I.J. Image, Graphics and Signal Processing, **2024, 3, 52-63**
Published Online on June 8, 2024 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2024.03.05

Human Abnormal Activity Recognition from Video Using Motion Tracking

Manoj Kumar*
JSS Academy of Technical Education, NOIDA, India
National Institute of Technology, Kurukshetra, India
Email: manojkumar@jssaten.ac.in
ORCID iD: https://orcid.org/0000-0001-5259-4208
*Corresponding Author

Anoop Kumar Patel
National Institute of Technology, Kurukshetra, India
Email: akp@nitkkr.ac.in
ORCID iD: https://orcid.org/0000-0002-3249-3888

Mantosh Biswas
National Institute of Technology, Kurukshetra, India
Email: mantoshbiswas@nitkkr.ac.in
ORCID iD: https://orcid.org/0000-0001-9027-4432

Sandeep Singh Sengar
Department of Computer Science, Cardiff Metropolitan University, Cardiff, United Kingdom, CF5 2YB
Email: SSSengar@cardiffmet.ac.uk
ORCID iD: https://orcid.org/0000-0003-2171-9332

Received: 18 May, 2023; Revised: 17 June, 2023; Accepted: 12 August, 2023; Published: 08 June, 2024

Abstract: The detection of violent behavior in the public environment using video content has become increasingly important in recent years due to the rise of violent incidents and the ease of sharing and disseminating video content through social media platforms. Efficient and effective techniques for detecting violent behavior in video content can assist authorities with identifying potential hazards, preventing crimes, and promoting public safety. Violence detection can also help to mitigate the psychological damage caused by viewing violent content, particularly in vulnerable populations such as infants and victims of violence. We have proposed an algorithm to calculate new descriptors using the magnitude and orientation of optical flow (MOOF) in the video. Descriptors are extracted from MOOF based on four binary histograms each by applying various weighted thresholds. These descriptors are used to train Support Vector Machine (SVM) and classify the video as violent or nonviolent. The proposed algorithm has been tested on the publicly available Hockey Fight Dataset and Violent Flow dataset. The results demonstrate that the proposed descriptors outperform the state-of-the-art algorithms with an accuracy of 91.5% and 78.5% on the Hockey Fight and Violent Flow datasets, respectively.

Index Terms: Velocity, Orientation, Classifier, Thresholding, Optical Flow, Suspicious Activity, Adaptive thresholding, SVM.

1. Introduction

In public spaces, such as retail malls, train stations, airports, and so on, numerous security cameras have been deployed. It would take too much time and effort for a human to watch every video stream at the monitor center. To avoid danger, automatic identification of odd events in real-time is helpful. For instance, activities such as fighting, theft, burglary, and vandalism can activate immediate alerts to security personnel, allowing for prompt intervention and incident prevention. This framework improves the efficacy of security systems and contributes to public safety in areas such as airports, railway stations, retail centers, and public parks. This model can aid in crowd management and event security during large-scale events or gatherings by monitoring crowd movements and detecting potentially hazardous

situations, such as overcrowding or aggressive behavior. Security personnel are able to proactively prevent stampedes and altercations. This approach contributes to the successful organization of public events by ensuring the safety and well-being of participants. The objective of the proposed model is to develop a reliable algorithm capable of analyzing video data in real-time and providing immediate alarms or notifications when anomalous activities are detected. The proposed algorithm can automatically recognize and categorize violent and combative behavior exhibited by individuals. By detecting such anomalies, it is possible to take the necessary precautions against potential threats and maintain a secure environment. A useful method for motion identification and tracking is optical flow, which characterizes the synchronized motion of objects [1–3]. It is widely used in the fields of motion analysis and tracking. Optical flow is a promising theory because it provides a substantial displacement between consecutive frames. Most common methods[4, 5] assume that greyscale images are consistent, and that motion between neighboring pixels is smooth. A further approach, proposed by Sun et al.[6] to overcome these restrictions, including training a probabilistic model for optical flow estimations. However, this approach needs the incorporation of optical flow ground realities, which can be challenging to obtain for training purposes. A data term including gradient and grey value limitations, along with an additional smoothness term is commonly used in the standard optical flow model. However, most traditional formulations of optical flow only consider global uniformity and do not include methods that preserve local picture features [7]. Many optical flow methods attempt to solve problems only when the scene's motion is mostly fixed. One contemporary challenge that sheds light on this phenomenon is the matching of 3D vibrant facial sequences with very non-rigid changes [8, 9].To solve the issue, a group of photos must be loosely aligned to a reference, such as an expressionless face. A 3D mesh that corresponds to each image known as a UV map and has a different vertex topology is also included. Vertex correspondence can be imposed once the Ultra Violet maps have been registered to an original image. Construction of 3D morphable models frequently uses this method[10]. Mesh correspondence is approached in a somewhat different way by Beeler et al. [11] and Bradley et al.[12] in their solutions, a starting frame through a 3D series is deformed using a reference mesh by computing image displacement from camera views. The mesh eliminates artifacts, a constraint that prevents mesh faces from getting reversed or inverted, and the optical flow gives guidelines for altering pixel placements. Research is also being done on mesh and picture distortion in graphics. Such methods offer adaptable ways to introduce deformation while maintaining some desired qualities, including local geometrical specifics. As a result, it is also interesting to apply smoothness restrictions and other similar solutions to optical flow calculations, which serve as the foundation for the work mentioned. A hybrid optical flow framework[13] is presented that incorporates both a global smoothness term and a Laplacian mesh data component. Their energy, however, is extremely nonlinear and challenging to reduce. Frames as a whole and optical flow data are taken into account to accurately depict motion to detect human activity in global methods. A Violent Flow (ViF) descriptor, for instance, was proposed by Hassner et al.[14]; it takes into account the amount of the shift in optical flow between two consequent frames. The SVM classifier is then used to categorize the video based on the ViF descriptor as either normal or violent. Since the orientation of the optical flow isn't taken into account when describing motion in ViF, some techniques combine the two to create a more accurate descriptor of ViF. A new metric called Oriented Violent Flow (OViF) was proposed by Gao et al.[15]. The objective is to take advantage of both the direction and intensity of the optical flow, but it has a long computation time. This restricts the practical application of the algorithm in situations where numerous video segments must be processed in real scenarios addition, OViF fails to identify orientation changes between two pixels $p_t(x, y), p_{t+1}(x, y)$ with identical optical flow intensities ut may identify unique optical flow orientations when the pixels are placed in the same grid. Here t represents the frame index, while (x, y) represents the location of the pixel.

