Violence Detection and Localization in Surveillance Video

David Gabriel Choqueluque Roman
Department of Computer Science
Universidad Católica San Pablo
Arequipa, Perú
Email: david.choqueluque@ucsp.edu.pe

Guillermo Cámara Chávez
Department of Computer Science
Federal University of Ouro Preto
Ouro Preto, Brazil
Email: guillermo@ufop.edu.br

***Abstract*—Automatic violence detection in video surveillance is crucial for social and personal security. Due to the massive video data produced by surveillance cameras installed in different environments like airports, trains, stadiums, schools, etc., traditional video monitoring by humans operators becomes inefficient. In this context, develop systems capable of detect automatically violent actions is a challenging task. This study describes a method to detect and localize violent acts in video surveillance using dynamic images, CNN's, and weakly supervised localization methods. Experimental results demonstrate the effectiveness of our approach when applied to three public benchmark datasets: Hockey Fight [1], Violent Flows [2], and UCFCrime2Local [3].**

***Keywords*-Violence detection; violence localization; video surveillance; dynamic images; video summarization; saliency detection.**

I. Introduction

Nowadays, conventional video surveillance systems rely heavily on human operators to monitor activities and reduce crime. Unfortunately, many incidents are erroneously detected due to the inherent limitations of deploying humans to watch multiple surveillance videos. Such a problem promotes research of violence detection methods to analyze videos robustly and feasibly. Violence actions can be defined as those an 8-year-old child should not watch because of physical violence [4]. In the video-based violence detection literature, the approaches can be broken into two categories: hand-craft methods and deep learning methods.

Handcraft methods represent video actions using a predefined method to overcome specific problems like occlusions, variable illumination, and scales. Generally, these kinds of methods build descriptors around interest points [5]–[8], and dense information [9]–[13]. Then, the descriptors are codified using the *Bag-of-visual-words* technique. Finally, The codified descriptors are fed to an *SVM*, *Random Forest* or *AdaBoost* classifier.

On the other hand, deep learning methods train convolutional networks to extract low and high-level features. Multi-stream CNN architectures are proposed by [14]–[16] to obtain visual and motion information from violent videos. A 3D-CNN architecture is proposed in [17], [18] to extract spatio-temporal information. In [19], the authors propose a hybrid hand-crafted/CNN framework to encode motion information in images. In [20], [21] convolutional and recurrent neural networks are used to capture long spatio-temporal information. Although actual methods achieve good results detecting violent actions in videos, there are only a few methods that can localize violent regions in frames like [22] and [23].

In [14], the authors propose a two-stream CNN architecture and a SVM classifier. The approach consists of three stages: feature extraction, classification training, and label fusion. Each stream CNN uses an Imagenet pre-trained *VGG-f* architecture. The first stream extracts visual features, while the second extract motion features from consecutive frame differences. Then, two SVM classifiers are trained with visual and motion information, respectively. Finally, the detection result is obtained with a label fusion method. The main advantage of this method is its low processing time. However, this method does not detect violent actions between people at close range, making it challenging to detect violence in settings with many people. Meng *et al.* [15] took a similar approach of two-stream CNNs. They propose to integrate CNNs with improved trajectories to capture long temporal information. They use two *VGG-19* networks to extract spatial and temporal information. Spatial information is extracted from video frames, and temporal information is extracted from dense optical flow images.

Malveira *et al.* [16] focus on the idea of breaking down violence into more specific concepts such as fighting, shooting, explosions, presence of blood, fire, firearms, and knives. First, they detect movements and combine frames to train a CNN for each concept. The main disadvantage is their dependence on video quality and their high computational cost due to the large number of CNNs they use. In [24], the authors focus on detecting violence in dense scenarios; for that, they train two CNNs, one with RGB images and the other with Optical Flow images. Although his method achieves high levels of precision, one of the shortcomings of the technique is the high computational cost.

The authors in [17] modify a 3D convolutional network and propose a new preprocessing method based on keyframes and video length. Their method performs a uniform sampling in videos. A hybrid model of hand-crafted features and deep learning is proposed in [19], they extract features by combining the FAST corner detector, the BRISK descriptor, and Hough Forests classifier to obtain representative images and

train a CNN. The disadvantages of its method are that it can only detect a specific type of violence (fighting), and its best precision result depends on the scene condition with a static camera and a short distance between people and the camera.

Other studies explore temporal information that represents violent actions using recurrent networks. In [20], propose a method combining a pre-trained ResNet-50 by adding a pyramid grouping method in the final layer and feeding a bidirectional network LSTM. This method achieves low levels of precision compared to other methods focused on deep learning. Sudhakaran *et al.* [21] propose an improvement in the architecture of [20], by replacing LSTM cells with convolutional LSTM cells, achieving higher levels of precision.

