Fight Detection in Video Sequences Based on Multi-Stream Convolutional Neural Networks

Sarah Almeida Carneiro*†, Gabriel Pellegrino da Silva*†, Silvio Jamil F. Guimarães‡ and Helio Pedrini*
*Institute of Computing, University of Campinas
Campinas, SP, 13083-852, Brazil
†Semantix Brasil
São Paulo, SP, 03178-200, Brazil
‡Computer Science Department, Pontifical Catholic University of Minas Gerais (PUC Minas)
Belo Horizonte, MG, 30535-065, Brazil

***Abstract*—Surveillance has been gradually correlating itself to forensic computer technologies. The use of machine learning techniques made possible the better interpretation of human actions, as well as faster identification of anomalous event outbursts. There are many studies regarding this field of expertise. The best results reported in the literature are from works related to deep learning approaches. Therefore, this study aimed to use a deep learning model based on a multi-stream and high level hand-crafted descriptors to be able to address the issue of fight detection in videos. In this work, we focused on the use of a multi-stream of VGG-16 networks and the investigation of conceivable feature descriptors of a video's spatial, temporal, rhythmic and depth information. We validated our method in two commonly used datasets, aimed at fight detection, throughout the literature. Experimentation has demonstrated that the association of correlated information with a multi-stream strategy increased the classification of our deep learning approach, hence, the use of complementary features can yield interesting outputs that are superior than other previous studies.**

I. Introduction

The frequent violent scenarios lived in both public and private locations, nowadays, has influenced the development of more assertive security applications. It has become common not only for establishments, but also for public places where there are various people transiting and interacting (for instance, schools, subways, stores, airports and banks) to install surveillance cameras.

Investments are substantial to maintain a security team or even hire third-party services to guard a particular space. In addition, a security system team is required to take prompt and hasty actions in some situations to prevent a harmful event from being triggered. Therefore, for these operations to run smoothly, part of the job demands a thorough surveillance video checking. Even though there are trained professionals to perform this task, this is tiring and the large number of cameras can make it almost impossible to monitor the videos uninterruptedly [1]. Furthermore, some actions can be deceiving and complex to anticipate only by overseeing the surveillance recordings.

As a result of the mentioned struggle, human action analysis gained a significant interest in the field of machine learning as well as computer vision. Accordingly, a considerable number of methods for identifying these behaviors have been proposed in the literature. These approaches can be divided into (i) gesture; (ii) facial expression; (iii) pose recognition. Some of these studies only indicate if the video content contains a particular event, while others disclose what events transpire throughout the scene.

According to recent works, action recognition can be divided into two major subdivisions [2], [3]: (i) the single layered approaches; and (ii) hierarchical approaches. The main differences between these techniques are their complexity and the ability of each to recognize from much simpler actions to complex activities.

As its main concern, this work adopted a four-stream VGG-16 architecture and explored high-level handcrafted features as inputs to investigate their impact on fight detection in videos. As other studies mainly use similar features for this binary problem classification, we focus on finding distinct feature descriptors that can also be good investments for fight detection. Therefore, we examine the influence of using: (i) the optical flow; (ii) a depth estimation; (iii) the visual rhythm; and (iv) the RGB features for classification.

Although some of these features have already been used in works available in the literature, their combination for exploring spatial, temporal and spatio-temporal information from the video frames through RGB, depth, optical flow and visual rhythms is, in fact, novel and one of the main contributions of our work. In this manner, it is possible to use the temporal, depth, rhythmic and spatial information of a video in a complementary fashion.

We are able to observe that potentially weaker isolated features, are able to, when combined, provide as good results as other previous demanding approaches proposed by the literature for a fight classification problem. In addition, our study demonstrates that the combined use of features that conceal the RGB information such as the optical flow, depth and visual rhythm can generate better results then combinations including the spatial (RGB) representation. Experiments were conducted on two datasets, Hockey Fight [4] and Movie Fight [4].

This paper is organized as follows. In Section II, we discuss some of the recent works associated to fight detection in videos. In Section III, important concepts used in this work are clarified. In Section IV, the proposed multi-stream methodol-

ogy is explained. In Section V, we describe the experiments performed and compare the achieved results to other published methods. Finally, some concluding notes and suggestions for future work are presented in Section VI.

II. Related Work

In this section, we discuss some of the works that also specifically addressed fight detection.

Concerned with anomaly identification in videos, Sultani *et al.* [5] decided to tackle a multiple instance learning (MIL) approach to cope with the problem. Another study was made by Li *et al.* [6], which proposed a depth image information based framework to recognize human interaction focusing on key frame extractions. The problem of finding the most representative frames was treated as a dictionary selection problem using sparsity consistency. Therefore, these frames had the proposed spatio-temporal image motion feature and a local edge feature extracted (3D Gabor filters and optical flow) and sent to a Support Vector Machine (SVM) to be recognized. Other studies, such as Keçeli and Kaya [7], also investigated the SVM behavior using high-level features, such as optical flow and transfer learning for violence detection on both crowded and uncrowded environments. In addition, Lejmi *et al.* [8] addressed the violence scenario by feature extracting points of interest from the inputs on the SBU Kinect Interaction dataset. They used an SVM in an ensemble combination with other learning algorithms, such as K-means and random forests.

