Fight Recognition in Video Using Hough Forests and 2D Convolutional Neural Network

Ismael Serrano, Oscar Deniz, *Senior Member, IEEE*, Jose Luis Espinosa-Aranda, and Gloria Bueno, *Member, IEEE*

***Abstract*— While action recognition has become an important line of research in computer vision, the recognition of particular events, such as aggressive behaviors, or fights, has been relatively less studied. These tasks may be extremely useful in several video surveillance scenarios, such as psychiatric wards, prisons, or even in personal camera smartphones. Their potential usability has led to a surge of interest in developing fight or violence detectors. One of the key aspects in this case is efficiency, that is, these methods should be computationally fast. "Handcrafted" spatio-temporal features that account for both motion and appearance information can achieve high accuracy rates, albeit the computational cost of extracting some of those features is still prohibitive for practical applications. The deep learning paradigm has been recently applied for the first time to this task too, in the form of a 3D convolutional neural network that processes the whole video sequence as input. However, results in human perception of other's actions suggest that, in this specific task, motion features are crucial. This means that using the whole video as input may add both redundancy and noise in the learning process. In this paper, we propose a hybrid "handcrafted/learned" feature framework which provides better accuracy than the previous feature learning method, with similar computational efficiency. The proposed method is compared to three related benchmark data sets. The method outperforms the different state-of-the-art methods in two of the three considered benchmark data sets.**

***Index Terms*— Fight recognition, violence recognition, Hough forests, deep learning, 2D convolutional neuronal network.**

I. Introduction

In RECENT years, the task of human action recognition from video has been tackled with computer vision and machine learning techniques, see surveys [1]–[3]. Experimental results have been obtained for recognition of actions such as walking, jogging, pointing or hand waving [4]. However, action detection has been devoted comparatively less effort. Violence detection is a task that can be leveraged in real-life applications. While there is a large number of studied datasets for action recognition, specific datasets with a relevant number of violent sequences (fights) were not available until [5], where the authors created two specific datasets for the fight/violence problem testing state-of-the-art methods on them. The main task of large-scale surveillance systems used in institutions such as prisons, schools and psychiatric care facilities is generating alarms of potentially dangerous situations. Nevertheless, security guards are frequently burdened with a large number of cameras where manual response times are frequently large, resulting in a strong demand for automated alert systems. Also, this type of systems must be very efficient because there is generally a large number of surveillance cameras which must be processed. Similarly, there is increasing demand for automated rating and tagging systems that can process large amounts of videos uploaded to websites. Since smartphones are often used to record beatings, efficient mobile implementations are desired too.

This work is based on the assumption that fights in video can be reliably recognized by kinematic cues that represent violent motion. This idea is inspired by a body of research on human perception that has shown that the kinematic pattern of movement is sufficient for the perception of other's actions [6]. More specifically, empirical studies in the field have shown that relatively simple dynamic features such as velocity and acceleration correlate to emotional attributes perceived from the observed actions [7]–[10], albeit the degree of correlation varies for different emotions. Thus, features such as acceleration and jerkiness tend to be associated with emotions with high activation (e.g. anger, happiness), whereas slow and smooth movements are more likely to be judged as emotions with low activation (eg. sadness). This same essential idea has been also supported by research on the computer vision side [11], [12]. These authors demonstrate that kinematic patterns of movements and dynamic features are representative for the perception of high-energy actions. In other words, motion carries most of the information useful to discriminate fight/violence sequences. Moreover, motion information could be much more important than appearance in this task. Following these experiments we propose to leverage the high-motion areas in this type of sequences using spatial features combined with a spatio-temporal classifier to learn when and where the fight/violence actions could be occurring.

Still, in line with the classical approach to machine learning, "handcrafted" features have been mostly used in previous work related to this task. In this work, features are learned using a Convolutional Neural Network trained with images that summarize the content of video sequences. In this respect, the proposed method can be considered hybrid in the sense that

Manuscript received June 28, 2017; revised February 10, 2018 and March 26, 2018; accepted June 5, 2018. Date of publication June 8, 2018; date of current version June 27, 2018. This work was supported by the Spain's Ministry of Economy and Competitiveness under Project TIN2011-24367. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alin M. Achim. *(Corresponding author: Ismael Serrano.)*

The authors are with the VISILAB Group, E. T. S. I. Industriales, University of Castilla–La Mancha, 13071 Ciudad Real, Spain (e-mail: Ismael.Serrano@uclm.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2845742

Fig. 1. Some sample frames from the Hockey dataset (first row), Movies dataset (second row) and Behave dataset (third row). In each row, the two left columns are non-fight sequences, while the two right columns are fight sequences.

features are learned, albeit from "handcrafted" input images that encode motion (and appearance) throughout the video sequence. The proposed method improves upon accuracies reported with previous deep learning networks applied to this problem.

The main contribution of this paper is a novel method to recognize fight or aggressive situations in video sequences. More concretely, we propose a method that summarizes the content of a video sequence into a result image encoding appearance and leveraging motion areas that are very representative of fight or aggressive situations. We also design a simple and efficient 2-D Convolutional Neural Network that is able to discriminate the previous result images between fight/aggressive and normal situations. The proposed method is tested on three different benchmark datasets (some sample frames for each dataset are shown in Fig. 1) and compared with state-of-the-art related methods. The paper is organized as follows. Section II discusses previous related work. Section III describes the proposed method. Section IV provides experimental results. Finally, in Section V the main conclusions are outlined.

II. Related Work

One of the first proposals for violence recognition in video is Nam *et al.* [13]. They proposed a method that used flame and blood detection and capturing the motion degrees, as well as the characteristic sounds from violent events. Cheng *et al.* [14] recognized gunshots, explosions and car-braking in audios using a hierarchical approach based on Gaussian mixture and Hidden Markov models (HMM).

Giannakopoulos *et al.* [15] proposed a violence detector based on audio features. Zajdel *et al.* [16] proposed the CASSANDRA system that extracts motion features related from articulations in video and scream-like cues in audio to search aggressive actions in surveillance videos. Gong *et al.* [17] developed a violence detector that uses low-level visual, acoustic features and high-level audio sounds for identifying potential violent action in movies. Chen *et al.* [18] used binary local motion descriptors (spatio-temporal video cubes) and bag-of-words approach to detect aggressive behaviors. Furthermore, Giannakopoulos *et al.* [19] proposed a novel method for searching violence actions in movies clustering audio-visual features applying statistics, average motion and motion orientation variance features in video with a K-Nearest Neighbor classifier to decide whether there are violence actions. Chen *et al.* [20] proposed a method based on motion for detecting faces and nearby blood. Hassner *et al.* [21] recently tackled the problem of detecting violence outbreaks in crowds using an optical flow-based method. The authors of the Local Motion Patterns (LMP) method [22] claimed that it is both informative and efficient for action and violence recognition since the number of extracted descriptor vectors varies in each video sequence. This method is based on extracting simple statistics (variance, skewness and kurtosis) from temporal cuboids centered on tracked keypoints. Keypoints are located with a Harris detector. Proof of the growing interest in the topic was beside the MediaEval Affect Task, a competition whose goal is to search violence action in color movies [23].