In this research work, we propose and analyze a method for detecting and locating violence in video sequences. The method consists of four main stages. Initially, a video sequence is summarized into an image using the *rank pooling* method [25] in order to represent motion information through an image called *dynamic image* [25]. Then, a CNN classifier is trained on top of such dynamic images to learn the violent motion appearance. This *violence classifier* detects whether a video sequence (or dynamic image that summarizes it) has violence or not. Due to the lack of a publicly available violence dataset with spatial annotations, we propose a weakly-supervised approach as the third stage of our approach. This stage consists of generating a saliency mask from dynamic images using the knowledge of the *violence classifier* and produce region proposals with movement. Finally, a refinement method processes the saliency masks with region proposals to get bounding boxes with the violent regions only. This step is more detailed in Section II-D. The experiments were conducted on three public datasets like Hockey Fight [1], Violent Flows [2] and UCFCrime2Local [3].

In summary, the main contributions of this work are:

- A methodology for temporal detection of violent actions in videos.
- We propose an approach for identifying the spatial localization in the frame where the violent action is performed.
- Our detector is based on dynamic images, *i.e.* a video is resumed in one or very few images, making it possible for our proposal to be used in real-time applications.

II. Proposed Method

In this section, we present our approach for violence detection and localization in the video. The main stages of the proposed method are illustrated in Fig. 1. Firstly, a video is summarized into one or multiple dynamic images. Then, a deep neural network is used to learn a video classifier on top of such images. Next, if the classifier detects violence in the video, a weakly supervised mask model is used to manipulate the scores of the classifier by masking saliency regions of a violent dynamic image, generating region proposals with motion. Finally, a refinement method is applied to such region proposals to get accurate violent regions. Each stage is detailed in the following sections.

A. Video Summarizing

Initially, an input video is divided into N video segments with length T (number of frames). Then, each segment is summarized to a RGB image called *dynamic image*. Dynamic images are constructed with the *Rank Pooling* [26] method. Rank pooling represents a video as parameters of a linear ranking function that is able to order through time a sequence of frames $I_1, I_2, ..., I_T$. Precisely, let $\varphi(I_t) \in R^d$ be an operator that stacks RGB channels of each pixel in image I_t on a large vector and $V_t = \frac{1}{t}\sum_{T=1}^{t}\varphi(I_t)$ be the average of these large vectors up to time t. The ranking function $S(t|d)$ predicts a ranking score for each frame at time t denoted as $S(t|d) = \langle d, V_t\rangle$, where $d \in R^d$ are the parameters of the ranking function [26]. Learning d is posed as a convex optimization problem using the RankSVM formulation given as Equation 1.

$$\begin{aligned} d^* = \rho(I_1, ..., I_T; \varphi) = \underset{d}{\operatorname{argmin}} E(d) \\ E(d) = \frac{\lambda}{2}\|d\|^2 + \frac{2}{T(T-1)} \times \\ \sum_{q>t} \max\{0, 1 - S(q|d) + S(t|d)\} \end{aligned} \tag{1}$$

Optimizing Equation 1 defines a function $\rho(I_1, ..., I_T; \varphi)$ that maps a video sequence of length T to a single vector denoted by d^*. This output d^* has the same dimensions as input images, and it is called dynamic image. In [25], an a faster approximation of rank pooling called *approximate rank pooling* was presented. In this work, we used this approximation in order to process video sequences in a fast manner. Approximate rank pooling computes d^* by the following equation.

$$d^* = \sum_{t=1}^{T} \alpha_t I_t \tag{2}$$

the coefficients α are given by $\alpha = 2(T - t + 1) - (T + 1)(H_T - H_{t-1})$, where $H_t = \sum_{i=1}^{t} 1/i$ is the t-th Harmonic number and $H_0 = 0$.

B. Violence Classifier

After video summarizing, a CNN is trained on top of dynamic images in order to classify an input video as violent or not violent. The goal of this model is to determine if an input video has violent content or not, using violent motion appearance. The CNN architecture is illustrated in Fig. 2. The architecture is based on dynamic image networks proposed in [25]. After splitting and summarizing an input video into one or multiple dynamic images, the model extracts visual features from them using the convolutional layers of a pre-trained CNN. As illustrated in Fig. 2, the last convolutional layer is followed by a *temporal max pooling* layer [25] in order to aggregate feature maps over time into one, and it extracts long temporal information. Similar to [27], a batch normalization layer is added before the first fully-connected layer.

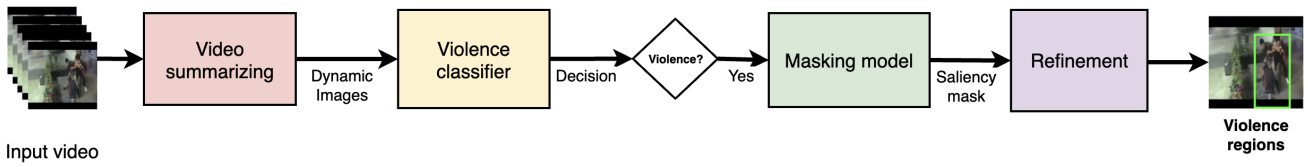


Fig. 1. Stages of the proposed method. In the first step, a video sequence is summarized into dynamic images using the *rank pooling* method. Then, a trained CNN classifies these images as a video sequence with violent or not violent content. If the video is categorized as violent, then a weakly-supervised approach generates saliency maps with violence region proposals from dynamic images. Finally, in the fourth step, a refinement method is applied to get bounded violent regions.