We plan to surmount forthcoming obstacles. To implement an effective system, it is necessary to relax the aforementioned restrictions. We proposed an algorithm that pre-processes the frame and calculates the optical flow to trace the motion of humans in a video by calculating the magnitude and orientation of the human in each successive frame. We derive descriptors from eight binary histograms by applying various weighted thresholds on the MOOF.
The main contribution of the research is as follows:

(a) Preprocessing of the video frames.
(b) Calculation of the MOOF of successive frames of video.
(c) Descriptors extracted from magnitude and orientation of optical flow based on four binary histograms on each by applying various weighted thresholds.
(d) These descriptors are used to train SVM which classifies whether a video is normal or abnormal.

In this work, we break down the framework and implementation for anomalous activity recognition across five sections. In section 2, we'll go over the literature review. Section 3 examines the suggested design, whereas Section 4 reflects the actual execution of the research and discusses all of its features and functionalities. The final section 5 provides the conclusions and discusses what comes next.

2. Literature Review

In the discipline of computer vision, aberrant activity detection is a difficult topic, requiring statistical approaches, machine learning, and Deep Learning techniques to identify abnormal human actions in video streams. Throughout the

past decade, researchers have relied on machine learning techniques as well as deep learning [16] to develop efficient systems for human activity recognition based on hand-crafted features as well as deep features. Extensive research has been conducted on both traditional and deep learning-based methods for identifying human action and activity. In light of the fact that video data analytics necessitates both spatial and temporal information for analysis, we have outlined some of the essential approaches for learning sequential video data using handcrafted features.

Mahmoodi et al.[7] presented a descriptor to identify potentially harmful activity using Histogram of optical flow's attributes. A measure of optical flow's intensity was suggested by Hassner et al. [14] to classify between typical and aberrant behavior using the support SVM classifier. The Oriented Violent Flow (OViF) description, introduced by Gao et al. [17] takes use of both magnitude and direction variations in the video's optical flow. While OViF is more effective in less-crowded settings, ViF excels in them. Cheng et al.[18] proposed a descriptor based on optical flow and log-polar histogram, and they utilized principal component analysis (PCA) to decrease the descriptor's dimensionality such that support vector machines (SVMs) and random forests (FFs) could identify human activity. Mabrouk et al.[19] presented the descriptor DiMOLIF, which uses SVM classifier to differentiate between violent and peaceful images based on the movements around key interest spots. Mabrouk et al. and C. Dhiman et al. present a comprehensive study[19, 20] of the many methods used to identify instances of suspicious human behavior. Pixel characteristics on the multi-level Gaussian distribution are used by Vishwakarma et al. [20] to offer a new descriptor. These features include color space values, Schmid filter responses, color moments, and gradient information. Ullah et. al. [21] propose a framework for activity recognition in surveillance videos captured over industrial systems. Ladjailia et.al. [22] study human activity recognition via optical flow: decomposing activities into basic actions. An optical flow descriptor is proposed for the recognition of human actions by considering only features derived from the motion.

Gesture recognition has been a topic of study for many years. Many solutions are suggested to deal with this issue. Early methods like HOF commonly used handmade features such as the histogram of oriented gradients (HOG) and the histogram of optical flow(HOF)[23–25]. The HOF data is used by Colque and Rensso [26] and Zhang et al. [27] as a motion descriptor. Human gestures are also modeled using the hidden Markov model used by Kumar et al.[28] and Kuehne et al.[29], the conditional random field method by Li et al.[10], and the temporal warping of Kuehne et al. [29]. Temporal features are also important to effectively handle video data. A 3D HOG feature is suggested by Patel et al. [30] for activity recognition. Yang et al. method's [31] of extracting HOG features and establishing a relationship between a depth map to a depth motion map simplifies activity recognition. For one-shot learning of gesture recognition, Ullah et al.[32] offer the mixed features around the sparse key points method.

Convolutional neural networks have recently demonstrated excellent promise in recognition tasks. A CNN-based approach is suggested by Neeraj et al.[33] to categorize films on huge datasets. Dai et al.[34] employ an LSTM network for label prediction after initially using a CNN to extract characteristics. By using multi-task learning, Simonyan et al. [35] increase the volume of training data and use two-stream convolutional networks to extract spatial and temporal concurrent information. Gupta et al. [36] use an autoencoder to solve the issue of multiple performance velocities. The classic CNN, on the other hand, focuses on feature extraction from images. The temporal data is never used for activities that use visual input. A temporal segment network is proposed by Wang et al. [37] where temporal information is acquired from a sequence of optical flow. Another approach to extracting spatiotemporal features is provided by certain researchers. The author uses a hardwired layer to first extract a few manually created characteristics, such as optical flow as well as gradient, and then sends the entire set of features through a 3D CNN for feature extraction. Fan et al. [38]'s innovative idea of using dynamic images to depict films is advantageous for the use of CNN. An updated version of Jia et al. BVLC 's caffe, called C3D, is proposed by Li et al. [39] and processes on real-time video.

Precisely, the work in concern focuses on constructing a model based on hand-crafted features and deep features to create an effective classification tool for cooperative/noncooperative activity. Optical flow is one of the best features which is used extensively because it keeps track of the displacement of pixels in a specific direction. We also exploit the optical flow histogram to craft features based on weighted thresholding to classify video as non-violent and violent. It is well-suited for implementing real-time surveillance systems with the goals of improving public safety, security, and crime prevention because to its dependence on optical flow for effective feature extraction, minimal processing complexity, quick inference, and flexibility to varied situations. The approach has practical usefulness in a variety of real-world surveillance applications due to its prompt identification of violent acts.