To detect anomalous violent actions in crowd scenes, Hassner *et al.* [9] captured the optical flow from the videos and the changes between frames. Since abnormality detection is not confined to a few actions, Antić and Ommer [10] decided to parse video frames and use a discriminative background classification method. Stephens and Bors [11] focused their studies on group activity recognition. In their research, the authors investigated the video motion flow association. Naikal *et al.* [12], concentrating their work on simultaneous detecting and recognizing human actions from both single camera or multiple cameras, extracted the histogram of oriented gradients (HOG) descriptor from the foreground region of each frame, along with the coordinates of the bounding box and used them as inputs for their deformable keyframe model framework (DKM). In order to detect anomalous events, Du *et al.* [13] experimented with structural multi-scale motion interrelated patterns (SMMIP) and a Gaussian mixture model (GMM). In their study, Wang *et al.* [14] applied wavelet transformations on traditional spatio-temporal features to acquire high-frequency information. Multiple Hidden Markov Models, allied to a mechanism to judge the inputted behavior type, are then used to detect video abnormality.

Given the observation that the majority of studies involving action recognition were related to simple detections, such as hand gesture recognition, Bermejo *et al.* [4] targeted their work on video fight detection. The main contribution of their work verified that the use of Bag-of-Words (BoW) associated to Space-Time Interest Points (STIP) and Motion SIFT (MoSIFT) could provide approximately 90% accuracy when dealing with fighting in videos. In addition, the author introduced two datasets aimed at fights: (i) Hockey Fights and (ii) Movie Fights.

Deniz *et al.* [15] concerned with time efficiency compared to previous work, that relied on costly feature extractions, decided to study fight detection using kinematic features. In their work, the use of extreme acceleration patterns calculated based on motion blur allied to an SVM classifier proved that less features could generate significant results for three datasets. Their experiments were conducted on the datasets proposed by Bermejo *et al.* [4] and the UCF101 dataset demonstrating to be 15 times faster than their other compared methods. Also concerned with practical implementations of fight detection, Gracia *et al.* [16] based their work on classification of motion blobs extracted from video frames. Although the method, depending on the dataset, could not outperform the compared approaches, the authors were able to maintain a 70% to 90% accuracy average and still be time efficient. The study was performed also using the datasets proposed by Bermejo *et al.* [4] and the UCF101 dataset associated to SVM, AdaBoost and Random Forest classifiers.

Convinced of the need of improvements in surveillance applications, Gao *et al.* [17] employed the Violent Flows (ViF) as a descriptor for fight detection. Since the ViF did not consider some information involving both motion magnitudes and motion orientations, the authors also proposed the Oriented Violent Flows (OViF) descriptor. By using the SVM and AdaBoost algorithms, the Violent Flow dataset and the Hockey Fight dataset [4], Gao *et al.* [17] was capable of obtaining an accuracy average of 94%. In their study, it was concluded that the proposed feature was more appropriated for violence detection in non-crowded scenarios and that the combination of learning algorithms improved classification.

Interested in detecting violent content in videos, Mukherjee *et al.* [18] compared two methods for fight detection in sports. The first used blur and radon transform with a feed-forward Neural Network. For the second, the performance was fine-tuned using pre-trained VGG16. The authors reported that after 550 epochs, using the Hockey Fight dataset [4], the performance did not change and the accuracy continued 75%, when dealing with the pre-trained version. Considering only the feed-forward Neural Network, after 200 epochs the accuracy remained 56%. Fu *et al.* [19] inspired by an ensemble learner and the fact that there were not sufficient data in human fighting datasets, proposed a cross-species learning method. In their experiments, the authors used local motion features (LMF), including the motion statistics and segment correlation to readjust animal fighting data to assemble a human fight detection model. Results were based on four datasets: (i) Hockey Fights [4]; (ii) Movie Fights [4]; (iii) Animal Fights [19] and (iv) Human Fights [20]. Results achieved 85% to 99% of accuracy depending on the dataset.

Serrano *et al.* [21] proposed a hybrid "handcrafted/learned" feature. Their method was based on summarizing the content of a video sequence into an image and afterwards identifying

representative motion areas of fighting scenarios. In their work, a designed 2-D Convolutional Neural Network was used to classify the summarized resulting images between violent and normal cases. Results, compared to other works that used handcrafted features, such as LMF and ViF, showed an above 90% of accuracy when applied to the Hockey Fight [4] and Movie Fight [4] datasets. A spatio-temporal elastic cuboid (STEC) trajectory descriptor was proposed and used as input to a Hough forest classifier by Serrano *et al.* [22]. This made possible an average result of 90% for both of the Bermejo *et al.* [4] study. Xia *et al.* [23] invested on a bi-channel with VGG networks and two SVMs to achieve violence detection by using a label fusion method. In their study, a pre-trained VGG-f model on ImageNet dataset was used for feature extraction of the original video frame and the difference of adjacent frames. For each of these channels, an SVM was used for appearance and motion classifier, respectively. Their approach yielded a 96% accuracy for the Hockey Fight [4] dataset.

