A Locally Weighted Linear Regression-Based Approach for Arbitrary Moving Shaky and Non-Shaky Video Classification

[1]Arnab Halder, [2]Palaiahnakote Shivakumara, [1]Umapada Pal, [3]Michael Blumenstein and [4]Palash Ghosal

[1]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India. Email: arnabhalder1997@gmail.com, umapada@isical.ac.in

[2]Faculty of Computer Science and Information Technology, University of Malaya, Malaysia. Email: shiva@um.edu.my

[3]University of Technology Sydney (UTS), Sydney, Australia. Email:

michael.blumenstein@uts.edu.au

[4]Dept of Information Technology, Sikkim Manipal Institute of Technology, Sikkim, India.

Email: ghosalpalash@gmail.com

Abstract

Classification and identification of objects are complex and challenging in pattern recognition and artificial intelligence if a shaky and non-shaky camera captures the videos at different distances during the day and nighttime. This work presents a model for classifying a given video as static, uniform, or arbitrarily moving videos such that the complexity of the problem can be reduced. To avoid the threat of different distances between the objects and the camera, the proposed work introduces new steps for estimating the depth of the objects in the video frames. We explore locally weighted linear regression for feature extraction from depth information based on the notion that the regression line fits almost all the points for uniformity and does not fit for arbitrary moving. The extracted features are fed to a random forest classifier to classify static, uniform, or arbitrary moving video. The results on a large dataset, which includes videos captured day and night, show that the proposed method successfully classifies static, uniform and arbitrary videos with 0.86, 1.00 and 0.67 F-measures, respectively. Overall, our method obtains 87% accuracy for classification of static, uniform and arbitrary video, which is superior to the state-of-the-art methods.

Keywords: Moving objects detection, Vehicle movements detection, Shaky camera detection. Arbitrarily moving objects detection.

1. Introduction

Automation is common for all fields to make the system cost-effective and accurate and to prevent human error during the night, especially in protected and sensitive areas. In general, in the case of exhibitions, marathons, processions, big events etc, it is necessary to trace the objects, humans and vehicles to study the behavior to detect suspicious activities. To protect and monitor such areas from robbery, stealing, and tampering, there is a need to develop a robust surveillance system in the field of pattern recognition and artificial intelligence. However, detecting intruders, including humans and

vehicles, is not easy at night because of poor quality and lighting effects. In addition, arbitrary movements of objects, such as a tree leaf in the same scene due to wind and a shaky camera, make detecting moving objects (actual) more complex and challenging. It can be seen in Fig. 1, where sample frames captured by a shaky and non-shaky camera are shown in Fig. 1(a) and Fig. 1(b), respectively. It is observed from Fig. 1(a) and Fig. 1(b) that the frames are suffering from poor quality, and the objects are not visible properly, including human movements.

There are several methods proposed in the literature for static and moving object detection and video classification [1-4]. For example, the method [1] proposes a kinematic theory of rapid human movements for neurodegenerative disease video classification. The model [2] uses video information for differentiating caption and scene text based on changes in the pixel values. The approach [3] explores temporal difference and the Otsu thresholding technique for detecting moving objects in video. The technique [4] also uses frame difference and training algorithms for moving object detection in video. It is noted from the above discussion that although video information has been used for moving object detection, the focus of the methods is limited to day video with high quality. In addition, the methods are suitable for detecting objects that move in a particular direction and speed, but not the video containing arbitrary movements and directions. Therefore, there is a dearth of a new method for the classification of static and arbitrary movements of objects in real-time environments.

(a) Sample frames of different videos containing static, uniform moving and arbitrary moving objects

(b) Center of the objects is detected to trace the object movements to classify it as static, uniform, or arbitrary moving video

Fig. 1. Sample video for classification of static, uniform, and arbitrary moving.

Therefore, this work aims to develop a model for the classification of static, uniform, and arbitrary video based on a locally weighted linear regression approach. This model works based on the fact that

the features, namely, pixel magnitude, orientation and speed, change according to object movements in the videos. For instance, in the case of static video, one can expect uniform pixel magnitude, and there is no change in the orientation and speed. However, for the objects in the uniform video, we can expect constant speed and uniform direction. For the objects in arbitrary video, one cannot predict changes in pixel magnitude, orientation and speed. To extract the above observations, we propose a locally weighted linear regression approach. Thus, the key contributions of the proposed model are as follows. (i) Proposing a simple and effective model for addressing the complex problem of classification of static, uniform and arbitrary video of day and night. (ii) Use of depth information of the video normalizing the distance between the camera and objects in the videos such that variations in distance do not affect feature extraction. (iii) Exploring locally weighted linear regression approach for studying the behavior of pixels in terms of their magnitude, orientation and speed for successful classification of static, uniform and arbitrary video.

The structure of the paper is as follows. The review of state-of-the-art methods for the classification of static and moving objects is presented in Section 2. In Section 3, the proposed plans for foreground and background separation, FFT for extracting frequency feature vectors, and the approach for classification of static video and arbitrary moving video are described. Section 4 provides a discussion of several experiments to validate the proposed and existing methods. The conclusion and summary are presented in Section 5.

2. Related Work

There are many methods developed in the literature for moving object detection in video [3-16]. For example, Boufares et al. [3] propose a plan for moving object detection using temporal difference and OTSU thresholding techniques. The approach uses input frame difference at the pixel level for detecting moving objects in the videos. Wang et al. [4] developed a model for moving object detection using frame difference and an algorithm for teaching video. The method uses the OTSU thresholding approach and median filter for moving object detection. Goyal et al. [15] aim to develop a method for detecting moving objects in complex scenes. The approach combines the Gaussian Mixture model with foreground matching for moving object detection. Rai et al. [16] explored thermal image processing to address some of the problems of surveillance applications. The method performs background modeling and background subtraction. However, none of the methods use shaky and non-shaky video for classification. In addition, the focus of the methods is to classify the objects but not the video in contrast to the proposed work, which focuses on video classification based on objects' behavior. Furthermore, the methods may not be effective for arbitrarily moving objects in the video because the features are not robust enough to handle complex situations.