3. Method for Proposed Descriptor

We described the proposed method through the system architecture shown in Fig.1. To begin, we must convert the grayscale of each frame of the input video. Then the MOOF of variations between the pairs of subsequent frames is determined. Following this, eight binary histograms are obtained by applying different threshold values to the computed MOOF changes. A feature descriptor is derived from these histograms. At last, the 160 descriptors are used as a vector feature in the SVM classifier. Fig.2. describes the steps to obtain descriptors using optical flow.

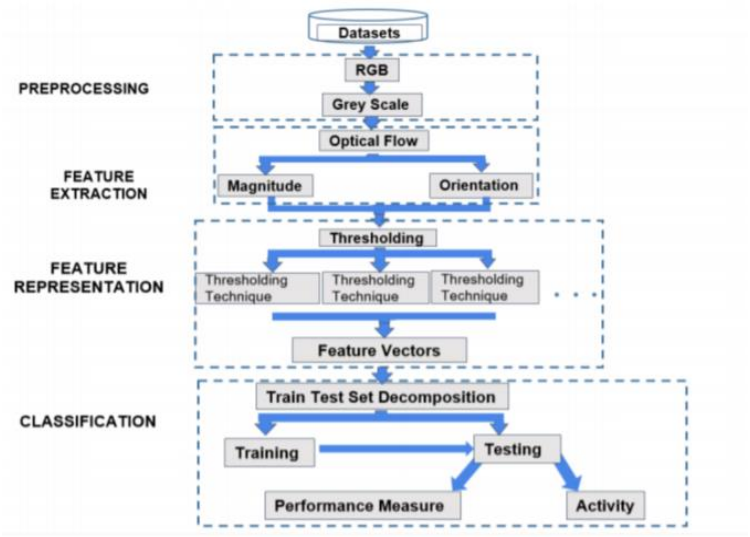

Fig. 1. Architecture of proposed method

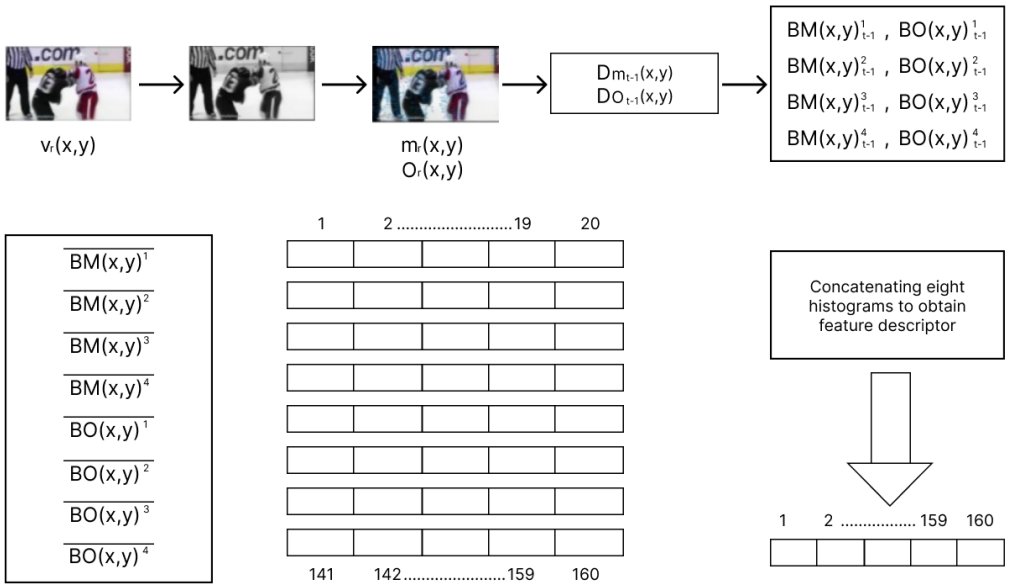

Fig. 2. Steps for producing descriptors

3.1 Extracting frames from the input video:

Firstly, a video is read from the standard database which contains a number of videos. Then various frames at different timestamps are extracted which will be further used for processing and training our model. The sequence of frames from a video stream $V(x, y)$ can be expressed by equation (1).

$$V(x,y) = \{F_1(x,y), F_2(x,y), F_3(x,y), \ldots . F_t(x,y)\} \quad (1)$$

In equation (1), $F_t(x, y)$ represents the frame of the stream with a dimension of H-W-3, where H and W are the frame's height and width, respectively. Integer 3 denotes the colored channel (Red, Green and Blue) of the frame.

3.2 Converting the frame to grayscale:

Frames are expressed as $S(x,y) = \{F_1(x,y), F_2(x,y), F_3(x,y), \ldots . F_t(x,y)\}$. Here, $F_t(x, y)$ indicates the t^{th} frame of the video. Applying equation (2), an RGB video frame can be transformed into a grayscale.

$$F_t(x,y) = 0.2989 * r_t(x,y) + 0.5870 * g_t(x,y) + 0.1140 * b_t(x,y) \quad (2)$$

The coefficients used in equation (2) are based on the photometric luminosity function, which is a mathematical model of how the human eye perceives brightness. The mathematical model is called the photometric luminosity function and defined by German physicist Max Planck. It describes how the human eye responds to different wavelengths of light and how it perceives brightness.

3.3 Optical flow's magnitude and orientation:

Optical flow is used to determine the direction and velocity of an object's motion from one frame to the next. Equations (3) and (4) can be used to calculate the amplitude and direction of the optical flow based on the obtained optical flow vectors.

$$\mathrm{M_t}(x, y) = \sqrt{{U_t}^2(x,y) + {V_t}^2(x,y)} \tag{3}$$

$$\mathrm{O_t}(x, y) = arctan\left(\frac{V_t(x,y)}{U_t(x,y)}\right) \tag{4}$$

Here, $U_t(x,y)$ and $V_t(x,y)$ represents the spatial derivation of optical flow vector, whereas $\mathrm{M}_t(x,y)$ and $\mathrm{O}_t(x,y)$ denote the magnitude and orientation of the optical flow. After then, the equation (5) and (6) is used to evaluate each frame's pixel's magnitude and orientation compared to those of the preceding frame.

