IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 July 2022, accepted 25 July 2022, date of publication 29 July 2022, date of current version 4 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3195035

RESEARCH ARTICLE

Architecture for Automatic Recognition of Group Activities Using Local Motions and Context

LUIS FELIPE BORJA-BORJA[1], **JORGE AZORÍN-LÓPEZ**[2], **MARCELO SAVAL-CALVO**[2], **ANDRÉS FUSTER-GUILLÓ**[2], **AND MARC SEBBAN**[3]

[1]Facultad de Ingeniería y Ciencias Aplicadas, Universidad Central del Ecuador, Quito 170129, Ecuador
[2]Computer Technology Department, University of Alicante, 03690 Alicante, Spain
[3]Laboratoire Hubert-Curien, UMR CNRS 5516, Université de Lyon, Université de Saint-Etienne, 42000 Saint-Étienne, France

Corresponding author: Jorge Azorín-López (jazorin@ua.es)

This work was supported by the Spanish State Research Agency (AEI) funded by the Ministry of Science and Innovation [Ministerio de Ciencia e Innovación (MCIN)]/AEI/10.13039/501100011033 under Grant PID2020-119144RB-I00.

ABSTRACT Currently, the ability to automatically detect human behavior in image sequences is one of the most important challenges in the area of computer vision. Within this broad field of knowledge, the recognition of activities of people groups in public areas is receiving special attention due to its importance in many aspects including safety and security. This paper proposes a generic computer vision architecture with the ability to learn and recognize different group activities using mainly the local group's movements. Specifically, a multi-stream deep learning architecture is proposed whose two main streams correspond to a representation based on a descriptor capable of representing the trajectory information of a sequence of images as a collection of local movements that occur in specific regions of the scene. Additional information (e.g. location, time, etc.) to strengthen the classification of activities by including it as additional streams. The proposed architecture is capable of classifying in a robust way different activities of a group as well to deal with the one-class problems. Moreover, the use of a simple descriptor that transforms a sequence of color images into a sequence of two-image streams can reduce the curse of dimensionality using a deep learning approach. The generic deep learning architecture has been evaluated with different datasets outperforming the state-of-the-art approaches providing an efficient architecture for single and multi-class classification problems.

INDEX TERMS Neural network architecture, one-class classification, multi-class classification.

I. INTRODUCTION

The growing global population makes it necessary to develop and improve automatic methods to analyse and recognize activities of people groups in public areas for many reasons including safety (e.g. recent restrictions on meetings due to COVID pandemic) and security (e.g. demonstrations, terrorism, etc.). Usually, human personnel visualize and analyse data from surveillance cameras in order to provide the required actions and decisions depending on the specific problem. This task is very time consuming and not always possible to perform online because, on the one hand, it can be expensive but, above all, the limitations related to human visual inspection: tiredness, fatigue, lack of attention, etc. As a result, camera monitoring is usually consulted after a fact. Recently, however, special attention is being paid to solving these problems with artificial vision and machine learning techniques.

Despite the progress of this area of research, there are many challenges on human activity recognition, mainly related to situations where more than one individual is analyzed. For example, how to distinguish between a street fight and a gather of friends playing, detect an abnormal behavior of an individual among a crowd, or how to analyse a group of people activity over a long period of time. Moreover, the context where the activity takes place has a great impact on the

The associate editor coordinating the review of this manuscript and approving it for publication was Francesco Piccialli.

final decision. It allows to determine the type of the activity of if it is normal or abnormal. For instance, a fight in a street is considered abnormal; however, a fight in a boxing ring can be considered as a normal activity. The context defines all the aspects that involve the situation, as the place as we have just mentioned, the time, the conditions, etc.

Researchers have proposed different strategies to address partial problems of the challenge of analysing group behaviour. In this way, the works propose partial solutions to determine movements, actions, activities of groups, other focus on small groups or crowds, etc. [8]. Therefore, it is necessary to advance in the proposal of a generic architecture that allows to automate the monitoring of groups of different size and with different levels of semantics. This motivates our proposal, a multi-stream deep learning architecture (see Fig. 1) able to learn activities of groups using local motion features (Section III). The two main streams of the architecture correspond to the deep-learning variant of the Activity Descriptor Vector (ADV) [4]. This is a simple descriptor capable of representing the trajectory information of a sequence of images as a collection of local movements that occur in specific regions of the scene. The use of motion features helps to reduce the curse of dimensionality using a deep learning approach. Moreover, additional streams are considered to incorporate context information (e.g. location, time, etc.) to enhance the classification of activities. As specific cases, in this paper we instantiate the generic architecture in Section IV-A for multi-class analysis with two above-mentioned streams corresponding to the deep-learning ADV variant, with a final decision stage that outperforms state-of-the-art proposals with over 2% accuracy. Also, the architecture is instantiated for one-class classification in Section IV-B, with three streams including the two ADV variant as in multi-class and a third of context awareness. As it will be in detail presented in Section V, one-class classification instance also improves previous works.