In summary, a significant number of previous state-of-the-arts methods required audio cues for recognizing violence or trust on color areas to detect cues such as blood or skin. In that regard, we note that there is an important number of applications, especially in surveillance, where audio and color features are generally unavailable. In other cases, it is possible and easy to obtain audio features, but audio information may increase false positive rates or decrease true positive rates because there are many violence actions where the audio features may be confused: to push, throw (something), knock down, attack with a knife, block (someone), etc. Moreover, while explosions, blood and running may be very useful cues to detect violence scenarios in action movies, they are unusual in real-world actions. Anyway, violence recognition is a very difficult issue, since violence is a subjective concept. Fight recognition, on the contrary, is a more specific violence-related topic that may be tackled using similar techniques.

To the best of our knowledge, only few studies have focused specifically on fight recognition. A fight is defined in our context as the use of physical force to try to hurt someone. This can involve two or more people. Large accelerations are obviously an important cue, as demonstrated in [24], which introduced the so-called Acceleration Measure Vectors (AMV). In that system, however, such vectors were obtained from tracking of body parts, which is in itself a challenging problem. In [25] a surveillance system was described in which average global motion histograms are learned from video segments, generating alarm events in the case of unusual (in direction, magnitude and location) motion. The authors did not present recognition results, however. Andersson *et al.* [26] fused optical and acoustic information to detect fights. The magnitude of whole-image optical flow was used to discriminate the visual side of fights. Even with such simple method the authors measured a 81% fight recognition rate (including audio information). Bermejo *et al.* [5] demonstrated encouraging results in applying generic action recognition methods to the specific task of fight recognition, achieving 90% accuracy using MoSIFT (Motion Scale-Invariant Features) descriptor ([27]). That work also introduced the "Hockey dataset", which has since been used by researchers to assess fight detectors (besides used in this work). MoSIFT descriptors are obtained from salient points in two parts: the first is an aggregated histogram of gradients (HoG) which describes the spatial appearance. The second part was an aggregated histogram of optical flow (HoF) which indicates the movement of the feature point. More recently, Jian-Feng and Shui-Li [28] proposed variance of optical flow along with a SVM classifier to detect fights. In [29] Kernel Density Estimation was exploited for feature selection on the MoSIFT descriptor. This system achieved an accuracy of 94.3% on the Hockey dataset. The Motion Binary Patterns MBP ([30]) were also used to detect fights. Although, MBP was originally proposed for action recognition, for estimating motion intensity using local changes in pixel intensity. Kaelon [31] proposed a new descriptor "grey level co-ocurrence texture measures, edge cardinality and pixel intensity difference" (GEP) for fight recognition. The method was comprised of four measures, texture energy, texture contrast, edge cardinality and pixel difference between adjacent frames.

Another recent method was described in [32]. In that case the authors focuses on measuring the fuzzy regions that appear in the image when sudden strong motion (from the fight) occurs and use this as a cue to discriminate between fight and non-fight sequences. Gracia *et al.* [33] considered motion blobs (taken after thresholding the absolute difference of consecutive frames) to describe the different parts of an action. Basic features (area, perimeter, etc.) were extracted from the largest K blobs and used as the main features to discriminate between fights and non-fights. Even more recently, Senst *et al.* [34] proposed a modified set of SIFT features that encoded both appearance and Lagrangian-based motion models, in a bag-of-words framework and followed by a SVM classifier. Mohammadi *et al.* [35] proposed a novel video descriptor based on the substantial derivative, a concept in fluid mechanics. After estimating the convective and local field from the optical flow, a bag-of-word procedure is used for each motion pattern separately, concatenating the resulting histograms to form the final descriptor.

In general, all work described above used "handcrafted" features. While in some cases such features can provide robust results, they can also be unnecessarily inefficient. Take for example the powerful MoSIFT features mentioned above. The computational cost of extracting such features were prohibitively large, taking nearly 1 second per frame on a high-end laptop. This precludes use in practical applications, where many camera streams may have to be processed in real-time.

As far as the authors know, an important attempt to use the deep learning paradigm in this particular problem is [36], where a 3D Convolutional Neural Network (3D CNN) was used to detect fights. In that work, the network processes the whole video as input. The 3D nature of the network attempts to encode temporal information. Such architecture is extremely demanding both in storage and computational efficiency terms. Besides, in the light of our discussion above, the use of raw video sequences as input can really act as a distractor if the task at hand relies heavily on kinematic information.

Tran *et al.* [37] also proposed spatio-temporal feature learning using 3D Convolutional Neuronal Networks (3D-CNN) for action recognition, called C3D. The authors argued that 3D-CNN is able to model temporal information thanks to the 3D pooling and 3D convolution layers. All 3D convolution kernels were set to $3 \times 3 \times 3$ with a stride 1 in both spatial and temporal dimensions. The C3D method used an architecture with 15 layers. The method achieved 85.2% action recognition accuracy on the challenging UCF101 dataset. Zhang *et al.* [38] proposed to extend IWLD features by adding a temporal component, called MoIWLD, to detect violent behaviors in real video scenes. The authors also proposed a modified sparse model to learn a dictionary for classification.

In this context, this paper describes a method which uses "handcrafted" input images to feed a much simpler 2D Convolutional Neural Network. The input images are obtained as a weighted sum of the frames in the sequence, where these weights are time stamps (or frame numbers). A spatio-temporal voting model is also applied to weight every frame. This way, for each video sequence a single image is obtained which is then used as input to a 2D Convolutional Neural Network. This method can be considered a hybrid approach that, given the specific characteristics of the problem at hand, can be exploited to our advantage providing accurate results in an efficient manner.

III. Proposed Method

The proposed method aims at classifying fight and non-fight video sequences. Firstly, we assume that an image can be used to summarize the sequence. Frames from the sequence are accumulated in order to build a representative image for each sequence. In addition, we propose to leverage those zones inside the frames that may be most important in terms of describing the motion in the sequence. In this way, the important motion parts are weighted more while the remaining regions receive less importance, the latter generally corresponding to noise and mostly static background.

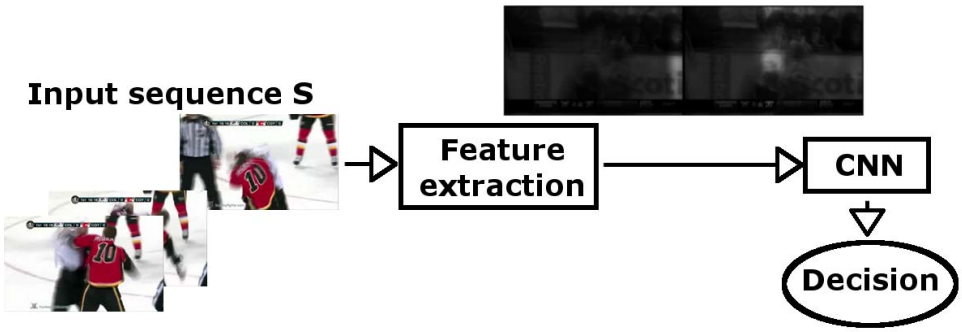

Fig. 2. General diagram of the proposed method.