$$\mathrm{DM}_t(x,y) = |\mathrm{M}_{t+1}(x,y) - \mathrm{M}_t(x,y)| \tag{5}$$

$$\mathrm{DO}_t(x,y) = |\mathrm{O}_{t+1}(x,y) - O_t(x,y)| \tag{6}$$

3.4 Algorithm I. For feature descriptor

$\boldsymbol{Input\ V(x,y)}$
$\boldsymbol{For}$ *each frame of* $V(x,y)\boldsymbol{do}$
$F_t(x,y) = \boldsymbol{rgb2gray}(V_t(x,y))$
$m_t(x,y) = \boldsymbol{calulate\ optical\ flow\ magnitude}(F_t(x,y))$
$O_t(x,y) = \boldsymbol{calulate\ optical\ flow\ orientation}(F_t(x,y))$
$\boldsymbol{End\ for}$

$\boldsymbol{For}$ *each* $m_t(x,y)$ $\boldsymbol{and}$ $O_t(x,y)$ $\boldsymbol{do}$
$Dm_t = |m_{t+1}(x,y) - m_t(x,y)|$
$DO_t = |O_{t+1}(x,y) - O_t(x,y)|$
$BM(x,y)_t^1 = Binary\ Value\ of\ Dm_t\ based\ on\ it's\ adaptive\ threshold\ value$
$BM(x,y)_t^2 = Binary\ Value\ of\ Dm_t\ based\ on\ it's\ Otsu\ threshold\ value$
$BM(x,y)_t^3 = Binary\ Value\ of\ Dm_t\ based\ on\ it's\ mean\ threshold\ value$
$BM(x,y)_t^4 = Binary\ Value\ of\ Dm_t\ based\ on\ it's\ Optimal\ threshold\ value$
$BO(x,y)_t^1 = Binary\ Value\ of\ Dm_t\ based\ on\ it's\ Adaptive\ threshold\ value$
$BO(x,y)_t^2 = Binary\ Value\ of\ Dm_t\ based\ on\ it's\ Otsu\ threshold\ value$
$BO(x,y)_t^1 = Binary\ Value\ of\ Dm_t\ based\ on\ it's\ mean\ threshold\ value$
$BO(x,y)_t^1 = Binary\ Value\ of\ Dm_t\ based\ on\ it's\ Optimal\ threshold\ value$
$\boldsymbol{End\ for}$
Mean value of eight binary indicator is obtained by equation (9) and (10)
$\boldsymbol{K = 1}$
$\boldsymbol{For\ j = 0\ to\ 1\ step\ 0.05}$(step is considered to be 0.05 to split $[0,1] into\ 20\ bins$)

$X1(K) = Find(j < BM(x,y)^1 < j + 0.05)$
$X2(K) = Find(j < BM(x,y)^2 < j + 0.05)$
$X3(K) = Find(j < BM(x,y)^3 < j + 0.05)$
$X4(K) = Find(j < BM(x,y)^4 < j + 0.05)$
$X5(K) = Find(j < BO(x,y)^1 < j + 0.05)$
$X6(K) = Find(j < BO(x,y)^2 < j + 0.05)$
$X7(K) = Find(j < BO(x,y)^3 < j + 0.05)$
$X8(K) = Find(j < BO(x,y)^4 < j + 0.05)$
$K = K + 1$
$\boldsymbol{End\ For}$
$X = [X1\ X2\ X3\ X4\ X5\ X6\ X7\ X8\ ]$
$Descriptor = X\ is\ divided\ by\ the\ number\ of\ pixel\ in\ F_t(x,y)$

3.5 Applying different thresholding

To determine whether an image pixel is white or black, the thresholding method is used. The pixel is made white i.e., one if and only if its brightness is higher than the threshold value. If not, it will be set to black i.e., zero. One and zero represent white and black respectively. Equation (7) expresses the four binary indicators of the optical flow 's magnitude for each pixel in every frame.

$$BM(x,y)_t^i = \begin{cases} 1, & DM_t(x,y) > \theta M^i(x,y) \quad i = 1,2,3,4 \\ 0, & otherwise \end{cases} \tag{7}$$

Here, we have four binary indicators denoted i.e. $BM(x,y)_t^1$, $BM(x,y)_t^2$, $BM(x,y)_t^3$, $BM(x,y)_t^4$ of each frame based on $\theta M^1(x,y), \theta M^2(x,y), \theta M^3(x,y), \theta M^4(x,y)$ i.e. adaptive threshold, Otsu threshold, mean threshold and optimal threshold of $DM_t(x,y)$, respectively.

Each pixel in each frame can be represented by one of four binary indications that indicate changes in orientation and it is represented by equation (8)

$$BO(x,y)_t^i = \begin{cases} 1, & DO_t(x,y) > \theta O^i(x,y) \quad i = 1,2,3,4 \\ 0, & otherwise \end{cases} \tag{8}$$

Here, $BO(x,y)_t^i$ signifies four binary indicators, with $BO(x,y)_t^1, BO(x,y)_t^2, BO(x,y)_t^3$ *and* $BO(x,y)_t^4$ are being binary indicators of each frame based on $\theta O^1(x,y), \theta O^2(x,y), \theta O^3(x,y), \theta O^4(x,y)$ i.e. adaptive threshold, Otsu threshold , mean threshold and optimal threshold of $DO_t(x,y)$ respectively.

3.6 Evaluating feature descriptor

After obtaining eight binary histograms for every frame, the mean of each histogram is calculated for all frames by equations (9) and (10).

$$\overline{BM(x,y)^i} = \frac{1}{T-1}\sum_{t=1}^{T-1} BM(x,y)_t^i \quad \text{i = 1, 2, 3, 4} \tag{9}$$

$$\overline{BO(x,y)^i} = \frac{1}{T-1}\sum_{t=1}^{T-1} BO(x,y)_t^i \quad \text{i = 1, 2, 3, 4} \tag{10}$$

Here, t is frame's number. Finally, $\overline{BM(x,y)^i}$ and $\overline{BO(x,y)^i}$ are calculated and combined to form the feature descriptor. Each histogram has 20 discrete bins. The number of features is calculated by multiplying the number of bins by the number of histograms. Algorithm I describe the steps to calculate the proposed descriptor.

