

Violence Detection using Linear SVM: An Experimental Study

¹Mr. Palash P. Feddewar and ²Prof. Mahendra P. Dhore

¹Assistant Professor, Department of Computer Science

²Pro Vice-Chancellor

¹J. M. Patel College, Bhandara, India,

²Sant Gadge Baba Amravati University, Amravati, India

Abstract: *Violence Detection is one of the challenging tasks in the field of computer vision. In this paper we have carried on an experiment where we have built a model for the task of violence detection. The experiment is carried out on the Hockey Fight dataset. Which is considered as a benchmarked dataset in the field of violence detection. The Experiment shows how the simplest and easiest Machine Learning Algorithm can be used to train a model to detect violence with great precision. Histogram of Oriented Gradients (HOG) is used as the Feature Descriptor to represent Features. Before extracting features, uniform temporal sampling was done to extract similar frames at a regular interval from all videos. This helps in adding temporal coverage of action across the clips. This is a typical two class classification problem where Linear SVM is used as a classifier to classify the input video into two classes as Violent and Non-Violent. This experiment uses a standard split size, where 80% of the data is used to train the model and 20% is used to test the model. The HOG + SVM pipeline is considered as a Machine Learning baseline pipe line. And it has given outstanding results with an accuracy of 99% on Hockey Fight Dataset.*

Keywords: Violence Detection, Machine Learning, Linear SVM, Hockey Fight.

I. INTRODUCTION

The amount of violence has increased drastically in recent years. There are various social, economic and psychological factors behind the increasing violent behavior. This raises the issue of public safety. To ensure public safety, surveillance cameras are installed at public places, malls, streets, etc. the work of surveillance cameras is to monitor human activity for any mishaps. This camera footage can be used to reduce violence and also the footage can be used as evidence for criminal proceedings [1]. But this becomes a tiring task for a human to manually supervise the surveillance cameras for hours. Due to this the major use of surveillance footage gets restricted to investigate after the violence has occurred. So, this leads to automation in the task of violence detection.

Violence detection has attracted a lot of researchers in a few years. But the task of detecting violence in video remains very complex in itself. Due to its complex nature and presence of spatial features for every temporal instance, this task of violence detection invited researchers to solve this problem using more powerful Deep Learning approaches such as CNN, 3D- CNN, ConvLSTM and Attention mechanisms. The Deep Learning Approaches show a clear dominance in recent years with a very high accuracy. But the accuracy achieved by these algorithms comes at a cost. This approach requires a lot of computation power, huge amounts of data and iterability. Deep learning models can automatically extract representative features from raw data, offering a major advantage over traditional machine learning approaches [2].

The major drawback of using Deep Learning is that, to run those algorithms it requires a lot of computational power. In DL algorithms the complete feature extraction process is done by the model itself. The programmer has no control on the type and no of features that are been considered by the algorithm for building the model. Also, it sometimes fails to capture more complex features in a complex environment. The traditional methods of video processing which is build



using a handcrafted feature pipeline are still relevant in the sense that firstly, they are super lightweight compared to DL models. The traditional model is more interpretable as they are built on much simpler mathematical logic which can be tuned, changed or altered with ease. The traditional model remains light weighted and low resolution which can be used as a baseline. Even though this technique remains baseline, it remains comparatively less explored with proper rigorous methodologies in recent literature.

In this paper we stated the experimental results carried on the benchmark HockeyFight [3] dataset using LinearSVM. The Histogram of Oriented Gradients HOG Descriptor [4] was used to extract features from the hockey fight dataset. The extracted features were then used to train the model using a LinearSVM classifier. This was a typical example of two class classification problem, where the input was classified into violent and nonviolent class. This paper is structured in the same pattern, the next section contains the previous related work in this field, followed by the methodology used for experimental setup including the evaluation, followed by result and discussion section and conclusion at the end.