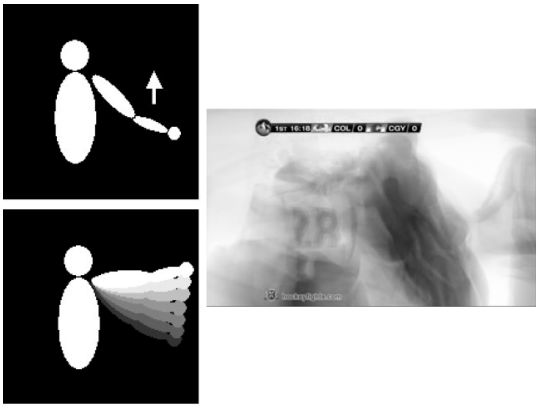

Fig. 3. Left top: example input sequence in which an individual moves his arm. Left bottom: a motion history image calculated from the input sequence. Right: a motion history image calculated from a Hockey fight sequence.

The proposed method is summarized in the following two steps, see Fig. 2:

1) The feature extraction step aims to obtain a representative image from each input video sequence.
2) A 2D Convolutional Neuronal Network is used to classify the representative image and obtain the final decision for the sequence.

The main step of feature extraction is performed as follows. Both motion and appearance in the video sequence are encoded in a representative image. The first step is to leverage the relevant motion parts and reduce the effect or irrelevant parts, such as background and noise. Then, a spatio-temporal voting model (Hough Forests model) is used to obtain a weighting image for each class and time. Afterward, to build the representative image, a technique similar to motion history images [39] is used and applied to the weighted frames. Frames are accumulated with a weight related to their temporal position (frame number in our case), see two examples in Fig. 3.

The resulting image is a concise representation of both appearance and motion throughout the video. Such representation can be also computed efficiently since it is readily parallelizable (all operations can be performed frame-by-frame). Given a frame sequence as:

$$S = \{I(1), I(2), \cdots, I(t), \cdots, I(T)\} \tag{1}$$

where $I(t)$ represents frame t and T is the number of frames for this sequence S with $W \times H$ frame sizes. The result representative image F is obtained by concatenating the following two images for each sequence as:

$$F = P' \| P \tag{2}$$

where

$$P = \frac{\sum_{t=1}^{T} t \cdot (M(t) \star I(t))}{n} \tag{3}$$

and

$$P' = \frac{\sum_{t=1}^{T} t \cdot (M'(t) \star I(t))}{n} \tag{4}$$

$$n = 255 \times \sum_{t=1}^{T} t \tag{5}$$

The $M(t)$ and $M'(t)$ images are obtained from the spatio-temporal voting Hough Forests model for the sequence S that correspond for each class (this will be explained below). Hough Forests models generally produce a sparse spatio-temporal voting map. In order to extend and soften these votes on the image, a 2D Gaussian filter is applied for each frame. This filter is widely used in computer vision, typically to reduce image noise and reduce detail. The Gaussian filter is commonly defined as:

$$G(x, y) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma}} \cdot e^{-\frac{x^2+y^2}{2 \cdot \sigma^2}} \tag{6}$$

where σ^2 is the variance of Gaussian filter, and the size of filter kernel ($l \times l$), and $-l \leqslant (x, y) \leqslant l$. The σ parameter is normalized with the frame height as $\sigma = \lceil H \cdot P_\sigma \rceil$, where $0 \leqslant P_\sigma \leqslant 1$. The size of the filter kernel l is defined as:

$$l = 4 \times \lceil \sigma \rceil + 1 \tag{7}$$

we have empirically fixed this value to 4 because it is necessary to produce an enough large filter in order to soften the effect of the spatio-temporal voting Hough Forests models. The $M(t)$ and $M'(t)$ images correspond with the space voting for each class and instant of time, specifically fight and non-fight classes, respectively. These images aim to weight the

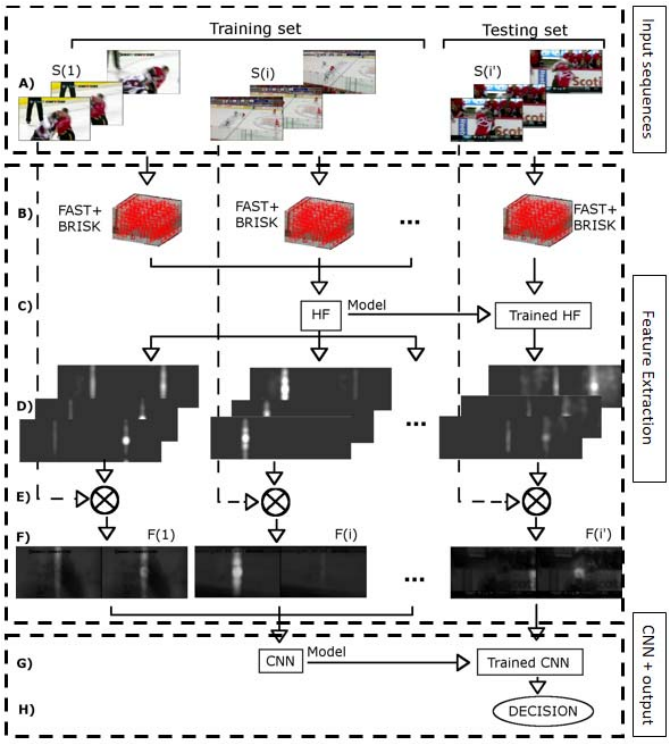

Fig. 4. Detailed diagram of the proposed method. From top to bottom rows: A) Input sequences. B) FAST detector and BRISK features (step 1.a). C) HF classifier and D) spatio-temporal voting models (step 1.b) combined with a Gaussian-filtering step. E) Combination step and F) representative images (step 1.c). These B), C), D), E), and F) steps correspond with the feature extraction phase. G) 2D Convolutional Neural Network with the H) final decision (step 2). The S(i) represents the i^{th} input sequence. The F(i) represents the i^{th} generated representative image for the input S(i).

action distribution in space-time, i.e. modeling where and when each action is more likely to occur.

Finally, each generated F is a single 2D image which will be used as input a 2D Convolutional Neural Network.

The main steps are next described in more detail:

1) The feature extraction step.
   a) Features are extracted for each sequence.
   b) Apply a Hough Forests classifier to the previously extracted features to create a spatio-temporal voting model of action distribution in space-time. This will provide $M(t)$ and $M'(t)$ images.
   c) Combine the previous spatio-temporal voting model with the input sequence in order to create a representative image of the video sequence, as shown in Ecs. (2), (3), (4) and (5).
2) Use a 2D Convolutional Neuronal Network to process the representative image and obtain the final decision for the sequence.

These four steps are shown in Fig. 4 and further described in what follows.

A. Feature Extraction

The feature extraction step aims to obtain a representative image from each input video sequence that correspond with the following three sections.

1) Extract Spatial Features: The first step (1.a) is to extract spatial features from each sequence. The Binary Robust Invariant Scalable Keypoints (BRISK) descriptor is used to extract features [40]. The BRISK descriptor extracts local image features in space characterized by a high variation of the image values. It uses a scale-space detector in combination with the assembly of a bit-string descriptor from intensity comparisons retrieved by dedicated sampling of each keypoint neighborhood. This method is rotation and scale invariant and very efficient. The BRISK descriptor is adjusted to extract a 64-dimension vector of features for each point.

Descriptors may be used with dense or sparse sampling. Sparse sampling is used along with the salient spatial keypoints and is computationally more efficient. The keypoints are obtained using the Features from Accelerated Segment Test (FAST) detector [41]. It requires consecutive pixels which are sufficiently brighter or darker than the central pixel to select it. In addition, the FAST detector is computationally very fast.