To detect violent actions in videos, after detecting people in frames using a trained MobileNet CNN model, Ullah *et al.* [24] used a sequence of frames as input to a 3D CNN model for spatial and temporal features extraction. Accordingly, the extracted features were passed to a Softmax classifier so their predictions could be obtained. Their proposed method was able of achieving an above 95% accuracy with both of Bermejo *et al.* [4] datasets. Febin *et al.* [25] used a movement filtering algorithm to check the existence of violence in videos. Only the frames that were assumed to have significant movement had their scale-invariant feature transform (SIFT), histogram of optical flow feature and motion boundary histogram extracted. The combined features formed the MoBSIFT descriptor used as inputs to an SVM, AdaBoost and Random Forest learning algorithms. Performance showed that classification ranged between 85% and 98% on Bermejo *et al.* [4] datasets depending on the classifying algorithm.

III. Theoretical Background

In this section, we explain some of the relevant concepts and techniques related to fight detection in video sequences.

A. High-Level Descriptors

In this work, we considered hand-crafted features as high-level descriptors. Therefore, the spatial, temporal, rhythmic and depth information extracted from video are clarified.

1) Spatial Descriptor: Since we intended to evaluate the impact of the raw information of a video frame, a feature chosen to be used as one of the multi-stream inputs was the unprocessed Red Green Blue (RGB) frames (Figure 1). This feature provides spatial relevant information regarding fight detection. Hence, it is a feature strongly associated to the RGB information of the videos such as location, clothing and actors.

2) Temporal Descriptor: The optical flow is a feature that captures an image object movement in a video. The extractors can generate the information relying on the movement among neighboring pixels or the modifications of pixel intensities between frames (Figure 2). Therefore, being able to describe

Fig. 1. Unprocessed RGB frame examples from the Hockey Fight dataset [4].

motion, it can support the networks recognition concerning classification and detection with temporal information.

Let I be a video frame and $I(x, y, t)$ be a pixel in an initial frame. In addition, compared to the next frame obtained dt time after, the pixel then moves a distance (dx, dy). Hence considering the mentioned pixels being equivalent and having static intensities, it is possible to consider Equation 1.

Subsequently, after applying a Taylor series approximation of right-hand side and divide by dt, the optical flow equation (Equation 2) is achievable, in which f_t is the gradient given time, f_x, f_y, u and v are given in Equation 3. Finally, to obtain the variable results of u and v, there are some methods that can be used, such as Lucas-Kanade [26] and Gunnar-Farnebäck [27].

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \tag{1}$$

$$f_x u + f_y v + f_t = 0 \tag{2}$$

$$f_x = \frac{\partial f}{\partial x}; \quad f_y = \frac{\partial f}{\partial y}; \quad u = \frac{\partial x}{\partial t}; \quad v = \frac{\partial y}{\partial t} \tag{3}$$

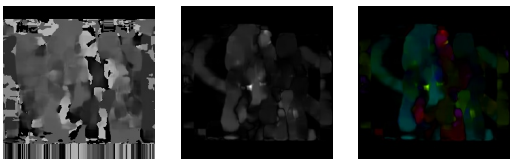

Fig. 2. Optical flow examples for (x, y, z) components.

3) Depth Descriptor: Much information can be obtained by calculating the video depth. Since specific fighting datasets have not already included the depth information (achievable with multiple camera shots) and it is a difficult task to compute this information based on a single 2D camera shot. For computing the depth descriptor, we used the depth estimator proposed by Godard *et al.* [28]. This estimator based on a deep learning approach is able to process a single 2D frame and estimate its depth values. Depth estimation frames can be seen in Figure 3.

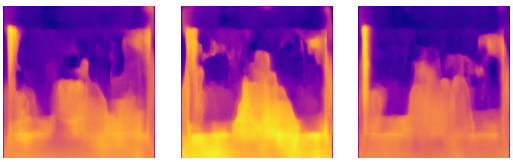

Fig. 3. Depth feature estimation examples.

4) Rhythmic Descriptor: The visual rhythm is an image of a full-length video and can describe both spatial and temporal information [29]–[31]. There are distinct forms to build a visual rhythm from the video, some of them are known as visual rhythm by histogram or by sub-sampling [32], [33].