Benaim et al. [17] proposed a model to detect the speed of the vehicles, such as faster, slower, and normal, based on a deep learning approach. Hosono et al. [18] focused on finding object positions in

the video to align the vehicles before classification or detection. Nandhini et al. [19] developed a CNN-based model for moving object detection in video. The approach explores the combination of CNN and the Gaussian Mixture Model for moving object detection. Aliouat et al. [20] aim to develop a method for coding for surveillance systems. The approach uses edge detection as well as frame difference and the sum of absolute differences for coding. In the end, the technique classifies moving and static blocks. Wang et al. [21] used a low-rank sparse representation network for satellite video scene classification. The method explores both spatial and temporal features. Zheng [22] proposed data mining-based techniques for sports video classification. The method uses an SVM classifier for the classification of moving video. Recently, Asadzadehkaljahi et al. [23, 24] developed models for arbitrary moving object detection in shaky and non-shaky video. The methods follow conventional approaches for object detection. The primary objective of the methods [23, 24] is to use video information for detecting objects. However, the proposed work is focused on detecting object behavior for classification moving, static and linearity video.

It is observed from the above review that a few methods addressed the challenge of shaky and non-shaky video and arbitrary moving video classification. Since the existing methods use specific properties of moving object detection, the methods may not be suitable for arbitrary moving object detection in the video. Furthermore, the scope of the existing methods is limited to day video but not night video, where one can expect enormous degradations. Therefore, classifying arbitrarily moving video from normal and uniform moving video is still considered an open challenge. Thus, this work aims to propose a new method for classifying arbitrary moving, uniform moving, and static video.

3. Proposed Method

Since input videos include day-night and are affected by shaky and non-shake cameras, one cannot expect constant quality, either poor or high. For handling degradations, there are enhancement techniques. However, proposing an enhancement method for handling unpredictable quality and degradations is a hard task. In addition, performing enhancement steps on all the temporal frames in the video is not feasible. It is true that temporal information in the video can overcome the above challenges for detecting moving objects in the video. Therefore, the proposed work does not prefer the enhancement step; instead, it explores temporal information for classification in this work. The scope of the classification is limited to three classes, namely static, uniform, and arbitrary video. The motivation to consider three classes is that if we consider any video with moving objects, objects either stay static, move in a particular direction, or move in an arbitrary direction with arbitrary speed. Therefore, any video with moving objects can be classified into three classes.

For a given video as input, to classify it as arbitrary moving, Uniform moving or static video, the method should study the content, including objects in the video. Therefore, the proposed work uses an existing model for object detection in video. Sample results for object detection can be seen in Fig. 1(b), where

it is noted that the method detects all the objects irrespective of video type and objects. For each detected object, we find a centroid and extract features based on pixel magnitude, speed, and direction to differentiate the different types of objects in videos. In other words, the proposed work focuses on something other than object detection. Instead, it studies the behavior in terms of pixel magnitude, speed, and direction. The steps detect even humans as objects. The extracted features are supplied to a random forest classifier for the final classification of arbitrary, uniform, and static video.

The number of training and testing samples is chosen according to a 10-fold cross-validation approach. The block diagram of the proposed work can be seen in Fig. 2, where one can see the steps and flow for classification.