To extract spatial features, the FAST detector is applied on the image difference between pairs of consecutive frames. For these located key points in space, the BRISK features are extracted. In summary, for each sequence a set of representative points is located using the FAST detector. Then, for each point a 64-dimension vector of features is extracted using the BRISK descriptor. After that, all these vectors of features (or descriptors) feed the Hough Forests classifier, as explained below.

2) Hough Forests: Hough Forests (step 1.b) ([42], [43] and [44]) consist of a set of random trees [45] that are trained to learn a mapping from densely-sampled D-dimensional feature *cuboids* to their corresponding votes in a Hough space $\mathbb{H} \subseteq \mathbb{R}^H$. The Hough space encodes the distribution of action classes in space and time. The term *cuboid* below is used in a generalized sense to represent a local image patch ($D = 2$) or video space-temporal neighborhood ($D = 3$) depending on the task.

Each tree τ in Hough Forests $\tau = T_{tree}$ is constructed from a set of feature *cuboids* $P_i = (F_i, c_i, d_i)$ that are randomly sampled from the image or video sequence; where F_i are the extracted features from a *cuboid* of fixed size (D) in $\mathbb{R}^D$; c_i is the class label for the sample and d_i is a displacement vector from the cuboid pointing toward the spatio-temporal center of the action. The negative classes have $d_i = 0$. In [42] and [43], cuboids are used with fixed dimensions 16×16 and $16 \times 16 \times 5$ for images and videos respectively. Then for each *cuboid* the grayscale intensity, absolute value of x, y and $time$ derivatives, absolute value of optical flow in x and y are obtained. In this work, Hough Forests are trained with the BRISK features.

Each leaf node L stores the probability of the cuboids belonging to the object class $\varphi(c \mid L)$, estimated as the proportion of cuboids per class label reaching the leaf after training, and $D_c^L = \{d_i\}_{c_i=c'}$ the cuboids respective displacement vectors. In this work, each non-leaf node is modified in order to assign a binary test from a set of input vector features (F). The binary test is now defined by a composition of two features values $(p, q) \in \mathbb{R}^D$ with some offset O_s. Where D is the dimension of input vector, p and q correspond with two different feature positions from the input vector. The binary

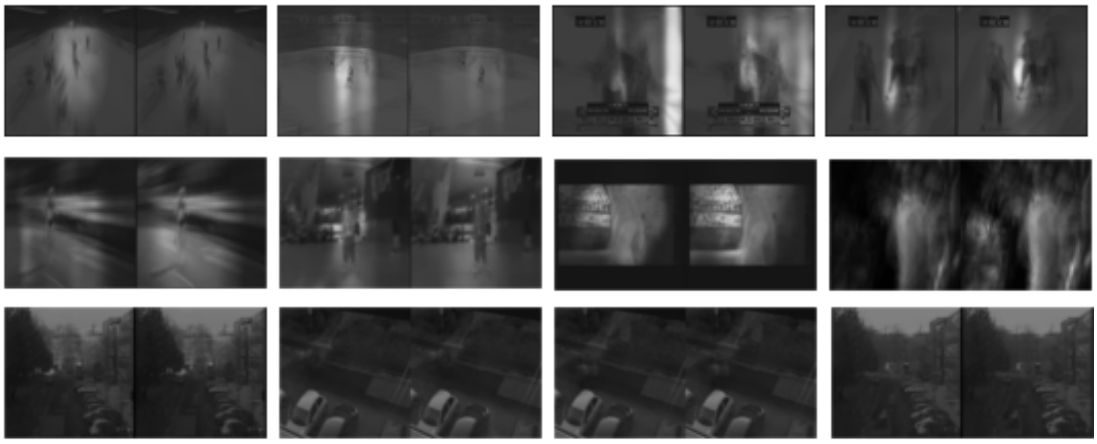

Fig. 5. The representative $F(s)$ images obtained from the Hockey dataset (first row), the Movies dataset (second row) and the Behave dataset (third row). These sequences are the same as in Fig. 1.

test (b) on a non-leaf node is adapted as:

$$b_{p,q,O_s}(F) = \begin{cases} 1, & if\ F(p) < F(q) \times O_s \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

The next step is to construct the Hough Forests according to [43]. The Hough Forests output for a sequence (set of features) is an image with their corresponding votes in Hough space $\mathbb{H} \subseteq \mathbb{R}^H$. Besides, the maxima class probability can be searched by applying a Parzen estimator [46]. The accumulation of the probabilities is made with the summation criteria in order to take the final decision.

The Hough Forests classifier creates a Hough Voting Space when a new sequence (set of features) is tested. Each feature produces a vote in space-time and class. Then, a spatial probability map image is created for each class in each time instant. Afterward, we decide to apply a Gaussian filter on these images to extend the votes on space and reduce noise, see Fig. 4(D). It now is possible to know where and when a fight/non-fight action ($M(t)$ and $M'(t)$, respectively) occurs. This is the spatio-temporal voting model that will be used for the following step. It is actually possible to make a final decision from this spatio-temporal voting model for the sequence. However, the accuracy is still worse than the proposed method as it will be shown in Section IV.

The original output of Hough Forests is a voting volume with the space-time probabilities for each class. The local maximum can be calculated for each class. The local maximum encodes the class, position and length of the action. Then, the temporal-voting model may be generated from the accumulation of these probabilities in time. The summation is preferred over multiplication for better stability, as shown in [42]. This approach is also considered in section IV.

3) Representative Image: Finally (step 1.c), the previous spatio-temporal voting models are combined with each input sequence, as it was explained before, to create a representative image for each video sequence, see some examples in Fig. 5, as Ecs. (2), (3), (4) and (5).

B. 2D Convolutional Neuronal Network

Images computed as described above do not appear natural, see Fig. 5, although both appearance and patterns of motion are recognizable. These images embed useful discriminant information that the Convolutional Neural Network, when properly dimensioned and trained, can leverage. While the whole input sequence matrix contains complete appearance and motion information, the images proposed here are a better representation of motion than appearance.

It is worth noting that the proposed method bears some resemblance with the method [32] described above. Fuzzy regions caused by strong motion were the main source of information in that work. In the proposed method, motion gives rise to fuzzy regions in the input image.

The neural network contains two convolutional layers, see Fig. 6, each followed by a MAX pooling layer. This results in a reduced-resolution output feature map which is robust to small variation in the location of features in the previous layer. Additionally, a fully-connected Inner Product layer is used followed by the ReLU operation, which adds non-linearity. The input representative images are rescaled before to enter the 2D Convolutional Neural Network. This is another parameter to adjust. Finally, the network is trained for the datasets separately.

IV. Experiments

In this section we compare the proposed method to evaluate its performance against different state-of-the-art related algorithms either implemented by the authors or cited from the literature, including "handcrafted" feature and deep feature learning methods. To evaluate the method, a 10-fold cross-validation is carried out. The results reported are the average of accuracy and standard deviation using 5 repetitions.