In order to understand this concept, it must be considered that $\mathbb{D} \subset \mathbb{Z}^2$, in which $\mathbb{D} = \{0, ..., H-1\} \times \{0, ..., W-1\}$, H and W are the height and the width of each video frame. Therefore, a video V, in domain $2\mathbb{D} + t$, is a sequence of frames F_t and can be described in Equation 4, where T is the number of frames contained in the video.

$$V = (F_t)_{t \in [0, T-1]} \tag{4}$$

A visual rhythm generated by the histogram B can be, considering $(H_{f_t})_{t \in [0, T-1]}$ the sequence of histograms, computed from all frames of V, described as a 2D representation of all frame histograms, where each vertical line represents a frame histogram, therefore, B is defined in Equation 5, where $z \in [0, L-1]$ and $t \in [0, T-1]$, such that T is the number of frames and L the number of histogram bins, whereas the sub-sampling technique consists of encoding videos into images by adding slices from every frame to it. Thus, the visual rhythm, in domain 1D + t, is a rendition of the video in which each frame f_t is transformed into a vertical line of the visual rhythm image A, defined in Equation 6, where $z \in \{0, ..., H_A - 1\}$ and $t \in \{0, ..., T-1\}$, H_A, T, r_x, r_y, a and b are the height and the width of the visual rhythm, the ratios of pixel sampling and shifts on each frame, respectively.

$$B(t, z) = H_{f_t}(z) \tag{5}$$

$$A(t, z) = f_t(r_x \times z + a, r_y \times z + b) \tag{6}$$

Informally, a slice is a one-dimensional column image of a set of linearly organized pixels that can be constructed based on the iteration over every pixel of the image in a diagonal path. All slices are horizontally concatenated to form an image with dimensions $W \times H$ pixels. In this manner, each column of a visual rhythm image represents an instant in time, while each row represents a pixel of the image, or some other visual structure, varying in time. Examples of visual rhythm outputs can be observed in Figure 4.

Fig. 4. Visual rhythm feature examples.

B. Multi-Stream Learner

A multi-stream is a learner based on an ensemble of multiple learning algorithm outputs. A stream is characterized as an individual learning process. Therefore, each stream is responsible for an individual classification for a given input. After obtaining these individual classifications an ensemble is used to assemble the information and to define the final classification result for a video.

The advantages of adopting a multi-stream is the fact that distinct yet complementary information can be learned individually. Hence, the chances of having a high-level feature information imposing over others reduces.

1) Transfer Learning: Transfer learning is a technique employed when it is possible to use previously trained weights based on other similar data sets to instantiate earlier layers in a learning architecture [34]. Since transfer learning has shown promising results, it is mostly used to reduce training time and when the original problem data size for training samples is insufficient to correctly tune the model weights.

The method consists in using a larger dataset containing similar instances of the problem to train the learning model and instantiate the first layers of this model. Eventually, to be able to deal with the specific study subject, the original dataset will be, then, used to train and generate the weights for the last convolutional layers of the model. The main idea of transfer learning is to allow that the information, learned from the initially trained dataset, can be useful to further adding to the learning process of the topic under investigation. Therefore, transfer learning is particularly common in the image recognition field problems, since there already are many previously trained weights publicly available, such as ImageNet's [35], making the specific learning process much faster and more robust to different input data.

2) Ensemble: To improve accuracy, new architectures have always been developed, thus, a useful approach is ensemble. Therefore, by arranging an ensemble, a number of different learning approaches are joined [36]. These architectures can be either equal or distinct; however, the concept of ensemble relies on training each of these networks according to a specific input. In this sense, multiple learners will specialize in a different input, hence, the input will be used in all of the ensemble networks. Finally, after each learner computes their individual results, the one that has the majority of votes will have its results associated to the mentioned input.

IV. Multi-Stream Fight Detection Approach

The methodology proposed in this project aims to implement and evaluate an architecture based on a multi-stream deep neural network to verify the existence of violent fighting actions in videos. A four-stream model can provide the information of which features are relevant to be considered during a binary fight detection problem. In addition, it is tested if the increasing number of used streams is proportional to the escalation in evaluation metric values. In this section, we describe each part of the model illustrated in Figure 5, as well as the techniques we will use in each stage.

A. High-Level Feature Extraction

Four descriptors were investigated in this work: (i) RGB frames; (ii) Optical Flow; (iii) Depth Estimation; (iv) Visual Rhythm. Initially, the video frames were extracted and went through a feature descriptor generator algorithm (discussed in