4. Result and Performance

This section elaborates on the SVM classifier, performance metric, and two benchmark datasets. The suggested technique is then compared to state of art methods, namely ViF[14], OViF[17], DiMOLIF [19], and HOMO[7]. All these models make use of the Optical flow feature. Finally, the impact of a variety of parameters is discussed.

4.1 SVM Classifier

Support Vector Machine (SVM) is a popular and powerful classification algorithm that can be used for both linear and nonlinear classification tasks, where the goal is to predict an outcome based on a set of features or input variables. we have used SVM classifier to classify video segments as violent or non-violent. For nonlinearly separable data, SVM maps the input data to a higher-dimensional feature space using a kernel function, which allows for a linear separation. The radial base function is evaluated for use as the SVM kernel. The SVM kernel has been set to automatic. Performance evaluation is an extremely essential duty in machine learning. We, therefore, rely on Classification accuracy, Receiver Operating Characteristics (ROC), Area Under ROC Curve (AUC), and Standard deviation when faced with a classification problem. The ROC is a graphical representation of the true positive rate versus the false positive rate. The AUC indicates the degree to which a classifier can distinguish different classes. To determine the accuracy of the SVM, each video is divided into five groups where at each checkpoint, one group is chosen at random for evaluation, while the other four serve as the training. This process is repeated five times. The performance of the classifier is evaluated after each step.

4.2 Datasets

The effectiveness of the aforementioned technique has been measured using the Hockey Fight dataset and the Violent Flows dataset, both of which are freely accessible to the public. The Hockey Fight dataset is used to evaluate the suggested approach in non-crowded scenarios. This dataset consists of 1,000 violent and non-violent hockey game video clips that are evenly distributed across the thousand clips. The shortest videos are 1.74 seconds long and the

longest are 2.06 seconds long, with 1.75 seconds being the median duration. The controlled setting, high contrast behaviors, clear-cut annotations, numerous situations, and potential for generalization make the Hockey Fight dataset useful for anomaly detection. The Violent Flows dataset is a dataset of 246 clips containing 123 violent and 123 non-violent scenarios from a crowded area. In football stadiums, cities, and schools, video clips are captured. The mean value of the number of frames is ninety. The video segment is 3.45 seconds long, with the lowest being 1.06 seconds and the largest lasting 6.46 seconds. The documented aberrant occurrences, high contrast behaviors, varied contexts, realistic representation, potential for generalization, and benchmarking capabilities make the Violent Flow dataset useful for finding anomalous behavior. The calculated optical flow on both Hockey Fight and Violent Flows dataset has been depicted in Figure 3 and Figure 4 respectively.

Fig. 3. Optical Flow of Hockey Fight dataset video

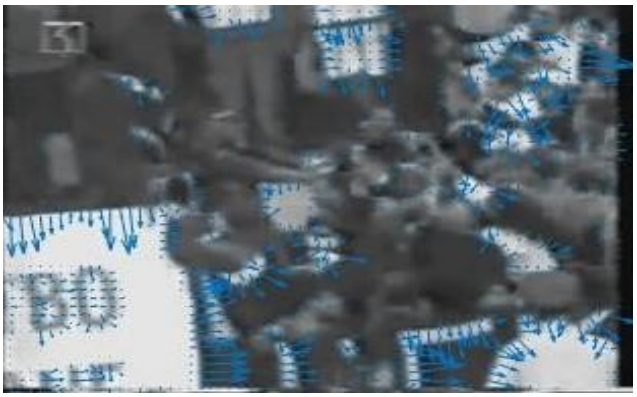

Fig. 4. Optical Flow of Violent Flow dataset video

4.3 Experimental Results

Our model was built in MATLAB (R2019a) and run on an i3 CPU with 8 GB of RAM. On the Violent Flow dataset and the Hockey Fight dataset, the effectiveness of the proposed model is compared to that of the existing method, which makes use of the HOMO [7], DiMOLIF[19], OViF [17], and ViF [14] descriptors based on optical flow. Classification Accuracy, Area Under ROC Curve (AUC), and Standard Deviation are used to evaluate the efficacy of our model. Use support vector machines (SVMs) with a radial base function kernel to estimate a classifier that can separate violent from non-violent video clips.

Table 1. Performance metrics on the Hockey Fight dataset

Method	Accuracy %	Standard deviation	AUC %
ViF	81.6	.22	88.01
OViF	84.2	.3.33	90.32
DiMOLIF	88.6	1.2	93.23
HOMO	89.3	0.91	95.18
Proposed Method	91.5	0.0082	96.43

The suggested technique achieves the best performance of all existing methods on the Hockey Fight dataset. In Table 1 we can observe the outcomes of our comparisons using the Hockey Fight dataset. The table shows that the suggested technique has an AUC of 96.43 and an accuracy of 91.5%. Figure 5 shows the ROC curves of the existing approaches and the proposed approach which shows the supremacy of the proposed approach on the Hockey Fight dataset.

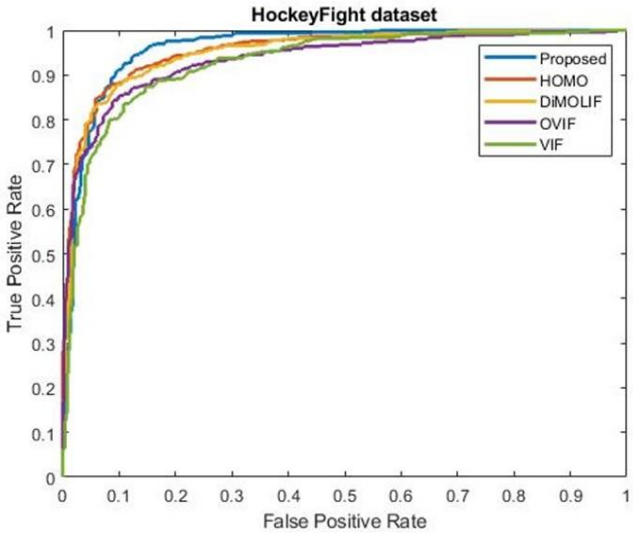

Fig.5. The ROC curves of different methods on the Hockey Fight dataset.