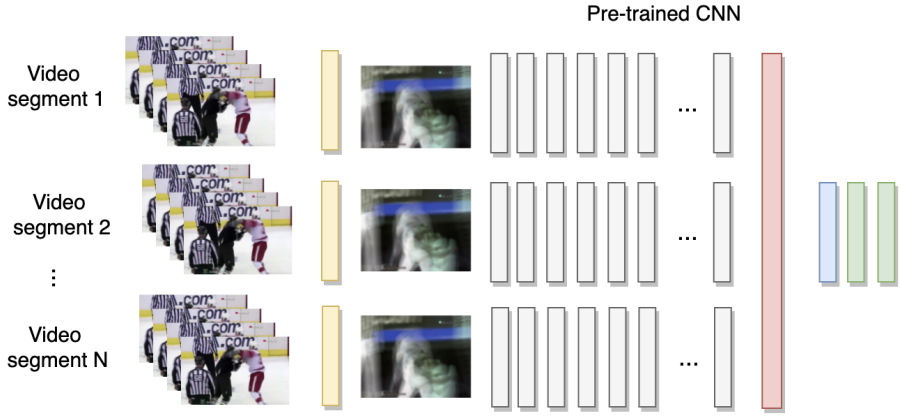


Fig. 2. The architecture of the violence classifier model. The model consists of a rank pooling layer (yellow), which can be thought of as a zero layer. The convolutional layers (gray) correspond to a pre-trained CNN model. The last convolutional layer is followed by a temporal max-pooling (red), and a batch normalization (blue) layer. The fully-connected layers are shown in green color.

C. Masking Model

Salient regions in a dynamic image are regions with movement in a sequence of video frames. In order to produce a saliency mask, we train a weakly supervised model proposed in [28]. Fig. 3 illustrates the architecture of the masking model. The model adapts a U-Net [29]. The encoder part is a ResNet-50 [30] pre-trained in ImageNet [31], and it downsamples the input by a factor of 32. The feature map from Scale 5 passes through a filter. The feature filter performs the initial localization attenuating spatial locations which contents do not correspond to the selected class (violence and non-violence). The class selector is just a class embedding that multiplies the feature map into the filter. The output of the feature filter Y at the spatial location i, j is given by:

$$Y_{ij} = X_{ij}\sigma(X_{ij}^T C_s) \tag{3}$$

where X_{ij} is the output of the Scale 5 at spatial location i, j. C_s is the embedding of the selected class s and $\sigma(.)$ is the sigmoid nonlinearity.

The initial localization, given by the filter, is fine-tuned by the decoder part of the model. The decoder part is composed of transposed convolutions that upsample a low-resolution feature map by a factor of two. Every upsampler block concatenates the upsampled feature map with the corresponding feature map from the decoder part. The model learns which parts of a dynamic image are considered salient by the violence classifier, minimizing the following objective function:

$$\begin{aligned} L(M) = \lambda_1 TV(M) + \lambda_2 AV(M) - \\ \log(f_c(\Phi(X, M))) + \lambda_3 f_c(\Phi(X, 1 - M))^{\lambda_4} \end{aligned} \tag{4}$$

where X is the original dynamic image, M is the mask, and f_c the *softmax* probability of the class c of the violence classifier. $TV(M)$ is the total variation of the mask defined as:

$$TV(M) = \sum_{i,j}(M_{i,j} - M_{i,j+1})^2 + \sum_{i,j}(M_{i,j} - M_{i+1,j})^2 \tag{5}$$

$AV(M)$ is the average of the mask elements and takes a value between 0 and 1. The constants λ_i are regularizers. The function Φ is the mask applied to the image, avoiding the introduction of artifacts. It is defined as:

$$\Phi(X, M) = X \odot M + A \odot (1 - M) \tag{6}$$

where A is a blurred version of X. This blurred image is useful to minimize introduced evidence during image saliency detection [28].

D. Refinement

Fig. 7 shows the outputs of our proposed refinement method. In contrast to existing methods, our localization method finds

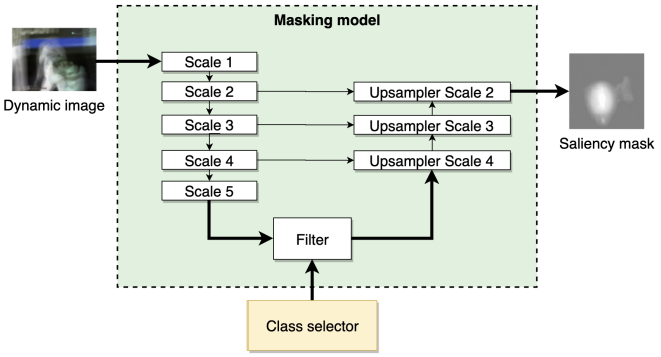


Fig. 3. Architecture of the masking model proposed in [28].

moving region proposals in dynamic images. After computing a saliency mask, we use it to generate region proposals. To detect moving regions under various motion contrast, we apply an adaptive thresholding method [32] over the saliency mask. To alleviate the problem of incomplete regions with some parts of the actors only, we apply morphological transformations to filter small regions and get complete moving region proposals. The results of the salient motion detection (sixth column of Fig. 7) are region proposals with violent movement. In most cases, these proposals are smaller than a person because the movements are centered in the arms and legs of the actors during violent actions. Therefore, to alleviate this problem, an object detector is applied to obtain regions with persons (seventh column of Fig. 7). Then, the salient regions close to a detected person are joined together to the person region. The final localization is the region with the greater area. Additionally, the person detector is useful to differentiate between noise movement and human movement. In this study, we experiment with two state-of-the-art pre-trained object detectors like *Yolo V3* [33] and *Mask R-CNN* [34]. The object detector is applied only in some frames of the video segment to reduce the computational cost.