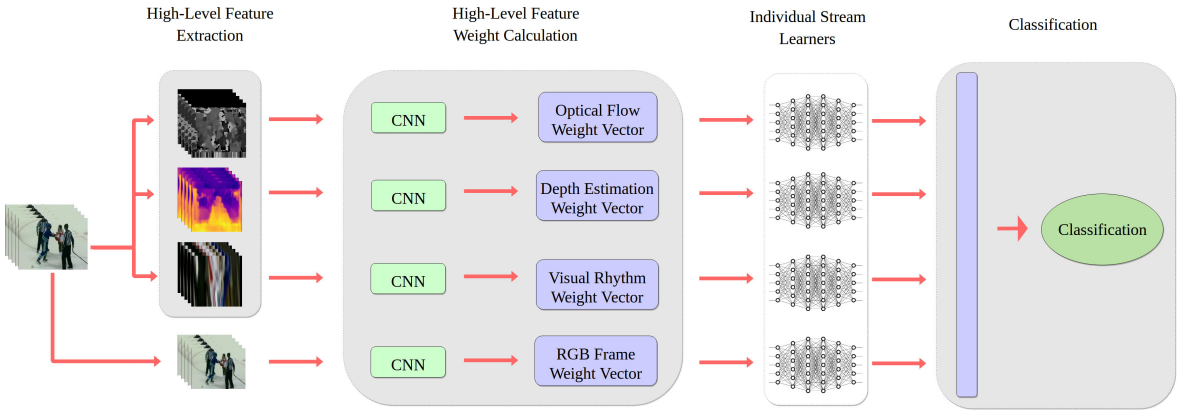

Fig. 5. Multi-Stream Fight Detection Approach.

Section III). Accordingly, the output of each generator was the associated high-level feature. The only feature that did not need to go through a generating algorithm was the RGB frames considering it was the raw frame itself.

B. High-Level Feature Weight Calculation

As it can be seen in Figure 5, each of the generated group of features went through a modified VGG so that a weight vector could be calculated and used in a posterior step of fine-tuning the streams. This weight vector generated by the CNN reduces the need of explicit feature engineering and it is able to make the method more independent.

C. Individual Stream Learner

Since we extracted four distinct features, it was needed a four-stream model for the proposed multi-stream architecture. As discussed in Section III, a multi-stream model has two general learning processes. The first process is the individual stream classifications and the second the final classification based on the previous step. Therefore, for each of the streams presented in Figure 5, a VGG-16 was used as the first individual classification. The VGG-16 model was chosen because of its simplicity as it is a classic convolutional approach and less deep than other networks, such as ResNet and Inception. In addition, it has yielded interesting results as it was found in previous works, such as Xia *et al.* [23]. A VGG-16 model has 16 layers and employs an architecture with small convolution filters, that can perform and output relevant results when coping with images [37].

In this work, the fighting data provided by the datasets did not have the ideal amount of information necessary to train an entire learning model. Therefore, a technique that was used to handle this dilemma was to pre-train each VGG-16 with the ImageNet dataset [35], and later with the UCF101 dataset [38]. This process ensures that the first layers are able to identify basic shapes and objects as well as the further layers are capable of distinguishing motion. This implies that the 14 first layers of the VGG-16 were trained with both ImageNet and UCF101. Finally, after assuring that the network had learned important basic information, the two last dense layers of the VGG-16 received one of the calculated high-level feature weights, based on the generated features designated as the input. This process allows the network to receive knowledge considered important to distinguish fighting and not fighting cases.

D. Classification

For the final classification, an ensemble was considered to merge the results of each individual stream. For this ensemble process, we propose to use three distinct approaches: (i) average and threshold; (ii) average and a support vector machine (SVM); and (iii) continuous values and SVM.

The average and threshold technique added the outputs of each stream and computed the average to compare it to a network parameter classification threshold. The second approach was similar to the previously discussed, but instead of empirically defining the threshold, an SVM was used for this purpose. Finally, the continuous values and SVM was responsible for generating a vector with each stream output so that an SVM could find this vector separation region.

Even though the results for each ensemble method were relatively close, we will only present the metrics achieved by the continuous values and SVM approach since the results were slightly higher.

V. Experiments

In this section, we discuss the experimental process setup to test our multi-stream methodology.

A. Datasets

In this work, we evaluate our method on Hockey Fight Dataset [4] and Movie Fight Dataset [4]. The Hockey Fight Dataset [4] is a gathering of 1000 video segments of action collected from the National Hockey League (NHL) hockey games. These segments have a dimension of 720×576 pixels and are composed of 50 frames each divided between fight and non-fight. In addition to this set, another dataset is considered in the experiments, the Movie Fight Dataset [4]. This set

contains 200 video snippets, separated in 100 fighting scenes and 100 normal events, collected from action movies.

B. Evaluation Metrics

The evaluations metrics used were accuracy, specificity and sensitivity. These were chosen since previous studies validated their works using them. The accuracy is a metric for the model performance evaluation that correlates both positive and negative classes and measures how accurate are the learner results. Specificity is a metric that provide information related to, given a negative example, the probability of a result being negative. Sensitivity, also known as recall, is, given a positive example, the classification result being indeed positive.

C. Quantitative Analysis

Throughout this study, we tested our method with a 10-fold cross-validation. The best results were achieved considering a 10^{-3} learning rate, a batch size of 1024, threshold of 0.5 and 500 epochs. Since most of the works presented in the literature, that tested with the same datasets as the current study, were not standardized while testing, we compare the best accuracy rates that the entire method could yield. The datasets were divided by 80% for training and 20% for testing. Results for our best configuration can be seen in Tables I and II.