The paper contribution is two-fold:

- The main contribution is a multi-stream generic architecture for human group activity recognition, independently of the number of people in the group, that uses local motions features and incorporate context information. The architecture has been instantiated for multi-class classification and for one-class classification. These two instances include different specific characteristics that, on average, cover a wide range of possibilities.
- The deep-learning ADV variant (D-ADV) as a simple descriptor that transforms a sequence of color images into a sequence of two-image streams that allow to represent spatio-temporal information using images. They can be used as inputs of deep neural networks models. Compared to end-to-end architectures, our proposal allows to train a classification network from features rather than from raw data reducing the space of solutions and, in consequence, less data is need to train.

II. RELATED WORK

The study of human activity has been an important goal of computer vision since its inception and has developed considerably in recent years. To address this problem, researchers have proposed several methods over the past years based on traditional Artificial Neural Networks and, more recently, on deep learning architectures.

The works use different strategies to address partial problems of the challenge of analysing group behaviour. In this way, the works propose partial solutions to determine movements, actions, activities of groups, other focus on small groups or crowds, etc. Applying robust solutions at each stage of the process allows higher level abstraction layers to be addressed with better performance. In general, the first stages deal with the extraction of motion information and video object segmentation [30]–[32]. In [14] the authors propose a vision-based solution to identify Activities of Daily Living (ADL), through skeletal data captured with an RGB-D camera. After the decomposition of a skeletal sequence into short time segments, the activities are classified through a two-layer network called Long-Short Term Memory Network (LSTM), which allows analyzing the sequence at different levels of temporal granularity. The proposal is evaluated in the dataset Watch-n-Patch [56], in which there are examples 11 different daily activities of people such as: bringing things from the fridge, turning things to the fridge, spilling liquid, drinking a liquid, leaving the kitchen, bringing things from the oven, putting things into the microwave, preparing food, filling a kettle, connecting a kettle to the electrical outlet, moving the kettle. The main contribution of the authors is a model of activities of multiple scales and time dependence, based on the comparison of the characteristics of the context that characterize the results of previous recognitions, and a hierarchical representation with a recognition layer of low level behavior units and another high level unit. That is, it is a solution that handles two different levels of semantics.

Human actions in video sequences are usually defined as three-dimensional (3D) space-time signals that characterize both the visual and dynamic appearance of the movement of the human beings and objects involved. Taking into account the success generated by the positive results of Convolutional Neural Networks (CNN) for image classification, recent attempts have been made to learn CNN 3D to recognize human actions in videos, but due to the high complexity of the training of 3D convolution cores and the need for large amounts of training videos that this type of networks requires for their training, there are only few success stories, being a broad topic still requiring maturity in research.

Regarding the study of groups of people, advances in analysing behaviour are limited to very concrete and simple activities or actions, usually of short duration (low semantic component) such as a actions in sport games [11], [35], [40], [50], detection interactions of people inside a group [13], [54], [57], inter-group violence [48], [61], [62], among others. If we increase the number of people in the group, becoming crowds, the level of semantics is even lower, being specifically limited to tasks such as counting people and calculating crowd density [6], [16], [23], [65] or detecting movements of a mass of people or crowd collisions [19], [37], [46], [68], mainly for the purpose of security tasks. It is important to highlight the work in [51], in which the authors present a model of learning based on contextual relationships that uses a deep neural network to recognize activities in a video sequence. The proposed model involves contextual learning using a bottom-up approach, learning from individual human actions to group-level activities, and learning from scene information. Taking into account that it can identify group and individual activities would be considered a progress in this field of research, since it is one of the first works that can identify behaviors according to the number of people in the scene, clarifying that it is not yet reached the level of identification of behaviors in crowds with the same application [51], [52].