II. RELATED WORK

- **Nievas et al. (2011)** has introduced the benchmark HockeyFight dataset, which contains the clips of the National Hockey League. Total 1000 clips were taken 500 for fight and 500 for non-fight. Each clip is of 50 frames captured at 720 x 576 resolution. Using this dataset they have gained accuracy of 90% using the existing methods.
- **Hassner et al.** — Violent Flows (ViF) proposed the following contributions: (1) a novel approach to real-time detection of breaking violence in crowded scenes. Considered statistics of how flow-vector magnitudes change over time. These statistics, were represented using the Violent Flows (ViF) descriptor. ViF descriptors are then classified as either violent or non-violent using linear SVM. (2) Presented data set of realworld surveillance videos.
- **Deniz et al. (2014)** Features extracted from motion blobs are used to discriminate fight and non-fight sequences. A frequency-based motion-intensity histogram method paired with AdaBoost, reported at 98.9% accuracy
- **Gao et al. (2016):** feature extraction method named Oriented Violent Flows (OVIF), which takes full advantage of the motion magnitude change information in statistical motion orientations. Using combined features with AdaBoost+Linear-SVM achieves improved performance over the state-of-the-art on the Violent-Flows benchmark.
- **S. Das et al. (2019)** HOG is applied to extract lower-level features. Support Vector Machine (SVM), Logistic Regression, Random Forest, Linear Discriminant Analysis (LDA), Naïve Bayes and K-Nearest Neighbors (KNN) are used for the classification purpose. Achieved the highest accuracy rate of 86% while using Random Forest classifier.
- **Zhou P et al. (2018)** Proposed low-level features are the Local Histogram of Oriented Gradient (LHOG) descriptor extracted from RGB images and the Local Histogram of Optical Flow (LHOF) descriptor. Extracted features are coded using Bag of Words (BoW). video-level vectors are classified by Support Vector Machine (SVM)
- **Ferdinando, Hany. (2018)** first detects foreground moving targets via the KNN, then preprocesses the detected targets via morphological processing. proposes a circumscribed rectangular frame integrating method to optimize the circumscribed rectangular frame of moving targets. Obtained approximately 89% accuracy using SVM.
- **Moaaz, Mostafa. (2020)** The model was evaluated on different datasets; like Real Life Violence Situations aka RLVS and Hockey Fight Detection datasets. The accuracy was 92% and 94.5% respectively.



- **-M. R. Abdali et al. (2019)** proposed model consists of CNN as a spatial feature extractor and LSTM as temporal relation learning method with a focus on the three-factor (overall generality - accuracy - fast response time). The suggested model achieved 98% accuracy with speed of 131 frames/sec.
- **Kavathia A et al. (2024)** a CNN was created to analyze optical flow frames of each video. The number of frames analyzed at a time would vary with one, two, three, ten, and twenty frames, and each model would be trained for 20 epochs. The greatest validation accuracy was 94.87%.

III. PROPOSED METHODOLOGY

3.1 Dataset

In this experimental study we have used publicly available HockeyFight Dataset [3] which was introduced by Nievas et al. (2011) in their paper "Violence Detection in Video Using Computer Vision Techniques." The dataset contains all total 1000 videos categories into two categories NonViolence and Violence; each category contains 500 videos. Each video is extracted from the National Hockey League. The extracted videos are of same frame size set as 50 and a fixed resolution of 360 x 288 pixels.

For the experimental purpose the dataset was split into a standard stratified split of 80: 20, where 800 videos (400 from each category) were used to train the model and 200 videos were used to test and validate the model. The major reason of using a stratified sampling is to prevent sampling bias. The complete data is split in proper proportion of 50:50 to maintain the balance. Also, the random state is set to a fixed integer of 42 so that rerunning the script gives the same sample split of train and test data every time.

3.2 Feature Extraction

For the feature extraction process 9 Frames were extracted from every video at an interval of 5 seconds. To reduce the data redundancy videos were sampled using uniform stride to extract nonconsecutive frame. Total 4504 frames were extracted from violence category and 4499 were extracted from Non-Violence category, both with a same frame size (288,360,3). For extraction process HOG Descriptor [4] was chosen.