It can be observed in Tables I and II that we showcase all of the possible combinations regarding our presented high-level features. Hence, we can observe all of the single feature until the four-streams combinations. Single streams do not have a prior ensemble step for their final classification since they already output the final result.

It is possible to observe that the use of individual features associated with our pre-trained VGG-16 as a single stream can already yield interesting results. However, when dealing with fight detection, it is important to have the best achievable metrics. As we suggested, the combination of non-straightforward features that are complementary can yield comparable results to the literature and even have higher accuracy metrics than some of the presented works. This is an important finding regarding studies of ideal features that one can quickly process and use in fight detection scenarios. In addition, it is possible to notice that the smallest accuracy rates for the single streams are related to the depth estimation descriptor. It is our understanding, since the dataset did not provide the original depth information of each video, that an estimation would not be as good as the ground truth. Moreover, as it can be seen in Table I, the combinations that had RGB did not surpass the three multi-stream combinations of the visual rhythm, optical flow and depth for the most challenging dataset. As expected, even though the RGB rates had a high performance, we believe that it is slightly overfitted. The RGB (spatial) information is a dependable feature, in other words, it is attached to all the objects that exist in a scene such as: colors, actors, and objects. Non-spatial information can further detach itself form a specific dataset and have, consequently, better results.

Multi-stream combinations results on the Hockey Fight dataset (Table I) demonstrate that our fight detection approach is comparable to some methods available in the literature ([16]–[19], [22], [25]), although it does not surpass the state of the art. The results on the Movie Fight dataset (Table II), as well as some of the methods cited in the paper, achieved high rates of sensitivity, specificity and accuracy. We conjecture that the dataset is not currently very challenging for a classification problem, although it is widely used to validate fight detection approaches.

All of our combinations had an above 80% in metrics indicating their relevance considering a classification problem. Our hypothesis lies on the fact that by learning complementary high-level features individually, it can help the networks to better balance the outlier classifications. Moreover, not necessarily, increasing the number of streams means to increase the accuracy values. In some cases, it can be observed that the numbers might decrease depending on the feature combination. However, by studying these features it is possible to observe which are conceivable descriptors to be used in this fight detection problem that might be less expensive, more reliable and which are the best feature combinations to cope with the presented situation.

Another remark that we were able to detect was that even though the accuracies were not higher then all of the presented works, our study was only carried out with 500 epochs compared to studies that ran on 1000 to 5000 epochs and demanded lots of processing. Studies such as Ullah *et al.* [24] based their progress on much more costly learning models compared to ours. Finally, it is possible to notice by our achieved sensitivity and specificity scores that the method has some trouble when dealing with negative fighting cases. Therefore, the learning model is able to better detect positive situations and, in sensitive scenarios such as health and security, it is best to detect a false positive than a false negative case.

D. Qualitative Analysis

According to the classification achieved with our method, we tried to define a video pattern of which the model had trouble to identify as a positive or negative scenario. The Hockey Fight dataset misleading videos did not have a specific situation, but we observed that the data that involved smaller commotions, such as grapple and non-magnified punch movements, were included in this group. Some of the incorrectly classified videos also had more than two actors and others might have been influenced by camera shifting. On the other hand, although the Movie Fight dataset had more variability than the Hockey Fight dataset, it was a less challenging dataset for the learning model. Therefore, the Movie Fight dataset had a smaller amount of inaccurate classification. However, since it is composed of staged fights, the videos classified as inaccurate were mostly the ones with significant camera shifts and not convincing fighting scenes.

TABLE I
HOCKEY FIGHT RESULTS.

	Sensitivity (%)	Specificity (%)	Accuracy (%)
Multi-stream (OF+RGB+VR+D)	91.36	85.87	88.62
Multi-stream (OF+RGB+VR)	91.30	**86.12**	88.71
Multi-stream (OF+RGB+D)	90.80	85.45	88.12
Multi-stream (OF+VR+D)	**92.64**	85.49	**89.10**
Multi-stream (RGB+VR+D)	90.84	85.62	88.23
Multi-stream (OF+RGB)	90.51	85.66	88.09
Multi-stream (OF+VR)	91.03	82.09	86.56
Multi-stream (OF+D)	91.61	79.65	85.53
Multi-stream (RGB+VR)	90.59	85.72	88.15
Multi-stream (RGB+D)	90.05	85.06	87.56
Multi-stream (VR+D)	88.68	84.00	86.34
Single-stream (OF)	86.30	77.37	81.84
Single-stream (RGB)	89.18	85.08	87.14
Single-stream (VR)	86.05	77.10	81.58
Single-stream (D)	85.12	85.14	81.49
Bermejo *et al.* [4]	-	-	90.90
Deniz *et al.* [15]	-	-	90.10
Gracia *et al.* [16]	-	-	72.50
Gao *et al.* [17]	-	-	86.30
Mukherjee *et al.* [18]	-	-	75.00
Fu *et al.* [19]	-	-	87.50
Serrano *et al.* [21]	93.80	95.4	94.60
Serrano *et al.* [22]	-	-	82.60
Xia *et al.* [23]	-	-	95.90
Ullah *et al.* [24]	96.67	95.43	96.00
Febin *et al.* [25]	-	-	86.50

TABLE II
MOVIE FIGHT RESULTS.