Finally, it is important to analyse the one-class classification approaches. They become particularly relevant when an adequate dataset is not available for all the classes that the classification models have to learn. There exists many situations in which it is necessary to identify one class among others but only having examples of this class. For example, in video surveillance, it is common to have information on normal actions or activities while it is very difficult to have examples of criminal actions. The same is valid for credit card transactions, where the large volume is related to legal transactions but the objective is to detect fraudulent ones. The models have to be trained without or scarce samples of abnormal situations, in which the goal is to learn from data the meaning of ''normal''. Deviations or data different from this definition are considered as anomalies or ''abnormal''. The problem of having most (or all) examples of a particular class becomes a bigger problem when using deep learning techniques as they are large consumers of data. Some proposals have been presented to address the one-class classification problem using deep neural networks. In particular, it is important to highlight the work of Chalapathy *et al.* [9] that proposed a model of a one-class neural network (OC-NN) to detect anomalies in complex datasets. Among the work specifically addressing activity recognition in groups or crowds, different neural network models have been used to solve the problem. For example, the use of Convolutional Neural Networks (CNN) is shown in the work of Li *et al.* [27] where a new colorization of images including other information as optical flow is used as input of CNNs to detect objects and their anomalies. Also, in [48], Su *et al.* integrate the one-class Support Vector Machine into a CNN proposing the Deep One-Class (DOC) model. One widely used model has been autoencoders (AE) where they attempt to extract features from images to form a new space in which to decide the existence of normal activities. In this way, Saokrou *et al.* [45] propose a cubic-patch-based method based on a cascade of AE to represent the information in the patches; Vu *et al.* [53] propose representation learning using Denoising Autoencoders (DAEs); and the works in [58], [59] are based on multiple Stacked Denoising AutoEnocders (SDAEs). Finally, Generative Adversial Networks (GAN) trained using normal data have been used to learn an internal representation of the scene normality [41], [42].

III. PROPOSED GENERIC ARCHITECTURE

Currently, the methods of Deep Learning are achieving great results that are revolutionizing the way to address the problems of Artificial Vision. These techniques can solve problems that previously could not be solved, especially in image recognition. End-to-end (E2E) architectures have various advantages, such as the simplicity of coding and the training process. Also, E2E networks internally learn features that describe the data in a way to produce an expected result. Nevertheless, they are highly dependent on the dataset diversity to be generally applicable. Also, a large number of examples is required in most cases to train a large number of parameters in the networks [5], [47]. Some human behaviour datasets are quite small, or not fully labelled. That hardens the training of E2E deep learning architectures. Some proposals have overcome this problem using transfer learning, such as in [70] where pre-trained weights of a VGG16 are used. However, this approach still depends on the previously trained parameters.

In order to cope with these problems, a different dual-stage solution is presented in this paper. It uses pre-defined features that convert the data to a different known space and train a network with those features. Using descriptors reduce the space of solutions, helping to learn tasks with fewer data. In this way, we detach the motion description (D-ADV in this case) from the behaviour classification and allow the system to train ones regardless of the specific dataset since we train the network from the features. Since those features are known (i.e. they always have the same shape and range) and not learnt from a specific dataset, and the network learns from them, the generality is implicit regardless of the original data. This idea has been used before with generic descriptors [20] or simple optical flow [49], but not with specific descriptors for behaviour classification.

Another advantage of this solution is that a smaller amount of data is required to train the architecture, which is important for both small and large datasets. In the case of small for the reasons explained before, and for large to avoid the need of fully labelling the entire dataset, which can be a tedious and time-consuming task.

Furthermore, the learning part is based on the paradigm of ''ensemble learning'' with, in the general case, any number of classifiers, and in the specific instances of this paper, two specific classifiers for the vertical and horizontal motion, and three for the one-class classification that includes the context. Ensemble learning [15], [60] has proven outstanding results in many papers and challenges.

With this idea in mind, the key contribution in this paper is a multi-stream architecture, depicted in Figure 1, that analyses group human behaviour using a movement descriptor and a classification stage to detect different activities. For the movement descriptor, in this paper, we present the D-ADV based on the Activity Descriptor Vector (ADV) for deep learning classification, as explained in detail in Section III-A2. The descriptor is attached to the two first streams of the architecture.