III. Experiments and Results

The proposed method in classifying and detecting violence in video evaluates its effectiveness in three benchmark datasets, reporting metrics like classification accuracy and localization error.

A. Datasets

The performance of the proposed method is evaluated on three standard public datasets namely, Hockey Fight [1], Violent-Flows [2] and UCFCrime2Local [3]. They contain real-world videos captured using CCTV cameras.

Hockey Fight dataset: This dataset is composed of 1000 video clips with a resolution of 360×228 pixels, collected from hockey games, and recorded by moving camera. All video clips have between 40 and 50 frames. Half of them (500 clips) are labeled as fight and another half as non-fight. Almost all the videos in the dataset have a similar background, duration, and subjects. Figure 4 shows some examples of the dataset.

Fig. 4. Frames captured from the Hockey Fight dataset. Frames in the first row are violence samples while non-violence in the second row.

Fig. 5. Frames captured from the Violent-Flows dataset. Frames in the first row are violence samples while non-violence in the second row.

Violent-Flows dataset: This dataset consists of videos in crowded scenarios where the number of people taking part in violent events is vast. Football matches compose most of the videos in this dataset. The dataset has 246 videos with a resolution of 320×240 pixels. Figure 5 shows some examples of the dataset.

UCFCrime2Local dataset: This dataset enriches a portion of UCF-Crime dataset [35] with spatio-temporal annotations (bounding boxes) with particular attention to human-based anomalies in six categories: arrest, assault, burglary, robbery, stealing, and vandalism. The dataset contains 100 anomalous videos and 200 normal videos. All the videos are long real-world surveillance videos with anomalies of a significant impact on public safety. We only use in our experiments the following violent categories: Arrest, Assault, Robbery, and Stealing. Because the videos have a long duration, they have positive and negative instances. Each positive sample has a max of three temporal instances of violence. During experiments, all positive instances were used as violent samples. For normal videos, we randomly chose videos from negative instances and original normal videos, respectively. Finally, the reduced dataset has 46 violent videos and 45 normal videos. Figure 6 shows some examples of the dataset.

B. Implementation details and Results

In this section, we present the specific experimental steps of the violence classifier, masking model, and refinement step training and results. All models are implemented using the Torch library.

Violence Classifier training: We fine-tuned three base models: AlexNet [36], ResNet-50 [30] and DenseNet [37], pre-trained on ImageNet to identify optimal values for the param-

Fig. 6. Frames captured from the UCFCrime2Local dataset. Frames in the first row are violence samples while non-violence in the second row.

eters N and T, which determine the number of segments and the segment length, respectively. We run this experiment on all datasets using 5-fold cross-validation. On the training subsets, we finetune all the layers of the respective models. During our experiments, we analyze the impact of the parameters N and T in violence detection.

Segment length (T): Figure 8a shows the classification accuracy using a dynamic image per video with different segment lengths using an AlexNet as the classifier. We can see that increasing the number of frames per dynamic image decreases the classifier performance (For the Hockey Fight dataset, the performance decreases using more than 40 frames). This is because a dynamic image can be contaminated by background motion when the number of frames increases. In Figure 9, we show some examples of dynamic images varying the parameter T. The best performance of the violence classifier is 95.5%, 82.0%, and 79.0% for the Hockey Fight, Violent Flows, and UCFCrime2Local datasets, respectively.

Multiple Dynamic images (N): Analyzing the previous results, we train a ResNet-50 classifier using short segments of video frames (10 frames), but now using more dynamic images per video. Figure 8b shows the classification accuracy using a different number of dynamic images per video. Using multiple short segments reduces background noise, and the classifier learns to model violent motions as a combination of complex motions. The classifier achieves a 96.4%, 92.0%, and 79.1% of accuracy for the Hockey Fight, Violent Flows and UCFCrime2Local datasets, respectively.

State of the Art comparison: Table I shows the classification accuracy for the Hockey Fight, Violent Flows and UCFCrime2Local datasets. We can see that the proposed method outperforms most of the methods and achieves competitive results with state of the art. Our method is outperformed by a margin of one to two percentage degrees by [10], [21], and [38] in the Hockey Fight, and Violent Flows datasets. However, most of the methods shown in Table I can only detect violence in videos in the temporal dimension. In contrast, our method can detect and localize the violence in a two-stage approach, identifying the region in the frame where the violence occurs. To the best of our knowledge, in the literature of violence detection, only a few methods like [38] can detect and localize violence.