In the experiments, some non-fight actions can be classified by error as fight actions and vice versa. These cases are denominated false positive (FP) and false negative (FN), respectively. When a fight action is classified as fight, this is called true positive (TP). However, when a non-fight action is classified as a non-fight action, this is called true

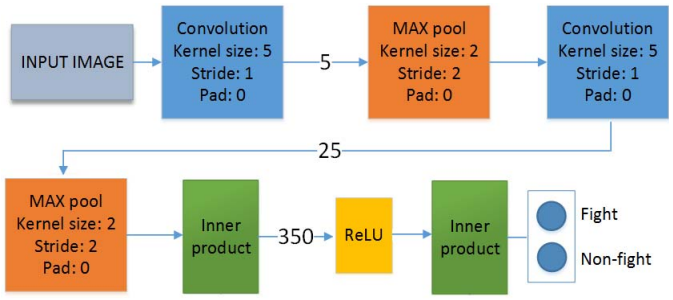

Fig. 6. Proposed 2D Convolutional Neural Network architecture. The kernel size represents the rectangular size of the filter, the stride is the step size in both directions to move the filter across the input image and pad (padding) extends the input volume with zeros around the border. The MAX pool represents the maximum sub-sampling (or pooling) layer. The number after the convolution and inner product layers is the number of generated filters.

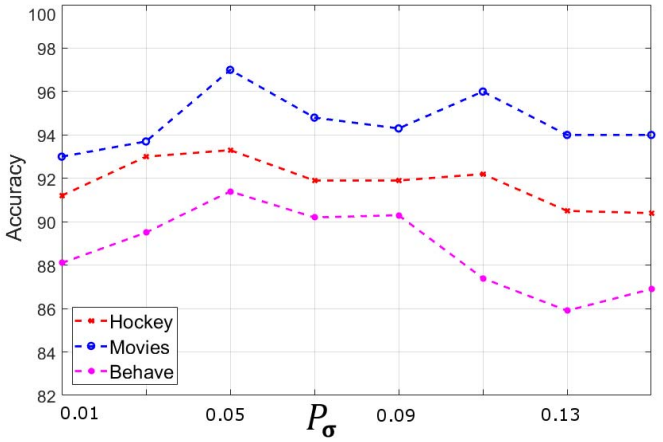

Fig. 7. Relation between accuracies and P_σ for the proposed method on the three considered datasets. The method achieves the best results for the three datasets with a 0.05 P_σ.

negative (TN). The classification accuracy (or simply accuracy) is used here to evaluate the method because it is widely used by other state-of-the-art methods. This measure is the fraction of all sequences that are correctly classified ($TP + TN$) to the total number of sequences. Additionally, note that our experiments have the same number of fight and non-fight sequences. In this scenario, the average of FP and FN is exactly 1− accuracy.

A. Datasets

The experiments of the proposed method are carried out on three different benchmark datasets, i.e., the Hockey [5], Movies [5] and Behave [47] datasets.

The Hockey dataset was explicitly designed for assessing fight recognition. The dataset consists of 1000 clips at a resolution of 360 × 288 pixels, divided into two groups, 500 fights (see Fig. 1 first row) and 500 non-fights, extracted from hockey games of the National Hockey League (NHL). Each clip was limited to 50 frames. The hockey dataset has become a reference for researchers working on this specific task.

The Movies dataset was also explicitly built for assessing fight recognition. The dataset consists of 200 video clips in which fights were extracted from action movies (see Fig. 1 second row). The non-fight videos were extracted from public action recognition datasets. Unlike the hockey dataset, which was uniform both in format and content, these videos depict a wider variety of scenes and were captured at different resolutions. The average resolution is 360 × 250 pixels. Each clip was limited to 50 frames. The Movies dataset generally contains first-person sequences with low or no camera motion. On the other hand, the Hockey dataset contains camera motion and the non-fight actions have a significant and abrupt motion, which makes it more challenging when we try to detect energetic fight actions.

The Behave dataset contains more than 200,000 frames at a resolution on 640 × 480 pixels and various scenarios, including walking, running, chasing, discussing in groups, driving or cycling across the scene, fighting and so on (see Fig. 1 third row). In this paper we have followed the experimental setup carried out in [38], in order to compare the proposed method on the same scenario. The dataset is partitioned into clips with various activities which have been manually labeled as violence or nonviolence. Each clip consists of 100 frames. Finally, 140 clips are randomly picked for violence detection, including 70 violence and 70 non-violence clips.

B. Parameters

Since the proposed method is based on strong motion patterns, energetic actions such as those in the above-mentioned Hockey dataset are particularly challenging. The proposed method uses the following parameters for each step. The FAST detector is used to locate salient point over the difference image for each two consecutive images, using the standard parameters proposed in [41]. The BRISK descriptor is applied on the detected points on the gray-converted original images, using the standard parameters proposed by BRISK authors in [40]. The Hough Forests classifier uses the following parameters: 10 trees, depth 15 for each tree and 2000 tests per node in the training step [42].

Furthermore, to assess the impact of P_σ, a set of experiments have been run using different values of this parameter. The frame size is initially adjusted to 128 × 256. The results are shown in Fig. 7 where the method accuracy rises until it reaches the best results at 0.05 P_σ for the three datasets, decreasing after that. Because of this observed behavior, it is decided to fit this parameter to 0.05 for each dataset in the rest of the experiments.

On the other hand, to assess the impact of the frame sizes for F Eq. (2), another set of experiments was carried out using different sizes of frames. The 60 × 90 size is also considered in order to allow for a better comparison with [36]. As shown in Fig. 8 the proposed method achieves the best results using 64 × 128 frame sizes on the first two datasets and 128 × 256 frame sizes on Behave dataset. As this experiment shows, the proposed method obtains the maximum accuracies rescaling the datasets using their resolutions as reference. It can be seen that the method can achieve the maximum accuracies by rescaling each frame, independently of the dataset, a 25% of its resolution.

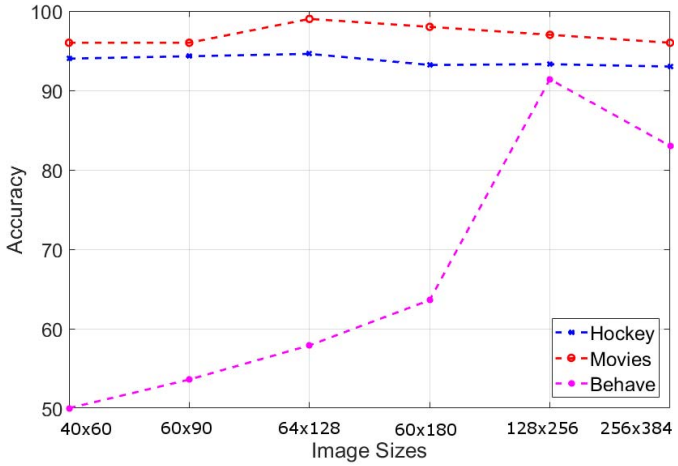

Fig. 8. Relation between accuracies and different frame sizes for the proposed method on the considered datasets. The method achieves the best results on the first two datasets with 64 × 128 and 128 × 256 frame sizes on Behave dataset.

Afterward, the previous rescaled images feed our 2D Convolutional Neuronal Network. The network architecture is shown in Fig. 6. The 2D Convolutional Neural Network is trained from scratch with a batch size of 64, 60 epochs, a base learning rate = 0.001, momentum = 0.9 and weight decay factor = 0.5 every 10 epochs. The learning rate is subject to inverse decay and a SoftMax loss function is used. The method described in [48] ("Xavier's method") is used to fill the network weights prior the starting of training. Finally, the network is optimized using the stochastic gradient descent method.