The classification stage is coupled to a learning block as it is shown in Figure 1, that classifies the activity taking into account the context. There are two main classification strategies, one distinguishes among multiple classes (multi-class classification) and the other between known and unknown activities (one-class classification). The second is common in human behaviour as most of the times people act normally, so it is highly unlikely to find enough data to train a network for abnormal activities. Thus, a system is trained to detect the known classes and ‘‘classify’’ the rest as abnormal or unknown. Using a motion descriptor let the system be adaptable to any of these two strategies, that we instantiated after in Section IV-B for one-class and Section IV-A for multi-class.

A. LOCAL DESCRIPTOR MOTION

The first block of the architecture aims to extract a representation of the movements that occur in the scene. Here a variant of the ADV descriptor that allows to be used in image-based deep learning systems is presented. Specifically, the Deep-ADV is proposed that allows describing a scene with images of local motions in regions of the scene. For the sake of completeness we first briefly describes the main aspects of the ADV and later the proposed D-ADV.

1) ACTIVITY DESCRIPTION VECTOR (ADV)

The ADV [2], [3] is a representation of the scene that discretizes the input data into a set of cells where the movement is computed. It has proven good classification capabilities independently of classifiers, in full sequences and in prediction [4]. G-ADV [1] is a variant specified to analyse group behaviour. The G-ADV describes the motion of the group and the individuals with three different components: the trajectory of the group, the coherence of the individual in the group and, the movement relations between the different groups in the scene.

The ADV assumes the input data to be a non-perspective set of images (i.e. images on the ground plane), in some cases a pre-process to correct them is required. Using homography (Eq. 1) can help to rectify the data, assuming that any point p_i in the image is transformed into a point p_g in the ground plane G.

$$p_g = H \cdot p_i \tag{1}$$

The ADV uses the information of pairs of consecutive points to find the ratio of movement in the four directions (up, down, left, right, and frequency). Each cell combines the movement information in the descriptor that will later be used to feed a classifier.

2) D-ADV

Deep-ADV (D-ADV) uses apparent motion of the individuals in the scene, in contrast to the original ADV that uses specific movements, i.e. displacement between frames. This gives a more abstract perception of what is happening. In order to do it, a sequence of images is used, and the dense optical flow is calculated from the sequence.

Specifically, this proposal uses a sequence of images as the input set. Unlike ADV, D-ADV does not rely on the specific, individual movements of a person in the scene and the occurrences in the scene (i.e., frequency). It considers the apparent motion of individuals in the visual scene and the appearance of individuals assuming a specific background. For the former, optic flow calculation is the initial stage of the process. It calculates the optical flow between two consecutive frames $(t, t + \delta t)$ of the sequence using the differential method as the most commonly used method [24]. It is based on the assumption of image brightness constancy: given a video sequence, the pixel intensity (x, y) of frame t, $I_t(x, y)$, remains the same despite small changes in position and time period. If $(\delta x,\delta y,\delta t)$ is expressed as a small change of motion, and assuming constancy of brightness and expansion as a Taylor series, it can be expressed and approximated as described in [24]):

$$I_{t+\delta t}(x + \delta x, y + \delta y) \approx I_t(x, y) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t,$$

solving and dividing the second term along it by δt, it is possible to obtain:

$$\frac{\partial I}{\partial x}\frac{\delta x}{\delta t} + \frac{\partial I}{\partial y}\frac{\delta y}{\delta t} + \frac{\partial I}{\partial t} = \frac{\partial I}{\partial x}U + \frac{\partial I}{\partial y}V + \frac{\partial I}{\partial t} \approx 0$$

where $U = \frac{\delta x}{\delta t}$ y $V = \frac{\delta y}{\delta t}$ are the two components of the optical flow in t.

If we assume the image as a ground plane and a static camera (i.e., apparent motion is only generated by the individuals, not by the observer), the difference between two points could be approximated as the derivatives of the pixels $(p_i - p_{i-1}) \approx (\frac{\delta x}{\delta t}, \frac{\delta y}{\delta t}) = (W, V)$. Based on the ADV, we can relate *Up* (U) and *Down* (D) motion components with vertical W and *Left* (L) and *Right* (R) with V. Hence, the components are calculated as in Eq. 2:

$$\begin{aligned}
U(I_t) &= \begin{cases} -V_t \text{ } if \text{ } V_t < 0 \\ 0 \quad others\ldots \end{cases}\\
D(I_t) &= \begin{cases} V_t \text{ } if \text{ } V_t > 0 \\ 0 \quad others\ldots \end{cases}\\
L(I_t) &= \begin{cases} -W_t \text{ } if \text{ } W_t < 0 \\ 0 \quad others\ldots \end{cases}\\
R(I_t) &= \begin{cases} W_t \text{ } if \text{ } W_t > 0 \\ 0 \quad others\ldots \end{cases}
\end{aligned} \tag{2}$$