The Violent Flows dataset is used to evaluate the performance of the proposed method in crowded scenarios. In the Violent Flows dataset, our proposed method yields superior outcomes compared to state of art algorithms. Table 2 shows the comparative analysis of accuracy, standard deviation and AUC of the Violent Flows dataset. As shown in Table 2, the AUC is 91.17% and Accuracy is 86.37% of the proposed method, respectively. Figure 6 illustrates the ROC of the existing method and proposed method on the Violent Flows dataset, which shows the efficacy of the proposed method.

Table 2. Accuracy of Violent Flow dataset

Method	Accuracy	Standard Deviation	AUC
ViF	81.2	1.79	88.04
OViF	76.8	3.9	80.47
DiMOLIF	85.83	4.26	89.25
HOMO	76.83	1.76	82.84
Proposed Method	86.37	.0167	91.17

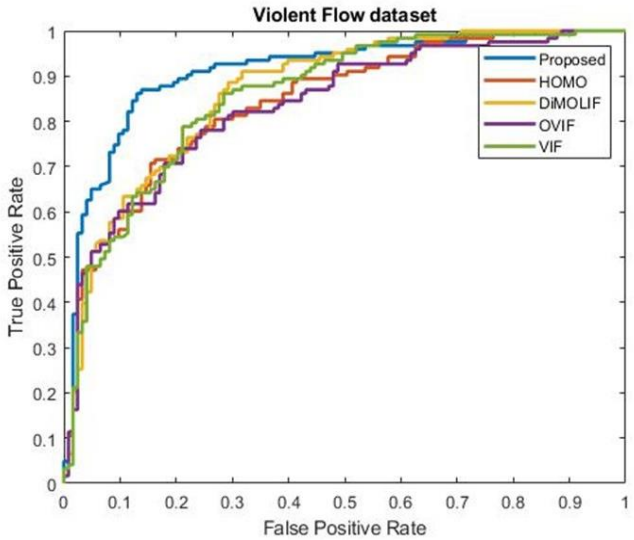

Fig. 6. The ROC of the Violent Flow dataset using the proposed method.

4.4 Number of Descriptor

Table 3 provides a dimensional analysis of the spatial complexity of the suggested and considered descriptors. By stringing together eight histograms, each with 20 bins, we have a length of 160, which is the intended length of our descriptor. Each frame is separated into non-overlapping portions in the currently available descriptors (ViF, OViF,

DiMOLIF, and HOMO). The histogram is then calculated for each block. Descriptor sizes may be calculated by multiplying the number of bins by the number of blocks.

Table 3. Length of descriptor

Descriptor name	Length	Numbers of bins	Number of blocks
ViF	320	20	16
OViF	144	9	16
DiMOLIF	128	8	16
HOMO	120	20	0
Proposed Method	160	20	0

4.5 Execution Time

We choose eight videos from the Violent Flows dataset and the Hockey Fight dataset to examine the computational time of descriptors. The suggested method's computing time is compared with that of the HOMO descriptor, whose source code is publicly accessible, for both datasets. Table 4 demonstrates that compared to HOMO, the suggested technique requires less computational time. It may be inferred that the suggested descriptor has a shorter processing time than the existing descriptors, since [1] HOMO is faster than ViF and OViF.

Table 4. Comparative analysis of Computational time for our and HOMO descriptor

Video Number	1	2	3	4	5	6	7	8
Duration	1	2	1	1	1	6	2	3
Violent clip	√	√			√	√		
Non – Violent clip			√	√			√	√
Number of Frames	41	49	41	41	46	159	59	118
Hockey Fight Dataset	√	√	√	√				
Violent Flow Dataset					√	√	√	√
Homo Execution time	1.32	1.36	1.22	1.32	1.07	3.22	1.26	2.21
Proposed Method execution time	1.19	1.41	1.17	1.26	0.93	3.1	1.18	1.18

4.6 Influence of Various Parameters

A second set of tests was run with varying bin sizes to see how the bin size affected the histograms. Tables 5 and 6 show the outcomes of the Hockey Fight and the Violent Flows analyses, respectively. These tables show that when using the suggested strategy, the optimal number of divisions is twenty.

Table 5. Comparison results of changing the bin size on the magnitude

Bin Size	Accuracy
20	94.5
30	93
40	93
50	93

Table 6. Comparison results of changing the bin size on orientation

Bin Size	Accuracy
20	85.71
30	83.67
40	82.75
50	81.67

5. Conclusion and Future Work

In this research, we suggest a strong descriptor that may be used in both crowded and uncrowded settings to identify instances of aggressive behaviour. It relies on the concept of threshold values, which are applied to changes in the amount and direction of the optical flow. Gray scaling the input frames comes first, followed by establishing the MOOF between two output frames. Then, we determine the median values of eight binary indicators that characterize the size and direction of the shifts between two consequence frames. After that, eight histograms of average binary indicators will be created. Binary feature vectors are generated by concatenating these histograms. The classification is then executed using a support vector machine classifier. The methodology leverages optical flow, as the primary feature which exhibits real-time capability. The reliance on optical flow for feature extraction and converting these features as binary values by applying various thresholds suggests a low computational complexity compared to complex optical flow features. This ensures that the methodology can be efficiently implemented on various devices with limited computational resources. Experiments conducted on the Hockey Fight and Violent Flow datasets showed that the suggested descriptor is better than the gold standard. Although the present study only employed a few datasets as its

starting standards, the authors acknowledge the need to broaden their review in future studies. We suggest further work to improve the system's accuracy in crowded circumstances by considering other features of the MOOF.

References

[1] S. Tanberk, Z. H. Kilimci, D. B. Tukel, M. Uysal, and S. Akyokus, "A Hybrid Deep Model Using Deep Learning and Dense Optical Flow Approaches for Human Activity Recognition," *IEEE Access*, vol. 8, pp. 19799–19809, 2020, doi: 10.1109/ACCESS.2020.2968529.