	Sensitivity (%)	Specificity (%)	Accuracy (%)
Multi-stream (OF+RGB+VR+D)	**100.0**	**100.0**	**100.0**
Multi-stream (OF+RGB+VR)	**100.0**	**100.0**	**100.0**
Multi-stream (OF+RGB+D)	**100.0**	**100.0**	**100.0**
Multi-stream (OF+VR+D)	100.0	98.48	99.32
Multi-stream (RGB+VR+D)	100.0	98.25	99.21
Multi-stream (OF+RGB)	**100.0**	**100.0**	**100.0**
Multi-stream (OF+VR)	**100.0**	**100.0**	**100.0**
Multi-stream (OF+D)	99.71	100.0	99.84
Multi-stream (RGB+VR)	100.0	99.76	99.68
Multi-stream (RGB+D)	**100.0**	**100.0**	**100.0**
Multi-stream (VR+D)	87.03	94.75	90.47
Single-stream (OF)	99.71	100.0	99.84
Single-stream (RGB)	**100.0**	**100.0**	**100.0**
Single-stream (VR)	83.92	94.52	88.65
Single-stream (D)	100.0	93.12	96.93
Bermejo *et al.* [4]	-	-	89.50
Deniz *et al.* [15]	-	-	82.50
Gracia *et al.* [16]	-	-	87.20
Fu *et al.* [19]	-	-	99.00
Serrano *et al.* [21]	98.00	100.0	99.00
Serrano *et al.* [22]	-	-	98.00
Ullah *et al.* [24]	100.0	100.0	99.90
Febin *et al.* [25]	-	-	76.60

VI. CONCLUSIONS

In this work, we considered the use of a deep learning approach based on multi-stream learners. Accordingly, this work focused on the an ensemble of individual VGG-16 streams to cope with the binary problem of fight detection in videos. Our study was performed upon a 10-fold cross-validation to test our method and it was possible to observe that utilizing a pre-trained VGG-16 and fine-tuning its last dense layers benefited even the single-stream approaches. In addition, the multi-streams also showed interesting results regarding classification. Our hypothesis lies on the fact that learning complementary high-level features individually can help the networks to better balance the outlier classifications. By studying the effects of each high-level feature on the classifier, we are more likely to understand relevant information for the network as well as observe which are the best descriptors to be used on a detection scenario. As a limitation of this work, we did not test if the VGG-16 was necessarily the best option to

cope with the presented problem. It has many parameters and can have a significant training time, might not being the best in terms of performance and effectiveness. As directions for future work, we intend to impose some constraints on the neural networks through regularization mechanisms and apply our method to more challenging datasets to observe its performance. In addition, it is intended to test the presented stream model using other learning networks.

ACKNOWLEDGMENTS

We are thankful to São Paulo Research Foundation (FAPESP grants #2017/12646-3 and #2014/12236-1), FAPEMIG (PPM-00006-16) and National Council for Scientific and Technological Development (CNPq #309330/2018-1, Universal #421521/2016-3 and PQ #307062/2016-3) for their financial support. We are also grateful to Semantix Brasil for the infrastructure and support provided during the development of the present work.

REFERENCES

[1] X. Sun, H. Yao, R. Ji, X. Liu, and P. Xu, "Unsupervised Fast Anomaly Detection in Crowds," in *19th ACM International Conference on Multimedia*. ACM, 2011, pp. 1469–1472.

[2] J. K. Dhillon and A. K. S. Kushwaha, "A Recent Survey for Human Activity Recognition based on Deep Learning Approach," in *Fourth International Conference on Image Information Processing*. IEEE, 2017, pp. 1–6.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM Computing Surveys*, vol. 41, no. 3, p. 15, 2009.

[4] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence Detection in Video using Computer Vision Techniques," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2011, pp. 332–339.

[5] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," *arXiv preprint arXiv:1801.04264*, 2018.

[6] J. Li, X. Mao, L. Chen, and L. Wang, "Human Interaction Recognition Fusing Multiple Features of Depth Sequences," *IET Computer Vision*, vol. 11, no. 7, pp. 560–566, 2017.

[7] A. Keçeli and A. Kaya, "Violent Activity Detection with Transfer Learning Method," *Electronics Letters*, vol. 53, no. 15, pp. 1047–1048, 2017.

[8] W. Lejmi, A. B. Khalifa, and M. A. Mahjoub, "Fusion Strategies for Recognition of Violence Actions," in *IEEE/ACS 14th International Conference on Computer Systems and Applications*. IEEE, 2017, pp. 178–183.

[9] 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*. IEEE, 2012, pp. 1–6.

[10] B. Antić and B. Ommer, "Video Parsing for Abnormality Detection," in *IEEE International Conference on Computer Vision*. IEEE, 2011, pp. 2415–2422.