The *Frequency* (F) refers to the time (frames) in which there are people in the scene regardless if they are moving or not (Eq. 3:

$$F = |I-B| > 2 \cdot std(I - B), \tag{3}$$

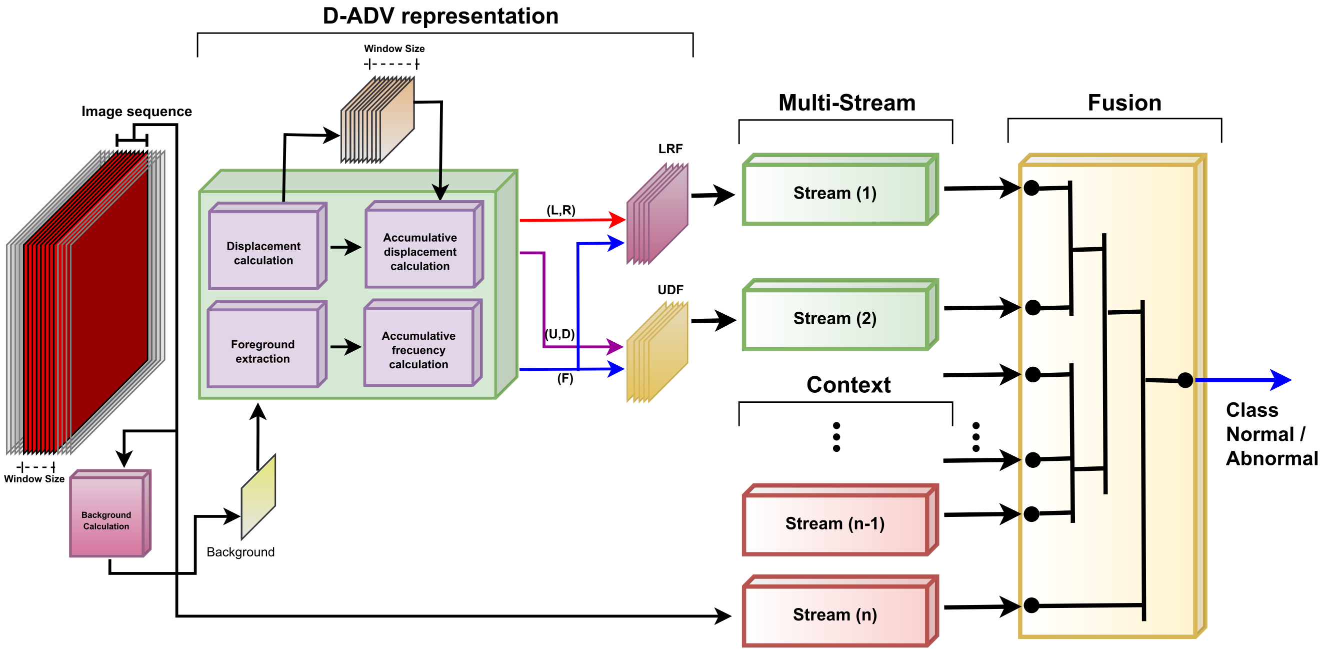


FIGURE 1. Overall proposed architecture. Local movement representation by D-ADV and multi-stream and fusion stages.

where B is the background, i.e. an image of the scene with no people. The std refers to the standard deviation, that in this case models the dispersion of the difference between the pixels of the image and the background. In order to reliably estimate the pixels that refer to people, by knowing the variability in the pixels of the background due to camera noise or other interferences, we can robustly define the segmentation threshold with the standard deviation specifically, setting it to twice the std.

Each component of the five is then accumulated over a sequence of frames determined by the windows size (ws) that depends on the accuracy requirements of the problem. Then, the accumulated U, D, L, R, and F, are passed to the classification stage, that is explained in Sec. III-B. In general, the classification stage is proposed to be a multi-stream network architecture. The input of each stream is not a concatenation of the five images, but one with LRF and the other one with UDF. In this way, one network takes into account horizontal displacements with the frequency, whilst the other vertical motions with the frequency, isolating the directions to allow the networks to focus better on specific directions. Figure 2 depicts this idea of multi-stream with the LRF and UDF inputs.