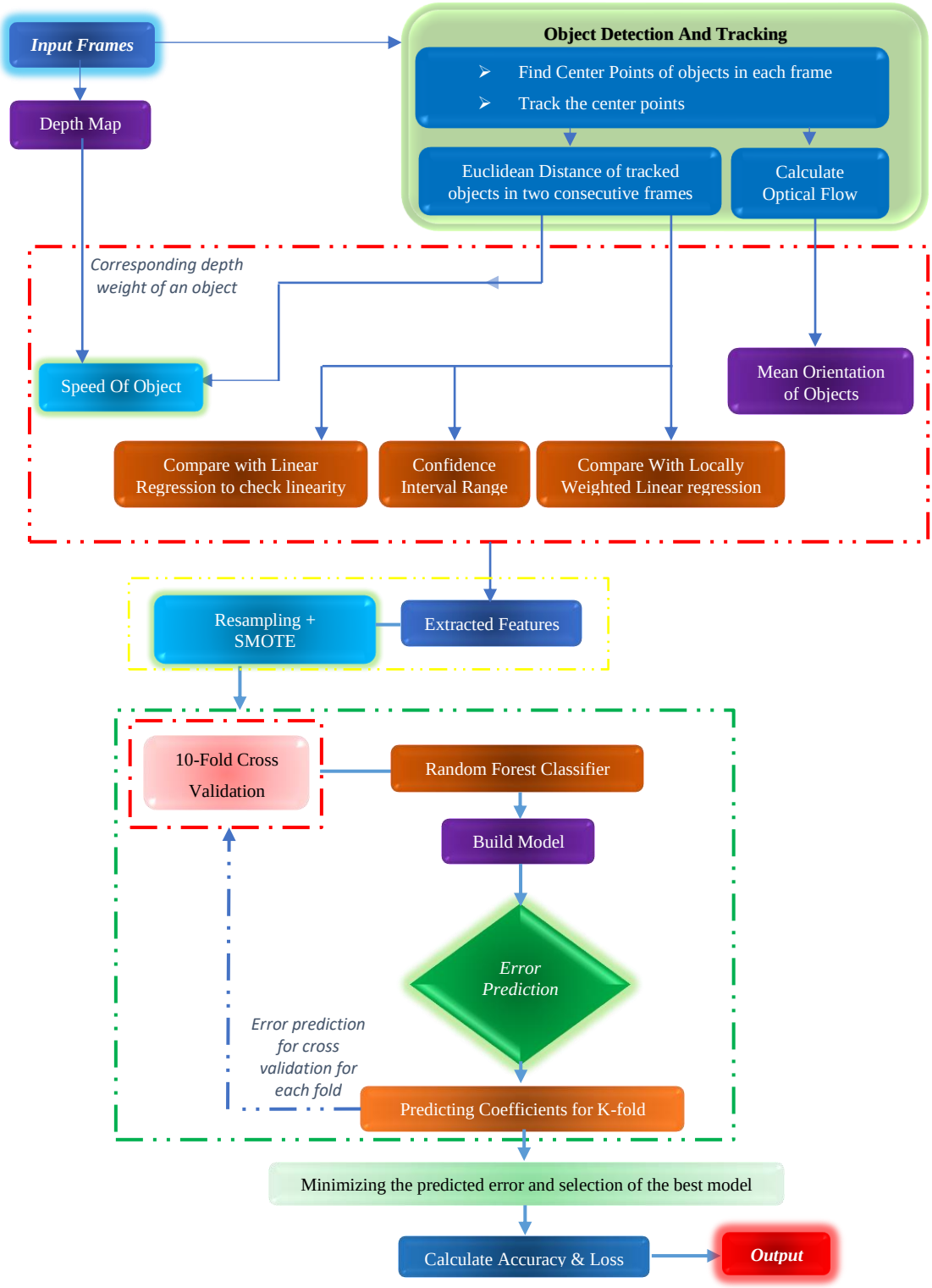


Fig. 2. Block diagram of the proposed method

3.1. Preprocessing

This work considers video captured by shaky and non-shaky cameras at different height distances. In addition, since objects are moving in the video, the distance between the camera and the objects varies greatly. When the distance changes abruptly, it isn't easy to estimate the speed and direction of the

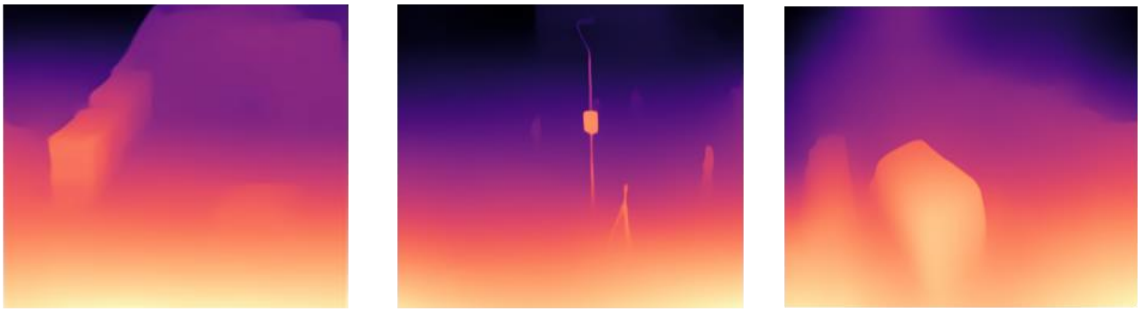
objects. Therefore, to alleviate this problem, we propose to estimate the depth information of the objects in the video. In this study, we use the inverted depth value for getting the accurate depth for an object. Thus, the inverse depth value is considered as a weight for obtaining the accurate speed for an object irrespective of the distance from a camera. The effect of depth estimation is illustrated in Fig. 3 for sample video frames shown in Fig. 3(a), where depth information in (b) normalizes the distance between the object and the camera.

Midas Depth Map: It is a machine learning model that estimates depth from an arbitrary input image, where it proposes to perform prediction in disparity space (inverse depth up to scale and shift) together with a family of scale and shift-invariant dense losses to handle the ambiguities. Let *M* denote the number of pixels in an image with valid ground truth and let be the parameters of the prediction model. Let be a disparity prediction and let be the corresponding ground truth disparity. Individual pixels are indexed by subscripts, which define the scale and shift invariant loss for a single sample as defined in Equation (1).

$$\mathcal{L}_{s\,si}(\hat{d}, \hat{d}^*) = \frac{1}{2M}\sum_{i=1}^{M} p(\hat{d}_i - \hat{d}_i^*), \tag{1}$$

Where $\hat{d}$ and $\hat{d}^*$ are scaled and shifted versions of the predictions and ground truth, and p defines the specific type of loss function.

(a) Objects are detected and labeled for illustrating feature extraction

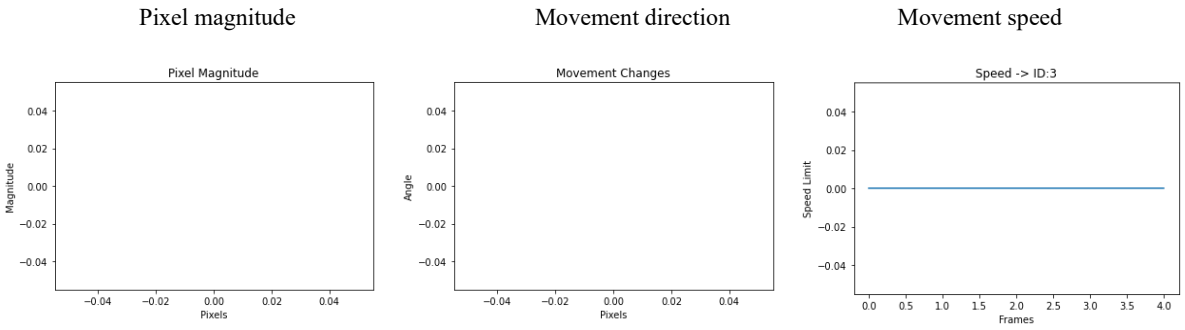


(b) Depth is estimated for sample frames of different videos.