The feature extraction was done using the standard HOG setup (8 x 8 pixel cell, 2 x 2 cell blocks, 9 orientation bins). Firstly, the pixels were converted to grid. Our pixel size was (288, 360), so the conversion becomes as:

Grid Cells:

$$\text{Vertical: } 288 / 8 = 36$$

$$\text{Horizontal: } 360 / 8 = 45$$

So, the grid size comes out to be 36 x 45

Then to calculate the overlapping blocks, a 2 x 2 block is slide across which decreases the size of block as:

$$\text{Vertical Step: } 36-2+1 = 35 \text{ blocks}$$

$$\text{Horizontal Step: } 45-2+1 = 44 \text{ blocks}$$

$$\text{Total Blocks: } 35 \times 44 = 1540 \text{ blocks}$$

To provide robustness, Euclidian normalization (L2 normalization) is used on each local vector independently. The total no of blocks inside each cell is 4. So, the total cell instance comes out to be:

$$4 \times 1540 = 6160 \text{ dim per frame}$$

The feature which are calculated per block is

$$4 \text{ cells} \times 9 \text{ bins} = 36 \text{ features}$$

So finally, the total dimension of HOG features comes out to be:

$$36 \text{ features} \times 1540 \text{ blocks} = 55440 \text{ dimension per video.}$$

Therefore, the shape of HOG features is (1000, 55440)

A video analysis requires both spatial as well as temporal features, so, temporal stride sampling technique is applied across the all-consecutive frame. And sampling frames at fixed strides reduces computation cost too, while preserving the important information. [14]. And lastly the descriptors are aggregated across time using sequential



feature concentration rather than using the mean pooling to preserve the coarse frame ordering as HOG features does not capture true motion dynamics or optical flow.

3.3 Classifier

A linear kernel is used as a classifier over a nonlinear kernel as the dimension is very high of the concentrated feature space. As the feature space dimension (55440) is more than the training sample, the data become sparse and is mostly linearly separable. And also mapping feature into higher dimension using nonlinear kernels provides no performance and there are high chances of model getting overfitted. In SVM the decision boundaries are made through just simple dot product of the support vectors rather than using complex function, this lowers the computation complexity.

LinearSVC optimize was chosen as an solver for SVM. The SVC finds the hyperplane

$$W^T x_i + b = 0$$

That separates the class with maximum margin.

The optimization problem is formulated as:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + c \sum_i \max(0, 1 - y_i(w^T x_i + b))$$

The function uses the empirical error minimization via hinge loss function to optimize hyperplane by maximizing geometric margin.

L2 regularization is used. Keeping the regularization parameter $C = 1$, controlling the penalization of the structural margin while keeping the weights small. To ensure the optimal convergence the maximum limit of iteration was set to 5000. Also, in this experiment no standard scaling was done for feature scaling.

3.4 Evaluation Protocol

Accuracy is considered as a standard evaluation protocol for classification problem. But when a model is trained for the specific task of violence detection, accuracy itself is not sufficient enough to evaluate the model. The reason behind is, when deployed such models shows a huge amount of class imbalance. So even a model with high accuracy can fail in real world situations. The accuracy metrics ignores the asymmetric cost of real-world classification error. False negative on a violent class carries an unacceptable high cost as any misclassification can lead to a life threatening event. So, the following other parameters are also considered for evaluating the models:

Precision: It is a important metrices for the violent class. As it deals with the fact that if everything id flagged as violence, what fraction of the data was actual violence. The higher the value of precision, lower is the rate of false alarm.

Recall: The second important metrices for evaluating violent class is Recall. It tells the value that among all the violence data, how much violence was actually detected by the model. A low recall rate is non accepted in the sensitive field of violence detection.

F1- Score: The value is dependent on the earlier two metrices i.e. precision and recall. But F -Score is calculated using harmonic mean, meaning if any of the value drops, it can adversely affect the F1- Score. Does helping us to know the imbalance.