Method generalization: In order to evaluate the capability of the method to generalize in the recognition of types of violence different from those viewed during training, we evaluate the classifier performance training on a dataset and testing in another one. Table II shows the results. We can see that using the Hockey Fight dataset during training, and testing on the Violent Flows dataset, the classifier achieves a 62.2% of accuracy. The poor performance is because of the big difference between the two datasets. Using two more similar datasets like Hockey Fight (train) and UCFCrime2Local (testing), the classifier achieves a 50.5% of accuracy. Increasing the training set with the Violent Flows dataset, increase the classifier accuracy to 58.24%. Analyzing the obtained results, we can conclude that the violence classifier based on dynamic images has a low capability of generalization in different kinds of violence such as fights, robberies, arrests, and crowd violence.

TABLE I
COMPARISON OF CLASSIFICATION RESULTS FOR THE HOCKEY FIGHT, VIOLENT FLOWS AND UCFCRIME2LOCAL DATASETS.

Method	Hockey Fight	Violent Flows	UCFCrime2Local
[39]	82.40%	-	-
[40]	90.10±0%	-	-
[9]	95.00%	94.31±1.65%	-
[38]	96.80±1.04%	93.19±0.12%	-
[13]	95.80%	95.11%	-
[11]	95.00%	-	-
[10]	**98.20±0.76**%	93.09±1.14%	-
[7]	88.60±1.2%	85.83 ±4.26%	-
[12]	92.79±3.05%	92.29%	-
[41]	81.25±0.59%	85.43±0.21%	-
[42]	89.30±0.91%	76.83±1.76%	-
[43]	89.10%	88.21%	-
[6]	96.50%	-	-
[20]	83.19%	-	-
[21]	97.10±0.55%	**94.57±2.34**%	-
[44]	95.90±3.53%	93.25±2.34%	-
[19]	94.60±0.6%	-	-
Our method	96.40±0.3%	92.0 ±0.14%	79.1 ±0.191%

TABLE II
GENERALIZATION EXPERIMENT RESULTS.

Dataset train	Dataset test	Accuracy
Hockey Fight	Violent flows	62.2 %
Hockey Fight	UCFCrime2Local	50.5 %
Hockey Fight + Violent flows	UCFCrime2Local	58.24 %

Mask Model training: The masking model training is about minimizing the function from equation 4. We use 30 epochs to train our model. The parameter values are $\lambda_1 = 8$, $\lambda_2 = 0.5$, $\lambda_3 = 0.3$ and $\lambda_4 = 0.3$. The training set consists of 73 random samples from the UCFCrime2Local dataset. The test set is composed of the 18 remaining samples.

In order to evaluate our violence localization approach, we use only the UCFCrime2Local dataset using the spatial annotations provided in it. Similar to [28], we evaluate our localization method using the localization error metric. Our method achieves a 46.4% of localization error using the *Yolo V3* detector [33] and a 35.35% of *localization error* using the *Mask R-CNN* detector [34]. In most cases, when our method fails, it is because the object detector does not detect persons

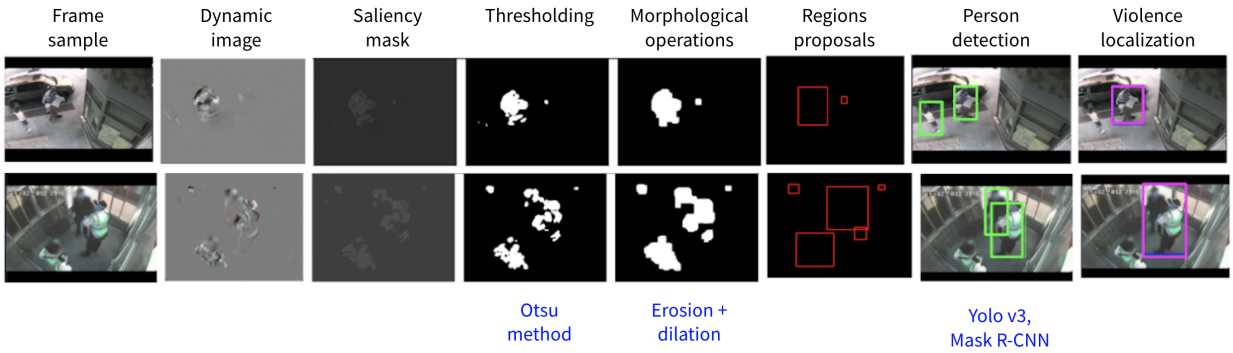
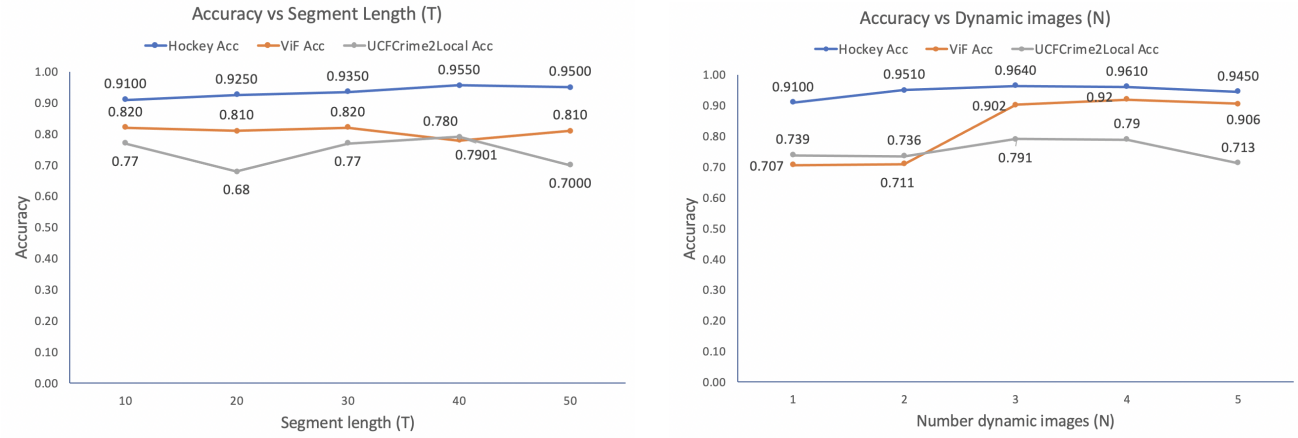