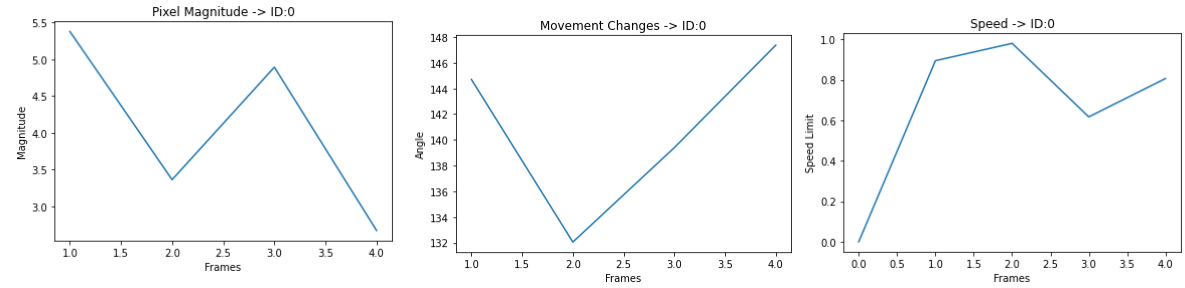
Fig. 3. Feature extraction for classification of static, uniform and arbitrary moving video.

To show the strength of depth information, the proposed method calculates the mean of pixel magnitude, orientation, and speed for the objects in the respective video frames shown in Fig. 3(a). The

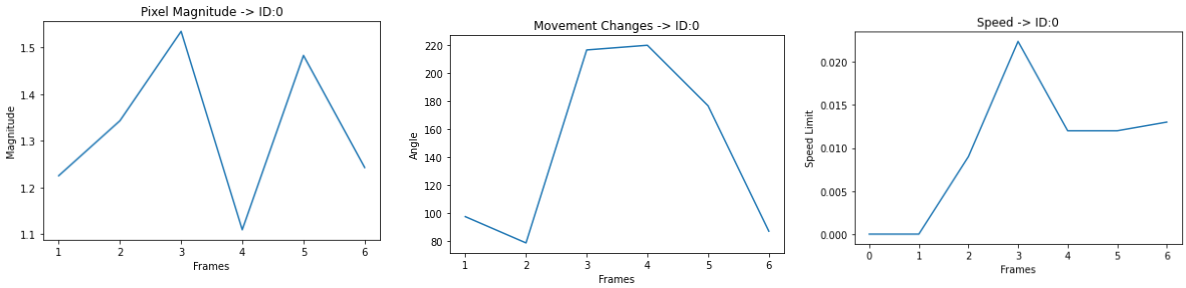
effectiveness of pixel magnitude, direction, and speed can be seen in Fig. 4(a)-(c) for static, uniform, and arbitrary moving video, respectively. It is observed from Fig. 4(a)-(c) that as time changes (frames), there is no change in pixel magnitude, orientation, and speed for static video, as shown in Fig. 4(a), gradual changes for uniform moving video as shown in Fig. 4(b) and arbitrary changes in the case of arbitrary moving video as shown in Fig. 4(c). Overall, Fig. 4 shows that the pixel magnitude, orientation, and speed information are sufficient to differentiate static, uniform moving, and arbitrary moving video. With this illustration, one can argue that depth information is helpful for classifying moving objects in the video despite distance changes randomly between the camera and objects.



(a) Cues for the static objects in the video (object-3-Car)



(b) Cues for the uniform moving objects in the video (object-0-Person)



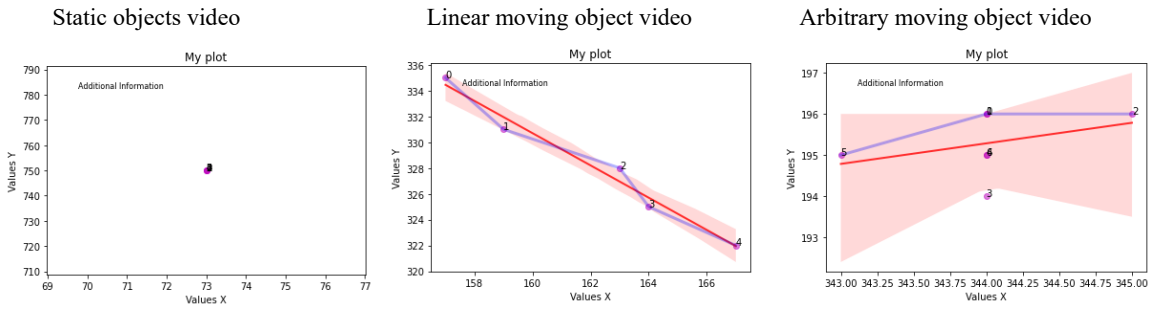
(c) Cues for the arbitrary moving objects in the video (object-0-Car)
Fig. 4. Illustrating cues for discriminating static, uniform and arbitrary moving video.