Confusion Matrix: The Confusion matrix tells that where error is exactly occurring.

IV. EXPERIMENTAL RESULTS

In this section we will discuss the quantitative results of the proposed a model which take spatio – temporal features extracted from HOG Descriptor and uses LinearSVM for classifying violent videos.

4.1 Quantitative Classification

The evaluation was carried on total 55440-D feature spare for total 200 videos, 100 for violence and 100 for non-violence. The complete per class classification is given in the table below



Table I: Classification Report

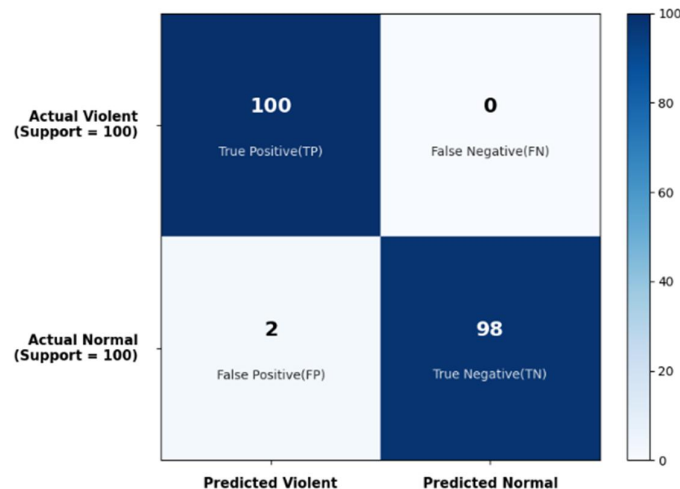
	Precision	Recall	F1- Score	Support
Class 0 (Normal)	0.9804	1.0000	0.9901	100
Class 1 (Violence)	1.0000	0.9800	0.9899	100
Micro Average	0.9902	0.9900	0.9900	200
Weighted Average	0.9902	0.9900	0.9900	200
Accuracy		0.99		200

The values of the micro average and weighted average are identical because our testing environment was perfectly balance. The results show that for the class 1 which is violence class the recall 0.9802 is nearly perfect. This proves that the structural gradient boundaries preserve a good amount of discriminative variance for the LinearSVM to isolate feature using the 8 x 8 pixel HOG pooling. Also, the score 1.000 for the precision of Class 1 indicate no false alarm were flagged during the test and validation. The experimental results shows that the model has got a nearly perfect accuracy of 0.99.

4.2 Confusion Matrix

The confusion matrix for the classification results is given below. Here we can clearly see that there are 2 clips which falls in the false negative category. Which means that the classifier has misclassified 2 clips. Those clips belong to the Normal Class but were miss classified to violent class. This type of miss classification is still considered safe than generating the false alarm.

Table II: Confusion Matrix
Your Custom Colorful Confusion Matrix



V. DISCUSSION

The clear class separation was possible due to the projection of HOG descriptor in 55440-dimension space. As the dimensionality (d) was very high than the training samples (n), the distance between any two points tends to maximize, which helps the sample vectors to form a geometric cluster. And statistically these highly sparse clusters are linearly separable. As the dimensionality was so high, there was no point in using a non- linear kernel, the linear SVM was easily able to create an hyperplane that was create a decision boundary in two classes.

5.1 Limitation

No doubt we have achieved very good results in our experiment, but there are several limitations that must be acknowledged.



The frame – order- preserving modeling was used for temporal feature. This mechanism though captures the coarse history of spatial transformation but it fails to truly incorporate true temporal tracking mechanism such as dense optical flow or frame differencing.

One more point to be mention is that the soft margin regularization strength parameter is fixed to $C = 10$ for the liner SVM classifier. The Hyperparameter grid search optimization was not executed.

Before fitting the data feature scaling like Standard Scaling, was not performed. The evaluation was done on single domain, cross domain validation was not performed.