Fig. 7. Outputs of the proposed refinement method.



(a) Accuracy vs segment length (AlexNet)

(b) Accuracy vs Dynamic images (ResNet-50)

Fig. 8. Violence classifier performance using (a) different segment lengths, and (b) multiple dynamic images per video.

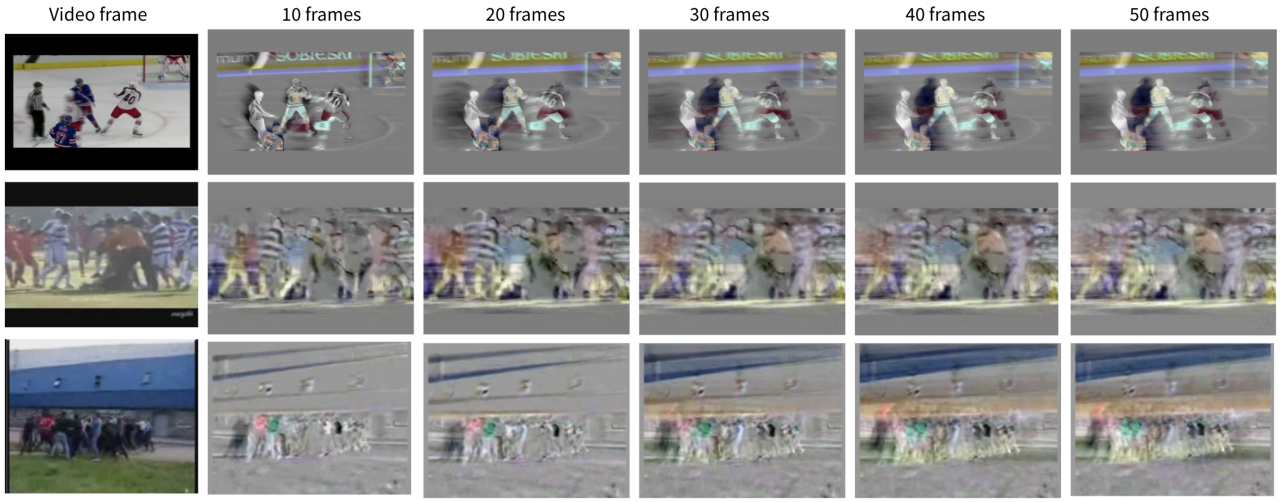


Fig. 9. Qualitative comparisons between dynamic images with different segment length (T). When the number of frames increases, the dynamic image is contaminated by background motion. Each row corresponds to a video sample. The first column shows a frame sample of the video, while the other columns show dynamic images generated with a different number of frames.

in the segment, or the binary segmentation fails.

Impact of the classifier in localization: Because our localization method depends on the classifier performance, we analyze the localization error training the mask model with different classifiers. Table III shows the localization errors using an AlexNet, ResNet50 and DenseNet, respectively. It

can be seen that localization has an improvement when the classifier achieves a good performance. This is because the masking model uses the knowledge of the violence classifier to produce the binary mask. Some examples of the localization results are showed in Figure 10. We can see that the mask model trained with AlexNet classifier has the best performance for the localization task. On the other hand, the mask model trained with a DenseNet classifier has poor performance. Last row of Figure 10 shows a fail localization in all models.

TABLE III
ACCURACY AND LOCALIZATION ERROR FOR THE UCFCRIME2LOCAL DATASET USING DIFFERENT CLASSIFIERS.