3.2. Locally Weighted Linear Regression for Classification

In the previous section, it is noted that the pixel magnitude, orientation, and speed are the critical observations for differentiating static, uniform, and arbitrary moving video. These observations are extracted through locally weighted linear regression for classification, which calculates a mean of local weights and a path of the closest point from the local mean. The notion of locally weighted linear regression is that it fits mostly all the points in the case of static and uniform movements, and for

arbitrary movements, the locally weighted linear regression can fit only one of the higher weighted points. This shows that there exists at least one point which can never be fit on the regression line. This indicates arbitrary movements. Based on this notion, the proposed method extracts features based on confidence intervals, and it calculates slope, intercept, and mean error margin.

The same is illustrated in Fig. 5(a)-(c), where it is noted that for the static video, most of the parameters received none, while for uniform and arbitrary moving video, the parameter received uniform and random values, respectively. In addition, for static video, there is no regression line; for uniform moving video, the regression line fits all the points; and for arbitrary moving, the regression line does not fit all the points. The features extracted using regression line behavior are fed to a random forest classifier for the classification of static, uniform, and arbitrary moving video. The reason for choosing the random classifier is that it is evident from the illustration shown in Fig. 5 that the extracted features are capable of distinguishing different videos accurately; a simple classifier is sufficient for successful classification rather than deep learning models.



(a) (b) (c)

For (a), Slope: None, Intercept: None, R-Squared: None, Mean Error margin: ±2.27e-13, For (b), Slope: -1.25, Intercept: 530.7, R-Squared: 0.97, Mean Error Margin: ±6.29 and for (c), Slope: 0.52, Intercept: 226.94, R-Squared: 0.10, Mean Error margin: ±0.52.

Fig. 5. Feature extraction for classification of static, uniform and arbitrary moving video.

The formal steps to derive locally weighted linear regression are as follows.

Locally weighted linear regression is a non-parametric algorithm; that is, the model does not learn a fixed set of parameters as is done in ordinary linear regression. Rather, parameters are computed individually for each query point x. Rather than parameters θ, a higher "preference" is given to the points in the training set lying in the vicinity of x than the points lying far away from x. The modified cost function is defined in Equation (2).

$$j(\Theta) = \sum_{i=1}^{m} w^{(i)}(\theta^T x^{(i)} - y^{(i)})^2 \tag{2}$$

where, $w^{(i)}$ is a non-negative. "weight" is associated with training point $x^{(i)}$. For $x^{(i)}s$ lying far away from x, the value of $w^{(i)}$ is small. A typical choice of $w^{(i)}$ is defined as Equation (2).

$$w^{(i)} = \exp\left(\frac{-\left(x^{(i)} - x\right)^2}{2T^2}\right) \tag{3}$$

where T is called the bandwidth parameter and it controls the rate at which $w^{(i)}$ falls with distance from x clearly, if $|x^{(i)} - x|$ is small $w^{(i)}$ is close to 1, and if $|x^{(i)} - x|$ is large $w^{(i)}$ is close to 0. Thus, the training set points lying closer to the query point x contribute more to the cost $j(\theta)$ than the points lying far away from x.

Confidence Interval: A Confidence Interval is the $mean\ of\ estimate \pm the\ variance$ in that estimate. Confidence level = $1-\propto$

$$\text{Confidence Interval:} \quad CI = \bar{x} \pm z * \frac{\sigma}{\sqrt{n}} \tag{4}$$

Where, $\bar{x} = population\ mean,$ $z = Critical\ value\ of\ distribution,$ $\sigma = Std.Dev.of\ Population,$ $\sqrt{n} = The\ square\ root\ of\ Population\ size$

The speed of the objects can be calculated as follows,

$$s = \sqrt{{p_2}^2 + {p_1}^2} * dpt \tag{5}$$

Where s denotes speed of an object, p_2 is the position of object at time t_2 as well p_1 is the position of object at time t_1 , and the dpt is the following numerical depth information.

The angle of the object movement from time stamp t_1 to t_2 we can measure by using Farneback Optical Flow,

Consider an object with intensity $I(x, y, t), after\ time\ dt$, it moves to $dx\ and\ dy$, now, the new intensity would be, $I(x + dx, y + dy, t + dt)$. We assume the pixel intensities are constant between two frames, i.e.,

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \tag{6}$$

Where Taylor approximation comes to take place on the RHS side, resulting in,

$$\frac{dI}{dx}\delta x + \frac{dI}{dy}\delta y + \frac{dI}{dt}\delta t = 0 \tag{7}$$

On dividing by δt, we obtain the Optical Flow, where we can easily specify the magnitude and the angle of the movement as follows:

$$\frac{dI}{dx}u + \frac{dI}{dy}v + \frac{dI}{dt} = 0 \tag{8}$$

Where, $u = \frac{dx}{dt}\ and\ v = \frac{dy}{dt}$,

Also, $\frac{dl}{dx}$ is the image gradient along the horizontal axis, $\frac{dl}{dy}$ is the image gradient along the vertical axis and $\frac{dl}{dt}$ is along the time.

So, the computed $\frac{dx}{dt}$ and $\frac{dy}{dt}$ are the arrays of respective magnitude and the direction of optical flow from the flow vectors.

4. Experimental Results

Since there is no standard dataset for experimentation, we construct our dataset for evaluating the proposed method. Our dataset consists of a total of 2959 videos, of which 817 Shaky Camera samples and 2142 Non-shaky Camera samples. The non-shaky camera samples provide arbitrarily moving videos and hence 817 are considered arbitrary video class. Out of 2142 Non-shaky Camera samples, 1268 are static videos, and the rest, 874 videos, are uniform moving videos.