[2] X. Zhang, S. Yang, J. Zhang, and W. Zhang, "Video anomaly detection and localization using motion-field shape description and homogeneity testing," *Pattern Recognit.*, vol. 105, p. 107394, 2020, doi: 10.1016/j.patcog.2020.107394.

[3] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image Vis. Comput.*, vol. 60, pp. 4–21, Apr. 2017, doi: 10.1016/J.IMAVIS.2017.01.010.

[4] H. S. Park and J. Shi, "Force from Motion: Decoding Control Force of Activity in a First-Person Video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 622–635, 2020, doi: 10.1109/TPAMI.2018.2883327.

[5] S. H. Kumar and P. Sivaprakash, "New approach for action recognition using motion based features," *2013 IEEE Conf. Inf. Commun. Technol. ICT 2013*, no. Ict, pp. 1247–1252, 2013, doi: 10.1109/CICT.2013.6558292.

[6] L. Sun, Y. Chen, W. Luo, H. Wu, and C. Zhang, "Discriminative Clip Mining for Video Anomaly Detection," *Proc. - Int. Conf. Image Process. ICIP*, vol. 2020-October, pp. 2121–2125, Oct. 2020, doi: 10.1109/ICIP40778.2020.9191072.

[7] J. Mahmoodi and A. Salajeghe, "A classification method based on optical flow for violence detection," *Expert Syst. Appl.*, vol. 127, pp. 121–127, 2019, doi: 10.1016/j.eswa.2019.02.032.

[8] J. Sochor, J. Spanhel, and A. Herout, "BoxCars: Improving Fine-Grained Recognition of Vehicles Using 3-D Bounding Boxes in Traffic Surveillance," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 97–108, Jan. 2019, doi: 10.1109/TITS.2018.2799228.

[9] J. Liu, Z. Wang, and H. Liu, "HDS-SP: A novel descriptor for skeleton-based human action recognition," *Neurocomputing*, vol. 385, pp. 22–32, 2020, doi: 10.1016/j.neucom.2019.11.048.

[10] W. Li and D. Cosker, "Video interpolation using optical flow and Laplacian smoothness," *Neurocomputing*, vol. 220, pp. 236–243, 2017, doi: 10.1016/j.neucom.2016.04.064.

[11] T. Beeler *et al.*, "High-quality passive facial performance capture using anchor frames," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 1–10, Jul. 2011, doi: 10.1145/2010324.1964970.

[12] D. Bradley, W. Heidrich, T. Popa, and A. Sheffer, "High resolution passive facial performance capture," *ACM Trans. Graph.*, Jul. 2010, doi: 10.1145/1778765.1778778.

[13] S. Kamal and A. Jalal, "A Hybrid Feature Extraction Approach for Human Detection, Tracking and Activity Recognition Using Depth Sensors," *Arab. J. Sci. Eng.*, vol. 41, no. 3, pp. 1043–1051, Mar. 2016, doi: 10.1007/S13369-015-1955-8/METRICS.

[14] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 1–6, 2012, doi: 10.1109/CVPRW.2012.6239348.

[15] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks", Accessed: May 11, 2022. [Online]. Available: https://github.com/liuzhuang13/DenseNet.

[16] M. Kumar and M. Biswas, "Abnormal human activity detection by convolutional recurrent neural network using fuzzy logic," *Multimed. Tools Appl.*, no. 0123456789, 2023, doi: 10.1007/s11042-023-15904-x.

[17] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using Oriented VIolent Flows," *Image Vis. Comput.*, vol. 48–49, pp. 37–41, Apr. 2016, doi: 10.1016/J.IMAVIS.2016.01.006.

[18] C. Dai, X. Liu, J. Lai, P. Li, and H. C. Chao, "Human behavior deep recognition architecture for smart city applications in the 5G environment," *IEEE Netw.*, vol. 33, no. 5, pp. 206–211, Sep. 2019, doi: 10.1109/MNET.2019.1800310.

[19] A. Ben Mabrouk and E. Zagrouba, "Spatio-temporal feature using optical flow based distribution for violence detection," *Pattern Recognit. Lett.*, vol. 92, pp. 62–67, Jun. 2017, doi: 10.1016/J.PATREC.2017.04.015.

[20] C. Dhiman and D. K. Vishwakarma, "A review of state-of-the-art techniques for abnormal human activity recognition," *Eng. Appl. Artif. Intell.*, vol. 77, no. September 2018, pp. 21–45, 2019, doi: 10.1016/j.engappai.2018.08.014.

[21] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," *Br. Mach. Vis. Conf. BMVC 2009 - Proc.*, pp. 124.1-124.11, Sep. 2009, doi: 10.5244/C.23.124.

[22] A. Ladjailia, I. Bouchrika, H. F. Merouani, N. Harrati, and Z. Mahfouf, "Human activity recognition via optical flow: decomposing activities into basic actions," *Neural Comput. Appl.*, vol. 32, no. 21, pp. 16387–16400, Nov. 2020, doi: 10.1007/S00521-018-3951-X/METRICS.

[23] N. Yudistira and T. Kurita, "Multiresolution Local Autocorrelation of Optical Flows over Time for Action Recognition," *Proc. - 2015 IEEE Int. Conf. Syst. Man, Cybern. SMC 2015*, pp. 1930–1935, 2016, doi: 10.1109/SMC.2015.337.

[24] Y. Shi, W. Zeng, T. Huang, and Y. Wang, "Learning Deep Trajectory Descriptor for action recognition in videos using deep neural networks," *Proc. - IEEE Int. Conf. Multimed. Expo*, vol. 2015-August, Aug. 2015, doi: 10.1109/ICME.2015.7177461.

[25] H. Wang and C. Schmid, "Action recognition with improved trajectories," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 3551–3558, 2013, doi: 10.1109/ICCV.2013.441.

[26] R. V. H. M. Colque, C. Caetano, M. T. L. De Andrade, and W. R. Schwartz, "Histograms of Optical Flow Orientation and Magnitude and Entropy to Detect Anomalous Events in Videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 673–682, Mar. 2017, doi: 10.1109/TCSVT.2016.2637778.