VI. CONCLUSION & FUTURE SCOPE

In this experimental study, we have presented a framework for violence detection. Histogram Oriented Flow (HOG) descriptor along with the LinearSVM classifier is used to build the model. The model prepared is highly efficient in terms of computation power. Standard HOG setup was used on Hockey Fight dataset containing 1000 videos which extracted total 55,440 features per video block. A standard LinearSVM classifier with regularization parameter $C=1$ was used as a classifier, which yields an accuracy and f1 score of 0.99. The feature scope is using different classifiers for the same task and comparing their performance using the same evaluation setup, also including a mechanism for tuning the hyperparameters and lastly evaluation on real life violence dataset.

REFERENCES

1. V. D. Huszár, V. K. Adhikarla, I. Négyesi and C. Krasznay, "Toward Fast and Accurate Violence Detection for Automated Video Surveillance Applications," in *IEEE Access*, vol. 11, pp. 18772-18793, 2023, doi: 10.1109/ACCESS.2023.3245521.
2. H. Khan, X. Yuan, L. Qingge and K. Roy, "Violence Detection From Industrial Surveillance Videos Using Deep Learning," in *IEEE Access*, vol. 13, pp. 15363-15375, 2025, doi: 10.1109/ACCESS.2025.3531213.
3. Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R. (2011). Violence Detection in Video Using Computer Vision Techniques. In: Real, P., Diaz-Pernil, D., Molina-Abril, H., Berciano, A., Kropatsch, W. (eds) *Computer Analysis of Images and Patterns. CAIP 2011. Lecture Notes in Computer Science*, vol 6855. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23678-5_39
4. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 886-893 vol. 1, doi: 10.1109/CVPR.2005.177.
5. Hassner, Tal, Yossi Itcher, and Orit Kliper-Gross. 2012. "Violent Flows: Real-Time Detection of Violent Crowd Behavior." In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1-6. IEEE.
6. Deniz, Oscar & Bueno, Gloria & Kim, Tae-Kyun. (2015). Fast Fight Detection. *PLOS ONE*. 10. e0120448. [10.1371/journal.pone.0120448](https://doi.org/10.1371/journal.pone.0120448).
7. Yuan Gao, Hong Liu, Xiaohu Sun, Can Wang, Yi Liu, Violence detection using Oriented Violent Flows, *Image and Vision Computing*, Volumes 48–49, 2016, Pages 37-41, ISSN 0262-8856, <https://doi.org/10.1016/j.imavis.2016.01.006>.
8. S. Das, A. Sarker and T. Mahmud, "Violence Detection from Videos using HOG Features," 2019 4th International Conference on Electrical Information and Communication Technology (EICT), Khulna, Bangladesh, 2019, pp. 1-5, doi: 10.1109/EICT48899.2019.9068754.
9. Zhou P, Ding Q, Luo H, Hou X (2018) Violence detection in surveillance video using low-level features. *PLoS ONE* 13(10): e0203668. <https://doi.org/10.1371/journal.pone.0203668>
10. Ferdinando, Hany. "A Video-Based DT-SVM School Violence Detecting Algorithm." *Sensors*, n.d. doi:10.3390/S20072018.
11. Moaaz, Mostafa. "Violence Detection In Surveillance Videos Using Deep Learning," 2020. doi:10.21608/FCIHIB.2020.42233.1003.



12. -M. R. Abdali and R. F. Al-Tuma, "Robust Real-Time Violence Detection in Video Using CNN And LSTM," 2019 2nd Scientific Conference of Computer Sciences (SCCS), Baghdad, Iraq, 2019, pp. 104-108, doi: 10.1109/SCCS.2019.8852616.
13. Kavathia, Aarjav & Sayer, Simeon. (2024). Optimizing Violence Detection in Video Classification Accuracy through 3D Convolutional Neural Networks. 10.48550/arXiv.2411.01348.
14. J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund and A. Clapés, "Video Transformers: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 11, pp. 12922-12943, 1 Nov. 2023, doi: 10.1109/TPAMI.2023.3243465.