C. Results

The proposed method is compared with different "handcrafted" and deep learning based methods on the litereture, all reviewed in Section II. With regards to the "handcrafted feature" category, the Violent Flows (ViF) method [21] and other more recent methods such as [22], [32], and [33] were considered on the first two datasets. The Violent Flows (ViF) and LMP methods were executed and compared on these datasets on [33]. The MoIWLD [38] method was considered on Hockey and Behave datasets. For this category three different state-of-the-art classifiers were applied: SVM (linear kernel), Adaboost (100 classifiers, weighted voting) and Random Forests (50 trees). With regards to the "deep feature learning" category, the 3D Convolutional Neural Network (3D-CNN) method [36] was considered on the Hockey dataset. Besides, we have tested the C3D method [37] on the first two datasets. This method achieved the best results with 100 × 120 frame sizes, maximum of 2000 iterations, base learning rate of 0.0001, momentum of 0.9 and weight decay factor of 0.005. The optimization is done using stochastic gradient descent. These parameters are obtained from the original paper [37].

Additionally, we have isolatedly tested the BRISK descriptor with Hough Forest classifier on the first two datasets with the same parameters as the proposed method. Besides, we have isolatedly tested our 2D-CNN architecture without using the spatio-temporal Hough Forests voting. To achieve that, the $M(t)$ Eq. (3) and $M'(t)$ Eq. (4) images are fixed to 1.

TABLE I

ACCURACIES AND STANDARD DEVIATION ON THE FIRST TWO CONSIDERED DATASETS. IN BOLD THE BEST AVERAGE OBTAINED ACCURACIES ARE SHOWN ON EACH DATASET

Features/Classifier		Dataset	
		Movies	Hockey
ViF [33]	SVM	96.7±0.3%	82.3±0.2%
	Adaboost	92.8±0.4%	82.2±0.4%
	Random Forests	88.9±1.2%	82.4±0.6%
LMP [33]	SVM	84.4±0.8%	75.9±0.3%
	Adaboost	81.5±2.1%	76.5±0.9%
	Random Forests	92±0.1%	77.7±0.6%
Deniz [32]	SVM	85.4±9.3%	90.1±0%
	Adaboost	98.9±0.2%	90.1±0%
	Random Forests	-	-
Serrano [33]	SVM	87.2±0.7%	72.5±0.5%
	Adaboost	81.7±0.2%	71.7±0.3%
	Random Forests	97.8±0.4%	82.4±0.6%
MoIWLD [38]	Classifier [38]	-	**96.8±1%**
3D-CNN [36]		-	91± - %
C3D [37]		93.6±0.8%	87.4±1.2%
FAST+BRISK	Hough Forests	95.5±0.6%	90.3±0.7%
Only our 2D-CNN		93.1±0.3%	87.8±0.3%
Proposed method		**99±0.5%**	94.6±0.6%

TABLE II

ACCURACIES AND STANDARD DEVIATION (STD) ON BEHAVE DATASET. IN BOLD THE BEST OBTAINED ACCURACY IS SHOWN

Algorithm	Accuracy±std
HOG + BoW [5]	58.97±0.34%
HOF + BoW [5]	60.03±0.28%
HNF + BoW [5]	58.24±0.31%
ViF [5]	83.62±0.19%
MoSIFT + BoW [5]	62.78±0.23%
MoWLD + BoW [38]	81.65±0.18%
MoIWLD + BoW [38]	81.98±0.15%
RVD [49]	85.29±0.16%
AMDN [50]	84.22±0.17%
SRC [51]	82.7±0.14%
MoWLD + Sparse Coding [52]	87.07±0.13%
MoIWLD + Modified dictionary [38]	88.83±0.11%
Proposed method	**91.42±0.14%**

Finally, the proposed method is run using the previous explained parameters on the first two datasets and compared to the state-of-the-art approaches. These results are shown in Table I reporting the accuracies and standard deviation. Table II shows the results obtained by the proposed method on Behave dataset, reporting the accuracies and standard deviation (std). The Movies dataset only contains 200 videos clips, a 10-fold cross validation has been used to get 180 video clips for training and 20 for test. As this is the procedure followed by the authors, it allows to compare the results of the proposed method with the state-of-the-art methods.

Overall the proposed method obtains the best accuracies in Movies and Behave datasets. To put the results in context, Tables I, II and III show our results and other accuracies obtained in the literature in the Movies, Hockey and Behave

TABLE III

RESULTS REPORTED IN THE LITERATURE ON THE FIRST TWO CONSIDERED DATASETS, SORTED BY YEAR OF PUBLICATION (OLDEST AT THE TOP) SHOWING THEIR ACCURACIES

Methods/Techniques	Accuracy
Movies Dataset	
STIP+BoW+SVM [5]	89.5%
Histograms of frequency-based motion intensities +Adaboost [32]	98.9%
Motion blobs+Random Forests [33]	96.9%
Optical flow+substantial derivative+BoW+SVM [35]	97.8%
STEC+Hough Forests [53]	92%
Hockey Dataset	
Dense optical flow+BoW+SVM [54]	74.1%
STIP+BoW+SVM [5]	91.7%
MBP+Random Forests [30]	76.6%
Variance of optical flow, SVM [28]	86.9%
MoSIFT+kernel density estimation+sparse coding [29]	94.3%
Histograms of frequency-based motion intensities +Adaboost [32]	90.1%
Co-ocurrence texture measures+edge cardinality and pixel intensity difference+Random Forests [31]	77.2%
Motion blobs+Random Forests [33]	82.4%
MoSIFT+BoW+SVM [34]	93.3%
STEC+Hough Forests [53]	82.6%

TABLE IV

CONFUSION MATRIX FOR EACH DATASET. THE COLUMNS ARE THE GROUNDTRUTH AND ROWS THE PREDICTED CLASS

	Movies		Hockey		Behave	
	Fight	NoFight	Fight	NoFight	Fight	NoFight
Fight	98	2	469	31	63	7
NoFight	0	100	23	477	5	65

datasets. The best absolute accuracy reported in the literature for Movies dataset was 98.9%, using histograms of frequency-based motion intensities as the main descriptor and an AdaBoost classifier. The best absolute accuracy reported in the literature for the Hockey dataset was 96.8%, using MoIWLD descriptor. The best absolute accuracy reported in the literature for the Behave dataset was 88.8%, using also MoIWLD descriptor. The proposed method provides an absolute accuracy of 99%, 94.6% and 91.4% on Movies, Hockey and Behave datasets, respectively.

In Table IV we provide the confusion matrix for each dataset given by the proposed method. We also considered the false acceptance ratio (FAR) and false rejection rate (FRR). FAR and FRR are defined as $FN/(TP + FN)$ and $FP/(TP + FP)$, respectively. Then, the proposed method provides a FAR of 0%, 4.7% and 7.4%, and FRR of 2%, 6.2% and 10% in the Movies, Hockey and Behave datasets, respectively.