	Accuracy	Localization Error(%)
AlexNet	0.79	35.35%
ResNet-50	0.69	42.85%
DenseNet	0.77	85.7%

IV. CONCLUSIONS AND FUTURE WORK

In this work, we have presented a method to detect and localize violent actions using only temporal video annotations. Our method follows a two-stage approach, firstly classifying a video as violent or non-violent, and finally localizing violent regions. Our method is based on dynamic images and convolutional networks, and it achieves close results to state of the art. We analyze the usefulness of dynamic images to represent violent motion in videos. Using dynamic images instead of optical flow frames to represent motion, allowed us to analyze long temporal information, and at the same time, avoid the high computational cost of optical flow. For three different datasets, the video length, and the number of dynamic images per video was studied. The results demonstrate that combining multiple dynamic images of short frame sequences improve the accuracy of violence classification. Due to the lack of a publicly available violence dataset with spatial annotations, we propose to localize violence spatially using the knowledge of the classifier. The localization error of the masking model and the refinement step is affected by the classifier performance.

Method Limitations: The proposed approach also has some limitations, which must be improved in future works. First, our two-stage approach for violence detection is limited by the classifier performance, *i.e.*, poor performance of the classifier will generate poor saliency regions during the localization step. Secondly, during the refinement step, we use a person detector under the assumption that human agents are performing violent actions. Therefore, the localization fails if the person detector does not detect a person in the frame. Finally, more experiments are needed to achieve a more robust method of violence detection in real-world videos, such as long videos and videos with two or more instances of violence in a frame. Future works include experiments with more recent convolutional network architectures, improve the refinement method, and propose an end-to-end architecture for violence detection and localization.

ACKNOWLEDGMENT

This work was supported by grant 234-2015-FONDECYT (Master Program) from Cienciactiva of the National Council for Science,Technology and Technological Innovation (CONCYTEC-PERU).

REFERENCES

[1] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Computer Analysis of Images and Patterns*, 2011, pp. 332–339.

[2] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012, pp. 1–6.

[3] F. Landi, C. G. Snoek, and R. Cucchiara, "Anomaly locality in video surveillance," *arXiv preprint arXiv:1901.10364*, 2019.

[4] C. Demarty, C. Penet, M.Schedl, I. Bogdan, V. Quang, and Y. Jiang, "The mediaeval 2013 affect task: Violent scenesdetection," *MediaEval 2013 Working Notes*, pp. 383–395, 2013.

[5] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "Temporal robust features for violence detection," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2017, pp. 391–399.

[6] I. P. Febin, K. Jayasree, and P. T. Joy, "Violence detection in videos for an intelligent surveillance system using mobsift and movement filtering algorithm," *Pattern Analysis and Applications*, May 2019.

[7] A. B. Mabrouk and E. Zagrouba, "Spatio-temporal feature using optical flow based distribution for violence detection," *Pattern Recognition Letters*, vol. 92, pp. 62 – 67, 2017.

[8] F. D. M. d. Souza, G. Cámara-Chávez, E. A. d. Valle Jr., and A. d. A. Araujo, "Violence detection in video using spatio-temporal features," in *2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images*, 2010, pp. 224–230.

[9] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violence detection in surveillance video using low-level features," *PLOS ONE*, vol. 13, pp. 1–15, 10 2018.

[10] T. Deb, A. Arman, and A. Firoze, "Machine cognition of violence in videos using novel outlier-resistant vlad," *17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 989–994, 2018.

[11] J. Yu, W. Song, G. Zhou, and J.-j. Hou, "Violent scene detection algorithm based on kernel extreme learning machine and three-dimensional histograms of gradient orientation," *Multimedia Tools and Applications*, vol. 78, no. 7, pp. 8497–8512, Apr 2019.

[12] F. D. Souza and H. Pedrini, "Detection of violent events in video sequences based on census transform histogram," in *30th Conference on Graphics, Patterns and Images (SIBGRAPI)*, Oct 2017, pp. 323–329.

[13] H. Cai, H. Jiang, X. Huang, J. Yang, and X. He, "Violence detection based on spatio-temporal feature and fisher vector," in *Pattern Recognition and Computer Vision (PRCV)*, 2018.

[14] Q. Xia, P. Zhang, J. Wang, M. Tian, and C. Fei, "Real time violence detection based on deep spatio-temporal features," in *Biometric Recognition (CCBR)*, 2018, pp. 157–165.

[15] Z. Meng and Z. Yuan, Jiabinand Li, "Trajectory-pooled deep convolutional networks for violence detection in videos," in *Computer Vision Systems*, M. Liu, H. Chen, and M. Vincze, Eds. Cham: Springer International Publishing, 2017, pp. 437–447.

[16] B. Malveira, S. Avila, Z. Dias, and A. Rocha, "Breaking down violence: A deep-learning strategy to model and classify violence in videos," in *13th International Conference on Availability, Reliability and Security (ARES)*, 08 2018, pp. 1–7.

[17] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, "A novel violent video detection scheme based on modified 3d convolutional neural networks," *IEEE Access*, vol. 7, pp. 39 172–39 179, 2019.

[18] C. Li, L. Zhu, D. Zhu, J. Chen, Z. Pan, X. Li, and B. Wang, "End-to-end multiplayer violence detection based on deep 3d cnn," in *VII International Conference on Network, Communication and Computing*, ser. ICNCC 2018. ACM, 2018, pp. 227–230.