[11] K. Stephens and A. G. Bors, "Group Activity Recognition on Outdoor Scenes," in *13th IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2016, pp. 59–65.

[12] N. Naikal, P. Lajevardi, and S. S. Sastry, "Joint Detection and Recognition of Human Actions in Wireless Surveillance Camera Networks," in *IEEE International Conference on Robotics and Automation*. IEEE, 2014, pp. 4747–4754.

[13] D. Du, H. Qi, Q. Huang, W. Zeng, and C. Zhang, "Abnormal Event Detection in Crowded Scenes based on Structural Multi-Scale Motion Interrelated Patterns," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2013, pp. 1–6.

[14] B. Wang, M. Ye, X. Li, F. Zhao, and J. Ding, "Abnormal Crowd Behavior Detection using High-Frequency and Spatio-Temporal Features," *Machine Vision and Applications*, vol. 23, no. 3, pp. 501–511, 2012.

[15] O. Deniz, I. Serrano, G. Bueno, and T.-K. Kim, "Fast Violence Detection in Video," in *International Conference on Computer Vision Theory and Applications*, vol. 2. IEEE, 2014, pp. 478–485.

[16] I. S. Gracia, O. D. Suarez, G. B. Garcia, and T.-K. Kim, "Fast Fight Detection," *PloS One*, vol. 10, no. 4, p. e0120448, 2015.

[17] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence Detection using Oriented Violent Flows," *Image and Vision Computing*, vol. 48, pp. 37–41, 2016.

[18] S. Mukherjee, R. Saini, P. Kumar, P. P. Roy, D. P. Dogra, and B.-G. Kim, "Fight Detection in Hockey Videos using Deep Network," *The Journal of Multimedia Information System*, vol. 4, no. 4, pp. 225–232, 2017.

[19] E. Y. Fu, M. X. Huang, H. V. Leong, and G. Ngai, "Cross-Species Learning: A Low-Cost Approach to Learning Human Fight from Animal Fight," in *ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 320–327.

[20] E. Y. Fu, H. V. Leong, G. Ngai, and S. C. Chan, "Automatic Fight Detection in Surveillance Videos," *International Journal of Pervasive Computing and Communications*, vol. 13, no. 2, pp. 130–156, 2017.

[21] 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, 2018.

[22] I. Serrano, O. Deniz, G. Bueno, G. Garcia-Hernando, and T.-K. Kim, "Spatio-Temporal Elastic Cuboid Trajectories for Efficient Fight Recognition using Hough Forests," *Machine Vision and Applications*, vol. 29, no. 2, pp. 207–217, 2018.

[23] Q. Xia, P. Zhang, J. Wang, M. Tian, and C. Fei, "Real Time Violence Detection Based on Deep Spatio-Temporal Features," in *Chinese Conference on Biometric Recognition*. Springer, 2018, pp. 157–165.

[24] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network," *Sensors*, vol. 19, no. 11, pp. 1–15, 2019.

[25] I. 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*, pp. 1–13, 2019.

[26] C. Jain and D. Gautam, "Abnormal Behaviour Detection at Traffic Junctions using Lucas Kanade and Harris Corner Detector," in *4th International Conference on Recent Advances in Information Technology*. IEEE, 2018, pp. 1–5.

[27] A. Lowhur and M. C. Chuah, "Dense Optical Flow based Emotion Recognition Classifier," in *IEEE 12th International Conference on Mobile Ad Hoc and Sensor Systems*. IEEE, 2015, pp. 573–578.

[28] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised Monocular Depth Estimation with Left-Right Consistency," in *Conference on Computer Vision and Pattern Recognition*, 2017.

[29] B. S. Torres and H. Pedrini, "Detection of Complex Video Events through Visual Rhythm," *The Visual Computer*, vol. 34, no. 2, pp. 145–165, 2018.

[30] T. Moreira, D. Menotti, and H. Pedrini, "First-Person Action Recognition Through Visual Rhythm Texture Description," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2017, pp. 1–4.

[31] F. B. Valio, H. Pedrini, and N. J. Leite, "Fast Rotation-Invariant Video Caption Detection based on Visual Rhythm," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2011, pp. 157–164.

[32] S. J. F. Guimarães, M. Couprie, A. A. Araújo, and N. J. Leite, "Video Segmentation based on 2D Image Analysis," *Pattern Recognition Letters*, vol. 24, no. 7, pp. 947–957, 2003.

[33] S. J. F. Guimaraes, A. A. Araújo, M. Couprie, and N. J. Leite, "Video Fade Detection by Discrete Line Identification," in *Object Recognition Supported by User Interaction for Service Robots*. IEEE, 2002, pp. 1013–1016.

[34] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[36] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *International Workshop on Multiple Classifier Systems*. Springer, 2000, pp. 1–15.

[37] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[38] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild," *arXiv preprint arXiv:1212.0402*, 2012.