Z. R. Mahayuddin and A. F. M. S. Saif, "Augmented Reality Based Ar Alphabets Towards Improved Learning Process In Primary Education System," Journal Of Critical ReviewS, vol. 7, 2020.

[37] Z. R. Mahayuddin, A. S. Saif, and A. S. Prabuwono, "Efficiency measurement of various denoise techniques for moving object detection using aerial images," in 2015 International Conference on Electrical Engineering and Informatics (ICEEI), 2015, pp. 161-165.

[38] A. Saif and Z. R. Mahayuddin, "Moving Object Segmentation Using Various Features from Aerial Images: A Review," Advanced Science Letters, vol. 24, pp. 961-965, 2018.

[39] A. Saif, A. Prabuwono, and Z. Mahayuddin, "Adaptive long term motion pattern analysis for moving object detection using UAV aerial images," International Journal of Information System and Engineering, vol. 1, pp. 50-59, 2013.

[40] A. S. Saif, A. S. Prabuwono, and Z. R. Mahayuddin, "Adaptive motion pattern analysis for machine vision based moving detection from UAV aerial images," in International Visual Informatics Conference, 2013, pp. 104-114.

[41] A. S. Saif, A. S. Prabuwono, and Z. R. Mahayuddin, "Real time vision based object detection from UAV aerial images: a conceptual framework," in FIRA RoboWorld Congress, 2013, pp. 265-274.

[42] A. S. Saif, A. S. Prabuwono, and Z. R. Mahayuddin, "Motion analysis for moving object detection from UAV aerial images: A review," in 2014 International Conference on Informatics, Electronics & Vision (ICIEV), 2014, pp. 1-6.

[43] A. S. Saif and Z. R. Mahayuddin, "Vehicle Detection for Collision Avoidance Using Vision based Approach: A Constructive Review," Solid State Technology, pp. 2861-2869, 2020.