Furthermore, the dataset includes the video captured day and night of protected and sensitive areas, which includes indoor and outdoor scenes. In addition, the video captured by a shaky camera generates the video of arbitrary moving objects. It also includes the video, which contains leaf and tree movements along with the objects. Therefore, the dataset is complex and challenging for the classification of arbitrary moving video and static video.

For evaluating the performance of the proposed and existing methods, we consider the following standard measures. Precision, Recall, F1-Score, and Accuracy.

Accuracy: In a given dataset consisting of (TP+TN) data points, the accuracy is equal to the ratio of total correct predictions (TP + TN + FP + FN) by the classifier to the total data points. The model's accuracy can be calculated as defined in Equation (10).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad 0.0 < \text{Accuracy} < 1.0 \tag{10}$$

TP-True Positive; TN- True Negative; FP-False Positive; and FN-False Negative.

Precision: This is equal to the ratio of the True Positive (TP) samples to the sum of True Positive (TP) and False Positive (FP) samples, which is defined as in Equation (11).

$$\text{Precision} = \frac{TP}{\text{TP} + \text{FP}} \tag{11}$$

Recall: Recall is the evaluation metrics equal to the ratio of the True Positive (TP) data samples to the sum of True Positive (TP) and False Negative (FN) data samples, which is defined as in Equation (12).

$$\text{Recall} = \frac{TP}{\text{TP} + \text{FN}} \tag{12}$$

F1 Score: F1 Score is equal to the harmonic mean of Recall value and Precision value. The F1 Score gives the perfect balance between Precision and Recall thereby providing a correct evaluation of the model's performance. F1 Score can be calculated as defined in Equation (13).

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

For a comparative study, we implement two state-of-the-art methods, namely, Boufares et al. [3] and Wang et al. [4]. The reason for choosing these two methods is that the objective of both methods is the same as the proposed method. In addition, the methods [1, 2] consider temporal information for moving object detection, which is similar to the proposed method.

Implementation: The following Hyperparameter values are used for successful classification.

n_estimators = 800, *, criterion = 'gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='sqrt', max_leaf_nodes=None, min_impurity_decrease=0.0, bootstrap=True, oob_score=False, n_jobs=-1,random_state=7, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None.

4.1. Ablation Study

In this work, we used the random forest classifier for the classification of arbitrary moving video and static video. To test the contribution of the random forest classifier, we compare the performance of the proposed method-random classifier with other well-known classifiers. For this experiment, the proposed work calculates accuracy for replacing the random forest classifier with the following classifiers on our dataset, and the results are reported in Table 1.

Decision Tree Classifier: A tree-structured classifier where internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome. There are two nodes in decision trees, which are the Decision Node and Leaf Node; Decision nodes are used to make any decision and can have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed based on features of the given dataset and is a graphical representation to get all the possible solutions to a problem based on given conditions.

Gradient Boosting Classifier: It is one of the boosting algorithms used to minimize the bias error of the model. Gradient boosting algorithms can be used for predicting not only continuous target variables (as a Regressor) but also categorical target variables (as a Classifier). When it is used as a regressor, the cost function is Mean Square Error (MSE), and when it is used as a classifier, then the cost function is Log loss.

Support Vector Classifier: Support vector classifier (SVC) is usually preferred for data analysis because of its computational capability within a very short time frame. This classifier works on the decision boundary concept Recognized as a hyperplane. The hyperplane is used to classify the input data into the required target group. The support vector classifier is not affected by overfitting problems, which makes it more reliable.

Table 1. Accuracy for the proposed method with different classifiers

Classifiers	Random Forest	Decision Tree	SVC	Gradient Boosting
Accuracy	87.14	78.57	82.85	81.19

It is noted from Table 1 that the proposed method with a random forest classifier reports the best accuracy compared to all other classifiers. The reason for the best performance of RF is as follows. The RF is an ensemble learning technique that builds many decision trees during training and outputs the mean prediction for regression or the majority vote of the classes for classification. As a result, there is less overfitting, which often impairs the effectiveness of a single decision tree (DT). Secondly, RF performs better in our study than SVC since it does not require the scaling of input features and can handle high-dimensional spaces and a large number of training samples. Finally, RF can operate more efficiently in parallel across several processors than GB, which might make it quicker, especially on more enormous datasets. This could lead to more effective model tweaking and, ultimately, higher performance. Nevertheless, it is essential to note that the effectiveness of these algorithms is highly dependent on the type of data, the particular problem being solved, and the tuning of hyperparameters. Therefore, one can infer that the proposed method with a random forest classifier is suitable for this work. When we compare the results of the proposed method with other classifiers, the Logical Regression approach is better than other classifiers. This is because the Logical Regression can cope with the imbalanced feature vectors, and it avoids the overfitting problem. However, other methods are good when the data is simple, but not for non-linear data.

4.2. Experiments on Classification of Arbitrarily Moving Video

Confusion matrix and Accuracy of the proposed and existing methods are reported in Table 2 and Table 3 on our dataset. Table 2 shows that the proposed approach obtains the highest classification rate for uniform video while lowest for Arbitrary video. For static, our method obtains neither high nor low. This shows that when objects in the video move in uniform direction, the regression step works well. The reason is that separating object which moves in a particular direction with constant speed from the background is easier than the static and arbitrary objects. Therefore, the method achieves the best classification rate for uniform video compared to other two videos. For arbitrary video, since predicting movements is difficult, the method achieves the lowest classification rate and hence arbitrary videos are misclassified as uniform video. However, when we look at overall accuracy, one can conclude that the method is promising and impressive for the classification of static, uniform and arbitrary videos.