In Table V we compare the times required to process an input sequence of the reference Hockey dataset with other state-of-the-art methods. The proposed method reports the computational time for each step: extract spatial features, Hough forests classifier, create the representative image and the 2D-CNN. This comparative is carried out using only CPU, however the method can be executed in GPU. The proposed method is fast enough for practical application in real time obtaining a good a tradeoff between accuracy and

TABLE V

TIME REQUIRED TO PROCESS A SEQUENCE OF THE REFERENCE HOCKEY DATASET USING AN INTEL I7 3.4Ghz

Method	Time	Conditions
MoSIFT [29]	43.22 s	Time to extract features, Time to extract features, this does not include time for feature reduction and/or classification
ViF [21]	16.27 s	-
LMP [22]	5.43 s	-
Deniz [32]	1.51 s	-
Serrano [33]	0.95 s	-
3D-CNN [36]	0.19 s	-
C3D [37]	0.41 s	-
Proposed	0.46 s	Took 151.2, 105.3, 67.4 and 138.2 ms for each step (the first and latter parts the codes were interpreted)

computational time. To achieve this goal, the proposed method is adjusted using the minimum frame sizes with the highest accuracy, see Figure 8.

Finally the proposed method outperforms state-of-the-art accuracies in two of the three considered datasets, specifically improving upon the previous use of deep learning in this task. Besides, the framework of the proposed 2D Convolutional Neural Network and its "handcrafted" inputs is efficient. The proposed method achieved the best results in the challenging Behave and Movies datasets because the camera position is mostly static. The proposed approach builds a better estimate of the spatio-temporal distribution maps in such scenarios. Still, with the Hockey dataset in which there is frequent and large camera motion the obtained results are very close to the related methods. The non-static camera does not necessarily imply bad accuracy (in the sense of higher confusion). In fact, in the Hockey dataset there are relatively distinct spatio-temporal patterns for the two classes: in non-fights there is a typically a play scene with the camera panning fast and players moving too, whereas in fights there is typically zoom and two players fighting covering almost the whole frame. On the other hand, the Behave dataset gave the poorest performance and we believe that is due to the fact that in that dataset the motion elements are very small (subjects captured at a long distance from the camera). This would be evidenced in Fig 8 which shows that the Behave dataset is clearly the most sensitive to input image resolution.

V. CONCLUSIONS

The task of fight recognition has attracted interest of many researchers in the last few years. Given the direct applications in surveillance and movie rating, the efforts are fully justified. "Handcrafted" features, some used for generic action recognition work, were used for this task before the advent of feature learning methods. The latter methods, represented by the powerful Convolutional Neural Networks, automatically learn features for the task at hand. The two compared works in this line used a 3D convolutional Neural Network that processes the entire video sequence as input. Inspired by psychophysics experiments suggesting that motion features may be more important for this specific task, this paper proposed a hybrid approach. The rich spatio-temporal voting information from

Hough Forests classifier is used to leverage the representative image for each sequence, which is fed with BRISK features that capture motion and appearance from a video sequence. Finally, a much simpler 2D Convolutional Neural Network is fed with the "handcrafted" images. The proposed method demonstrates superiority over different 'handcrafted" feature and 3D Convolutional Neural Network approaches for this binary recognition task. The proposed method provides the best absolute accuracy in two of the three considered datasets with the 99%, 94.6% and 91.4% accuracies in the Movies, Hockey and Behave datasets, respectively. In addition, the confusion matrix, FAR and FRR metrics are calculated for each dataset in order to provide a better evaluate of recognition performance. Besides, the framework is computationally very efficient, allowing for real-time processing. Future work will seek to a) compare the proposed method with other deep learning methods specifically designed to encode spatio-temporal information, and b) assess the validity of the method for generic action recognition.

REFERENCES

[1] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, 2010.

[2] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

[3] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, 2013.

[4] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. 9th Int. Conf. Comput. Vis.*, Oct. 2003, pp. 432–439.

[5] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Proc. 14th Int. Congr. Comput. Anal. Images Patterns*, 2011, pp. 332–339.

[6] R. Blake and M. Shiffrar, "Perception of human motion," *Annu. Rev. Psychol.*, vol. 58, pp. 47–73, Jan. 2007.

[7] M. Saerbeck and C. Bartneck, "Perception of affect elicited by robot motion," in *Proc. 5th ACM/IEEE Int. Conf. Human-Robot Interaction*, Piscataway, NJ, USA, Mar. 2010, pp. 53–60.

[8] T. J. Clarke, M. F. Bradshaw, D. T. Field, S. E. Hampson, and D. Rose, "The perception of emotion from body movement in point-light displays of interpersonal dialogue," *Perception*, vol. 34, pp. 1171–1180, Oct. 2005.

[9] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*. Berlin, Germany: Springer, 2007, pp. 71–82.

[10] S. Hidaka, "Identifying kinematic cues for action style recognition," in *Proc. Annu. Meeting Cogn. Sci. Soc.*, vol. 34, Jan. 2012, pp. 1679–1684.

[11] O. Oshin, A. Gilbert, and R. Bowden, "Capturing the relative distribution of features for action recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, Mar. 2011, pp. 111–116.

[12] A. Bobick and J. Davis, "An appearance-based representation of action," in *Proc. 13th Int. Conf. Pattern Recognit.*, vol. 1, Aug. 1996, pp. 307–312.

[13] J. Nam, M. Alghoniemy, and A. H. Tewfik, "Audio-visual content-based violent scene characterization," in *Proc. ICIP*, Oct. 1998, pp. 353–357.

[14] W.-H. Cheng, W.-T. Chu, and J.-L. Wu, "Semantic context detection based on hierarchical audio models," in *Proc. 5th ACM SIGMM Int. Workshop Multimedia Inf. Retr.*, New York, NY, USA, 2003, pp. 109–115.

[15] T. Giannakopoulos, D. Kosmopoulos, A. Aristidou, and S. Theodoridis, "Violence content classification using audio features," in *Advances in Artificial Intelligence* (Lecture Notes in Computer Science), vol. 3955. Berlin, Germany: Springer, 2006, pp. 502–507.

[16] W. Zajdel, J. D. Krijnders, T. Andringa, and D. M. Gavrila, "CASSANDRA: Audio-video sensor fusion for aggression detection," in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.*, Sep. 2007, pp. 200–205.

[17] Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao, "Detecting violent scenes in movies by auditory and visual cues," in *Proc. 9th Pacific Rim Conf. Multimedia*. Berlin, Germany: Springer-Verlag, 2008, pp. 317–326.

[18] D. Chen, H. Wactlar, M.-Y. Chen, C. Gao, A. Bharucha, and A. Hauptmann, "Recognition of aggressive human behavior using binary local motion descriptors," in *Proc. 30th Annu. Int. Conf. Eng. Med. Biol. Soc.*, Aug. 2008, pp. 5238–5241.

[19] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis, "Audio-visual fusion for detecting violent scenes in videos," in *Artificial Intelligence: Theories, Models and Applications*. Athens, Greece: Springer-Verlag, 2010, pp. 91–100.

[20] L.-H. Chen, C.-W. Su, and H.-W. Hsu, "Violent scene detection in movies," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 25, no. 8, pp. 1161–1172, 2011.

[21] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *Proc. Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2012, pp. 1–6.

[22] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1576–1588, Aug. 2012.

[23] C. Demarty, C. Penet, G. Gravier, and M. Soleymani, "The MediaEval 2012 affect task: Violent scenes detection," in *Proc. MediaEval Workshop*, Pisa, Italy, 2012, pp. 1–3.

[24] A. Datta, M. Shah, and N. Da Vitoria Lobo, "Person-on-person violence detection in video data," in *Proc. 6th Int. Conf. Pattern Recognit.*, vol. 1, Aug. 2002, pp. 433–438.