[27] Y. Zhang, H. Lu, L. Zhang, and X. Ruan, "Combining motion and appearance cues for anomaly detection," *Pattern Recognit.*, vol. 51, pp. 443–452, Mar. 2016, doi: 10.1016/J.PATCOG.2015.09.005.

[28] D. K. Singh, S. Paroothi, M. K. Rusia, and M. A. Ansari, "Human Crowd Detection for City Wide Surveillance," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 350–359, 2020, doi: 10.1016/j.procs.2020.04.036.

[29] H. Kuehne, A. Richard, and J. Gall, "A Hybrid RNN-HMM Approach for Weakly Supervised Temporal Action Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 765–779, Apr. 2020, doi: 10.1109/TPAMI.2018.2884469.

[30] C. I. Patel, S. Garg, T. Zaveri, A. Banerjee, and R. Patel, "Human action recognition using fusion of features for unconstrained video sequences," *Comput. Electr. Eng.*, vol. 70, pp. 284–301, Aug. 2018, doi: 10.1016/J.COMPELECENG.2016.06.004.
[31] L. Yang *et al.*, "A Refined Non-Driving Activity Classification Using a Two-Stream Convolutional Neural Network," *IEEE Sens. J.*, vol. 21, no. 14, pp. 15574–15583, 2021, doi: 10.1109/JSEN.2020.3005810.
[32] A. Ullah, K. Muhammad, K. Haydarov, I. U. Haq, M. Lee, and S. W. Baik, "One-Shot Learning for Surveillance Anomaly Recognition using Siamese 3D CNN," *Proc. Int. Jt. Conf. Neural Networks*, Jul. 2020, doi: 10.1109/IJCNN48605.2020.9207595.
[33] Neeraj, V. Singhal, J. Mathew, and R. K. Behera, "Detection of alcoholism using EEG signals and a CNN-LSTM-ATTN network," *Comput. Biol. Med.*, vol. 138, no. June, p. 104940, 2021, doi: 10.1016/j.compbiomed.2021.104940.
[34] C. Dai, X. Liu, and J. Lai, "Human action recognition using two-stream attention based LSTM networks," *Appl. Soft Comput. J.*, vol. 86, p. 105820, Jan. 2020, doi: 10.1016/J.ASOC.2019.105820.
[35] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," *Adv. Neural Inf. Process. Syst.*, vol. 1, no. January, pp. 568–576, Jun. 2014, doi: 10.48550/arxiv.1406.2199.
[36] P. Kumar, G. P. Gupta, and R. Tripathi, "An ensemble learning and fog-cloud architecture-driven cyber-attack detection framework for IoMT networks," *Comput. Commun.*, vol. 166, no. April 2020, pp. 110–124, 2021, doi: 10.1016/j.comcom.2020.12.003.
[37] Y. Hu, R. Song, Y. Li, P. Rao, and Y. Wang, "Highly accurate optical flow estimation on superpixel tree," *Image Vis. Comput.*, vol. 52, pp. 167–177, 2016, doi: 10.1016/j.imavis.2016.06.004.
[38] Y. Fan, M. D. Levine, G. Wen, and S. Qiu, "A deep neural network for real-time detection of falling humans in naturally occurring scenes," *Neurocomputing*, vol. 260, pp. 43–58, 2017, doi: 10.1016/j.neucom.2017.02.082.
[39] Y. Li *et al.*, "Large-scale gesture recognition with a fusion of RGB-D data based on optical flow and the C3D model," *Pattern Recognit. Lett.*, vol. 119, pp. 187–194, 2019, doi: 10.1016/j.patrec.2017.12.003.

Authors' Profiles

Manoj Kumar received his B.Tech degree in Computer Science and Engineering from BCE Bhagalpur, Bihar in 2007 and M. Tech degree in Information Technology from GGSIP University, New Delhi in 2012. He is pursuing a Doctorate of Philosophy from the National Institute of Technology, Kurukshetra, India. He is an assistant professor at JSS Academy of Technical Education Noida, U.P. His research interest includes computer vision, image processing, machine learning, deep learning, and Artificial Intelligence.

Anoop K. Patel, PhD (2020, NIT Kurukshetra), M.Tech. (2011, MNNIT, Allahabad, India), B.Tech. (2009, Computer Science & Engineering). He is an Assistant Professor in the Department of Computer Engineering at NIT Kurukshetra, India. He is actively involved in research and has more than 10 years' experience in teaching and research. His current research areas include image processing, medical imaging, machine learning, and image/video-based cardiovascular disease characterization.

Mantosh Biswas currently is working as Assistant Professor in the Department of Computer Engineering at National Institute of Technology, Kurukshetra, India. He earned his Ph.D. in Computer Science & Engineering from Indian Institute of Technology (Indian School of Mines), Dhanbad, India. He has published several research papers in International and National Journals including various International and National conferences of high repute. His research areas are Signal & Image Processing, X-let, and Soft Computing.

Dr. Sandeep Singh Sengar is a Lecturer in Computer Science at Cardiff Metropolitan University, United Kingdom. Before joining this position, he worked as a Postdoctoral Research Fellow at the Machine Learning Section of Computer Science Department, University of Copenhagen, Denmark. He holds a Ph.D. degree in Computer Science and Engineering from Indian Institute of Technology (ISM), Dhanbad, India and an M. Tech. degree in Information Security from Motilal Nehru National Institute of Technology, Allahabad, India. Dr. Sengar's current research interests include Medical Image Segmentation, Motion Segmentation, Visual Object Tracking, Object Recognition, and Video Compression. His broader research interests include Machine/Deep Learning, Computer Vision, Image/Video Processing and its applications. He has published several research articles in reputed international journals and conferences in the field of Computer Vision and Image Processing. He is a Reviewer of several reputed International Transactions, Journals, and conferences including IEEE Transactions on Systems, Man and Cybernetics: Systems, Pattern Recognition, Neural Computing and Applications, Neurocomputing.

How to cite this paper: Manoj Kumar, Anoop Kumar Patel, Mantosh Biswas, Sandeep Singh Sengar, "Human Abnormal Activity Recognition from Video Using Motion Tracking", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.16, No.3, pp. 52-63, 2024. DOI:10.5815/ijigsp.2024.03.05