Table 3 provides the precision, recall, F-measure and accuracy of the proposed and existing methods, where it is noted that the proposed method is the best in terms of accuracy compared to the two existing methods. This indicates that the proposed method is capable of classifying static, uniform, and arbitrary moving video irrespective of day, night, shaky, non-shaky camera, and distance variation between the object and camera. This makes sense because the key steps of depth estimation and feature extraction based on locally weighted linear regression are invariant to the challenges posed by the input video. On

the other hand, since the primary goal of the existing method is to detect moving objects, the methods are limited to particular objects and tracing. As a result, the methods are not effective for videos containing multiple objects with different speeds and directions. Therefore, the methods do not classify the video successfully as static, uniform, or arbitrary moving video. When we compare the results of the existing methods, the approach [6] is better than the other two existing approaches. This is because the approach [6] uses an adaptive mechanism which is robust to moving object separation from the background compared to the steps used in [3, 4]. However, when we compare the results of [6] with the results of the proposed approach, the performance is worse.

Table 2. Confusion matrix of the proposed method for classification of static, uniform and arbitrary video (Average classification rate is mean of diagonal elements of confusion matrix)

Classes	Static	Uniform	Arbitrary
Static	75.0	7.49	17.5
Uniform	12.01	87.98	0.0
Arbitrary	0.0	50.0	50.0
Average Classification Rate	70.99		

Table 3. Performance of the proposed and existing methods for classification of arbitrary moving video

Classes	Proposed									Boufares et al. [3]									Wang et al. [4]									Rahiminezhad et al. [6]								
	Static			Uniform			Arbitrary			Static			Uniform			Arbitrary			Static			Uniform			Arbitrary			Static			Uniform			Arbitrary		
Measures	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Results	0.75	1.00	0.86	0.88	1.00	0.93	1.00	0.50	0.67	0.60	0.65	0.63	0.58	0.57	0.58	0.62	0.58	0.60	0.59	0.73	0.65	0.62	0.50	0.57	0.59	0.54	0.58	0.62	0.64	0.60	0.68	0.64	0.69	0.58	0.64	0.60
Accuracy	**0.87**									0.60									0.60									0.62								

4.3. Limitation

Although the proposed method is robust to the classification of video of different situations, sometimes, it fails to perform well for the video frames shown in Fig. 6. In the case of night video, the fog and weather conditions may make the objects invisible as shown in Fig. 6. When the objects are not visible in the video, the proposed method fails to find visual cues for feature extraction. In the same way, if the video contains more number objects with arbitrary movement, direction, speed and shapes, the performance of the method degrades. It is evident from the results on arbitrary video in Table 2 and Table 3. Therefore, one can conclude that the number of classes may not increase, but the complexity of classification increases when video contains large variations, which is beyond the scope of the proposed work. This can be solved by extracting dynamic adaptive features, which will be our target of future work through an end-to-end deep learning model.

Fig. 6. An example of the limitations of the proposed method

5. Conclusion and Future Work

We have proposed a new method for the classification of static, uniform, and arbitrary moving video captured by shaky and non-shaky cameras day and night. The motivation for classification is to reduce the complexity of the problem such that the performance of classification and identification of objects in the video can be improved significantly. To neutralize the effect of variations in distance between objects and cameras located at different angles and directions, the proposed work introduced new steps for depth estimation. The features are extracted from the depth images based on locally weighted linear regression, and the features are fed to random forests for the classification of static, uniform, and arbitrary moving videos. The results on a large dataset, which includes video captured by shaky and non-shaky cameras day and night, show that the proposed method is superior to the existing methods in terms of accuracy. However, sometimes, when the video frames are affected by foggy, snow, and night effects, which may make objects invisible in the video, the proposed method does not perform well. Our future plan is to explore an end-to-end deep learning model which comprises a module for video enhancement and one more module for classification.

Acknowledgment

This project was funded by the Ministry of Higher Education of Malaysia for the generous grant Fundamental Research Grant Scheme (FRGS) with code number FRGS/1/2020/ICT02/UM/02/4, the National Natural Science Foundation of China grant no.62136001. The work received partial support from the Technology Innovation Hub, Indian Statistical Institute, Kolkata, India.

Declaration Section

Funding and/or Conflicts of Interests/Competing Interests

Conflict of Interest: The authors declare that they have no conflict of interest.

Data Availability Statement: The data and code will be made available based on request.

References

[1] N. Convertini, V. Dentamaro, D. Impedovo and G. Pirlo, “Sit-to-stand for neurodegenerative diseases video classification”, International Journal of Pattern Recognition and Artificial Intelligence, 2021. DOI : 10.1142/S021800142160003X

[2] L. Nandanwar, P. Shivakumara, U. Pal, T. Lu and M. Blumenstein, “A new hybrid method for caption and scene text classification in action video images”, International Journal of Pattern Recognition and Artificial Intelligence, 2021. DOI : 10.1142/S0218001421600090

[3] O. Boufares, M. Boussif and N. Aloui, “Moving object detection system based on the modified temporal difference and OTSU algorithm”, In Proc. SSD, pp 1378-1382, 2021. DOI : 10.1109/SSD52085.2021.9429516.

[4] Z. Wang, J. Wang and N. Wang, "Moving Object Detection and Marking Based on Frame Difference and Train Algorithm for Teaching Video," 2021 IEEE 15th International Conference on Anti-counterfeiting, Security, and Identification (ASID), Xiamen, China, 2021, pp. 61-65, DOI : 10.1109/ASID52932.2021.9651485.

[5] A. S. B. Sadkhan, S. R. Talebiyan and N. Farzaneh, "An Investigate on Moving Object Tracking and Detection in Images," 2021 1st Babylon International Conference on Information Technology and Science (BICITS), Babil, Iraq, 2021, pp. 69-75, DOI : 10.1109/BICITS51482.2021.9509887.