[25] Z. Szlavik *et al.*, "Behavior and event detection for annotation and surveillance," in *Proc. Int. Workshop Content-Based Multimedia Indexing*, Jun. 2008, pp. 117–124.

[26] M. Andersson, S. Ntalampiras, T. Ganchev, J. Rydell, J. Ahlberg, and N. Fakotakis, "Fusion of acoustic and optical sensor data for automatic fight detection in urban environments," in *Proc. 13th Conf. Inf. Fusion (FUSION)*, Jul. 2010, pp. 1–8.

[27] M.-Y. Chen, L. Mummert, P. Pillai, A. Hauptmann, and R. Sukthankar, "Exploiting multi-level parallelism for low-latency activity recognition in streaming video," in *Proc. 1st Annu. ACM SIGMM Conf. Multimedia Syst.*, New York, NY, USA, 2010, pp. 1–12.

[28] H. Jian-Feng and C. Shui-Li, "Detection of violent crowd behavior based on statistical characteristics of the optical flow," in *Proc. 11th Int. Conf. Fuzzy Syst. Knowl. Discovery (FSKD)*, Aug. 2014, pp. 565–569.

[29] X. Long, G. Chen, Y. Jie, W. Qiang, and Y. Lixiu, "Violent video detection based on MoSIFT feature and sparse coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 3538–3542.

[30] B. Florian, L. Jie, E. Arne, and R. Bodo, "Motion binary patterns for action recognition," in *Proc. 3rd Int. Conf. Pattern Recognit. Appl. Methods*, 2014, pp. 385–392.

[31] K. Lloyd, P. L. Rosin, D. Marshall, and S. C. Moore, "Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (GLCM)-based texture measures," *Mach. Vis. Appl.*, vol. 28, no. 3, pp. 361–371, May 2017, doi: 10.1007/s00138-017-0830-x.

[32] O. Deniz, I. Serrano, G. Bueno, and T.-K. Kim, "Fast violence detection in video," in *Proc. 9th Int. Conf. Comput. Vis. Theory Appl. (VISAPP)*, Lisbon, Portugal, vol. 2, Jan. 2014, pp. 478–485.

[33] 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.

[34] T. Senst, V. Eiselein, and T. Sikora, "A local feature based on Lagrangian measures for violent video classification," in *Proc. 6th IET Int. Conf. Imag. Crime Detection Prevention*, London, U.K., Jul. 2015, pp. 1–6.

[35] S. Mohammadi, H. Kiani, A. Perina, and V. Murino, "Violence detection in crowded scenes using substantial derivative," in *Proc. 12th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2015, pp. 1–6.

[36] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, "Violence detection in video by using 3D convolutional neural networks," in *Advances in Visual Computing* (Lecture Notes in Computer Science), vol. 8888. Cham, Switzerland: Springer, 2014, pp. 551–558.

[37] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[38] T. Zhang, W. Jia, X. He, and J. Yang, "Discriminative dictionary learning with motion weber local descriptor for violence detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 696–709, Mar. 2017.

[39] J. W. Davis and A. F. Bobick, "The representation and recognition of human movement using temporal templates," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1997, pp. 928–934.
[40] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 2548–2555.
[41] E. Rosten and T. Drummond, "Fusing points and lines for high performance tracking," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, vol. 2, Oct. 2005, pp. 1508–1515.
[42] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2188–2202, Nov. 2011.
[43] A. Yao, J. Gall, and L. Van Gool, "A Hough transform-based voting framework for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2061–2068.
[44] G. Garcia-Hernando, H. J. Chang, I. Serrano, O. Deniz, and T.-K. Kim, "Transition Hough forest for trajectory-based action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–8.
[45] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[46] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1065–1076, Sep. 1962.
[47] S. Blunsden and B. Fisher, "The BEHAVE video dataset: Ground truthed video for multi-person behavior classification," *Ann. BMVA*, vol. 4, no. 4, pp. 1–11, 2010.
[48] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist. Soc.*, 2010, pp. 249–256.
[49] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang, and X. He, "A new method for violence detection in surveillance scenes," *Multimedia Tools Appl.*, vol. 75, no. 12, pp. 7327–7349, 2016.
[50] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe. (2015). "Learning deep representations of appearance and motion for anomalous event detection." [Online]. Available: https://arxiv.org/abs/1510.01553
[51] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[52] T. Zhang, W. Jia, B. Yang, J. Yang, X. He, and Z. Zheng, "MoWLD: A robust motion image descriptor for violence detection," *Multimedia Tools Appl.*, vol. 76, no. 1, pp. 1419–1438, 2017.
[53] 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," *Mach. Vis. Appl.*, vol. 29, no. 2, pp. 207–217, 2018.
[54] H. Wang, A. Kläser, C. Schmid, and C.-L. Lin, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 3169–3176.

Ismael Serrano received the degree in computer science and the Ph.D. degree *(cum laude)* from the University of Castilla–La Mancha, Spain, in 2012 and 2016, respectively. His Ph.D. thesis was on fight detection in video using computer vision and machine learning techniques. He scored the highest mark in his final degree project about person detection. He has served as a Researcher Collaborator with Imperial College London, U.K., and Leica Biosystems, Ireland. He is currently a Ph.D. Researcher with the Industry and Advanced Manufacturing Department, Vicomtech Foundation. He has authored over 15 papers in journals and conferences. He has published two books on OpenCV. His research interests are mainly focused on computer vision and machine learning, especially on deep learning.

Oscar Deniz (SM'98) has been a Visiting Researcher with Carnegie Mellon University, USA, Imperial College London, U.K., and Leica Biosystems, Ireland. He is currently an Associate Professor with the University of Castilla La–Mancha and contributes to VISILAB. He is the Coordinator of the European H2020 Project Eyes of Things and a Partner in the European H2020 Project BONSEYES. He has authored over 50 refereed papers in journals and conferences. His research interests are mainly focused on computer vision and pattern recognition. He is with the AAAI, SIANI, CEA-IFAC, AEPIA, AERFAI-IAPR, and The Computer Vision Foundation. He serves as an Academic Editor of *PLOS One* journal. He is a Reviewer/Technical Expert for EU programs and an Advisory Board Member of the H2020 Project TULIPP.

Jose Luis Espinosa-Aranda received the degree in computer engineering and the Ph.D. degree in computer science from the University of Castilla–La Mancha, Spain, in 2009 and 2014, respectively. He is currently a Research Assistant of the VISILAB Group and also an Associate Professor with the University of Castilla–La Mancha. His current research interests include artificial intelligence and computer vision.

Gloria Bueno (M'99) received the degree in physics from UCM, Madrid, Spain, in 1993, and the Ph.D. degree in machine vision from Coventry University, U.K., in 1998. She has served as a Visiting Researcher with Carnegie Mellon University, USA, and Leica Biosystems, Ireland. She has experience as a principal researcher in several research centers, such as CNRS, Louis Pasteur University, Strasbourg, France, Gilbert Gilkes & Gordon Technology, U.K., and CEIT San Sebastian, Spain. She is currently a Professor with the Engineering School, University of Castilla–La Mancha. She has been a principal researcher of different national and international projects focused on artificial intelligence and image processing. She is the Coordinator of the European AIDPATH Project—Academia and Industry Collaboration for Digital Pathology. She has authored two patents, four registered software, and over 80 refereed papers. She is with several societies, such as ESDIP, CEA-IFAC, SEMF, and SEIB.