[19] I. Serrano, O. Deniz, J. L. Espinosa-Aranda, and G. Bueno, "Fight recognition in video using hough forests and 2d convolutional neural network," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 4787–4797, Oct 2018.

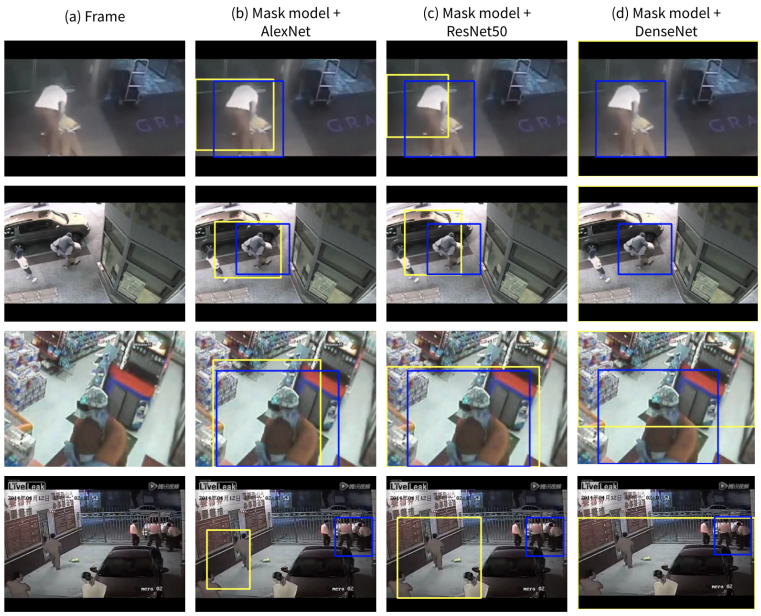


Fig. 10. Violence localization outputs generated by different classifiers. Column (a) shows a frame sample from the video, columns (b), (c) and (d) show the localization results generated by our localization method (masking + refinement) using different violence classifiers. The blue bounding boxes are the ground truth and the yellow bounding boxes are the localization results.

[20] E. Ditsanthia, L. Pipanmaekaporn, and S. Kamonsantiroj, "Video representation learning for cctv-based violence detection," *3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON)*, pp. 1–5, 2018.

[21] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," 09 2017.

[22] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang, and X. He, "A new method for violence detection in surveillance scenes," *Multimedia Tools and Applications*, vol. 75, no. 12, pp. 7327–7349, 2016.

[23] P. C. Ribeiro, R. Audigier, and Q. C. Pham, "Rimoc, a feature to discriminate unstructured motions: Application to violence detection for video-surveillance," *Computer Vision and Image Understanding*, vol. 144, pp. 121 – 143, 2016, individual and Group Activities in Video Event Analysis.

[24] C. Li, L. Zhu, D. Zhu, J. Chen, Z. Pan, X. Li, and B. Wang, "End-to-end multiplayer violence detection based on deep 3d cnn," in *VII International Conference on Network, Communication and Computing*, ser. ICNCC 2018. ACM, 2018, pp. 227–230.

[25] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, "Action recognition with dynamic image networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2799–2813, 2017.

[26] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 773–787, 2016.

[27] H. Kwon, S. Kwak, and M. Cho, "Video understanding via convolutional temporal pooling network and multimodal feature fusion," in *Proceedings of the 1st Workshop and Challenge on Comprehensive Video Understanding in the Wild*, 2018.

[28] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[32] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, 1979.

[33] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv.org*, pp. 1–6, Apr. 2018.

[34] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 2980–2988.

[35] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," *CoRR*, 2018.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12, 2012, pp. 1097–1105.

[37] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[38] T. Zhang, W. Jia, X. He, and J. Yang, "Discriminative dictionary learning with motion weber local descriptor for violence detection," *IEEE transactions on circuits and systems for video technology*, vol. 27, no. 3, pp. 696–709, 2016.

[39] I. Serrano, O. Deniz, G. Bueno, and T. Kim, "Fast fight detection," *PLOS ONE*, vol. 10, pp. 1–19, 04 2015.

[40] O. Deniz, I. Serrano, G. Bueno, and T. Kim, "Fast violence detection in video," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 2, Jan 2014, pp. 478–485.

[41] S. Mohammadi, H. K. Galoogahi, A. Perina, and V. Murino, "Physics-inspired models for detecting abnormal behaviors in crowded scenes," in *Group and Crowd Behavior for Computer Vision*. Elsevier, 2017, pp. 253–272.

[42] J. Mahmoodi and A. Salajeghe, "A classification method based on optical flow for violence detection," *Expert Systems with Applications*, vol. 127, pp. 121 – 127, 2019.

[43] P. Vashistha, C. Bhatnagar, and M. A. Khan, "An architecture to identify violence in video surveillance system using vif and lbp," in *2018 4th International Conference on Recent Advances in Information Technology (RAIT)*. IEEE, 2018, pp. 1–6.

[44] Q. Xia, P. Zhang, J. Wang, M. Tian, and C. Fei, "Real time violence detection based on deep spatio-temporal features," in *Biometric Recognition (CCBR)*, 2018, pp. 157–165.