[6] Rahiminezhad, M. Reza Tavakoli and S. Masoud Sayedi, "Hardware Implementation of Moving Object Detection using Adaptive Coefficient in Performing Background Subtraction Algorithm," 2022 International Conference on Machine Vision and Image Processing (MVIP), Ahvaz, Iran, Islamic Republic of, 2022, pp. 1-5, DOI : 10.1109/MVIP53647.2022.9738764.

[7] M. Sultana, A. Mahmood and S. K. Jung, “Unsupervised moving object detection in complex scenes using adversarial regularization”, IEEE Trans. Multimedia, pp 2005-2018, 2021. DOI : 10.1109/TMM.2020.3006419

[8] M. Tang and W. Liu, "A Moving Object Detection Algorithm for Removing Ghost and Shadow," 2021 40th Chinese Control Conference (CCC), Shanghai, China, 2021, pp. 7207-7212, DOI : 10.23919/CCC52363.2021.9550513.

[9] Y. Shu, Y. Sui, S. Zhao, Z. Cheng and W. Liu, "Small Moving Object Detection and Tracking Based on Event Signals," 2021 7th International Conference on Computer and Communications (ICCC), Chengdu, China, 2021, pp. 792-796, DOI : 10.1109/ICCC54389.2021.9674247.

[10] J. Wang, Y. Zhao, K. Zhang, Q. Wang and X. Li, "Spatio-Temporal Online Matrix Factorization for Multi-Scale Moving Objects Detection," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 2, pp. 743-757, Feb. 2022, DOI : 10.1109/TCSVT.2021.3066675.

[11] W. J. Kim, S. Hwang, J. Lee, S. Woo and S. Lee, "AIBM: Accurate and Instant Background Modeling for Moving Object Detection," in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 7, pp. 9021-9036, July 2022, DOI : 10.1109/TITS.2021.3090092.

[12] Y. Huang, Q. Jiang and Y. Qian, "A Novel Method for Video Moving Object Detection Using Improved Independent Component Analysis," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 6, pp. 2217-2230, June 2021, DOI : 10.1109/TCSVT.2020.3023175.

[13] Z. Deng, Z. Cui and Z. Cao, "Super Resolution Detection Method of Moving Object based on Optical Image Fusion with MMW Radar," IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 2022, pp. 1900-1903, DOI : 10.1109/IGARSS46834.2022.9883395.

[14] R. Kovalenko, A. Tashlinskii and M. Tsaryov, "Using Threshold Processing to Moving Object Detection in the Image Sequence," 2021 International Conference on Information Technology and Nanotechnology (ITNT), Samara, Russian Federation, 2021, pp. 1-4, DOI : 10.1109/ITNT52450.2021.9649334.

[15] Goyal, K., Singhai, J. Recursive-learning-based moving object detection in video with dynamic environment. *Multimed Tools Appl* **80**, 1375–1386 (2021). DOI : 10.1007/s11042-020-09588-w.

[16] Rai, M., Sharma, R., Satapathy, S.C. *et al*. An improved statistical approach for moving object detection in thermal video frames. *Multimed Tools Appl* **81**, 9289–9311 (2022). DOI : 10.1007/s11042-021-11548-x

[17] S. Benaim, S. Ephrat, O. Lang, I. Mosseri, W.T. Freeman, M. Rubinstein, M. Irani and T. Deke, “SpeedNet: Learning the speediness in videos”, In Proc. CVPR, pp 9919-9928, 2020. DOI : 10.48550/arXiv.2004.06130

[18] T. Hosono, K. Sawada, Y. Sun, K. Hayase and J. Shimamura, "Activity Normalization for Activity Detection in Surveillance Videos," 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 2020, pp. 1386-1390, DOI : 10.1109/ICIP40778.2020.9190884.

[19] T. J. Nandhini and K. Thinakaran, "CNN Based Moving Object Detection from Surveillance Video in Comparison with GMM," 2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), Chennai, India, 2022, pp. 1-6, DOI : 10.1109/ICDSAAI55433.2022.10028909.

[20] A. Aliouat, N. Kouadria, M. Maimour and S. Harize, "Region-of-Interest based Video Coding Strategy for Low Bitrate Surveillance Systems," 2022 19th International Multi-Conference on Systems, Signals & Devices (SSD), Sétif, Algeria, 2022, pp. 1357-1362, DOI : 10.1109/SSD54932.2022.9955963.

[21] T. Wang, Y. Gu and G. Gao, "Satellite Video Scene Classification Using Low-Rank Sparse Representation Two-Stream Networks," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-12, 2022, Art no. 5622012, DOI : 10.1109/TGRS.2022.3168602.

[22] R. Zheng, "Sports Video Classification and Labeling Algorithm Based on Data Mining Technology," 2023 International Conference on Distributed Computing and Electrical Circuits and

Electronics (ICDCECE), Ballar, India, 2023, pp. 1-7, DOI : 10.1109/ICDCECE57866.2023.10150741.

[23] M. Asadzadehkaljahi, A. Halder, U. Pal and P. Shivakumara, “Spatiotemporal edges for arbitrarily moving video classification in protected and sensitive scenes”, Artificial Intelligence and Applications, pp 1-8, 2023. DOI: 10.47852/bonviewAIA3202553

[24] M. Asadzadehkaljahi, A. Halder, U. Pal and P. Shivakumara, “Spatiotemporal edges for arbitrarily moving video classification in protected and sensitive scenes”, Artificial Intelligence and Applications, pp 1-8, 2023. DOI: 10.47852/bonviewAIA320526