B. CLASSIFICATION STRATEGY

The proposed architecture defines, for human behaviour analysis, a classification stage that takes local motion as input and provides a class, either in a range of classes or as a normal/abnormal classification. In the literature review, it has been seen that using multiple networks simultaneously for classification, and fusing their outputs to provide a final decision, achieves better results. This is because each behaviour may have specific features that are learnt by a specific network and then scored, rather than trying to extract all the features within a single global network. This idea is here applied, using a multi-stream strategy, that uses the previously calculated UDF and RLF. Also, in some cases, such as in the one-class classification instance of this general architecture (Section IV-B), another stream can be added for a specific purpose.

If prediction scores are assigned to each of the network streams (each stream emits scores from different classification tasks), we can weight the features from different points of view. It is very important to efficiently fuse all the streams to generate the final predictions. They can be fused using different methods such as Late fusion (LF), Early fusion (EF), Hybrid fusion (HF) used in [66]. Moreover, a hierarchical fusing can be done that scores two streams that have a common semantic meaning (motion behaviour in the example), and the result is merged with the context stream.

In this paper, the classification stage is proposed (see Figure 1) to include of two streams, each related to the D-ADV UDF and LRF descriptors respectively. The combination of both streams is carried out by a final fusion layer that concatenate the stream outputs and pass it to a fully connected layer. In addition, the proposed architecture considers, for the case of one-class classification, an extra stream for the context information in the scene. In this case, the fusion of the context stream is done in a weighted way to each of the flows and will depend on the specific application.

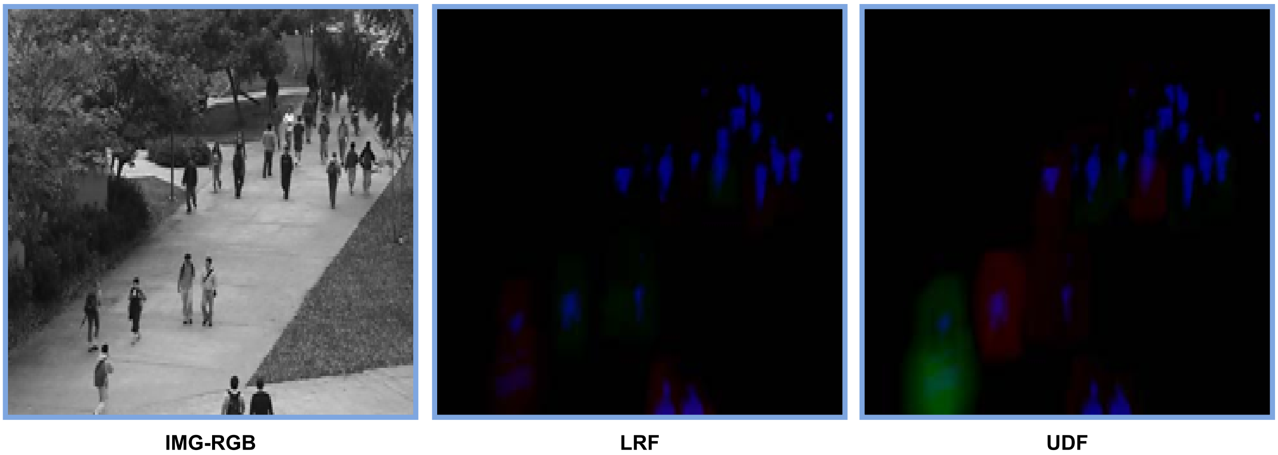


FIGURE 2. Example of ***IMG***, ***LRF***, ***UDF*** images from D-ADV representation of an abnormal activity in Ped1 dataset.

IV. ARCHITECTURE INSTANTIATED

For the development of this proposal, two neural network architectures have been proposed by instantiating the architectural model defined in the previous section: instance with One-Class classification (D-ADV-OC) and instance with Multi-Class classification (D-ADV-MC). These architectures are able to detect specific behaviors that are part of the training data set and abnormal behaviors from training normal data, respectively. Details about the classification and training stages, as well as some important configuration parameters of each of the architecture instances, are explained in the subsections IV-A and IV-B.

A. MULTI-ACTIVITY CLASSIFICATION (D-ADV-MC)

The architecture instantiated from the architectural model proposed in this paper for the classification of multiple group activities can be seen in Figure 3. This architecture uses two-stream activity classification and performs a late fusion as discussed in the previous section capable of classifying the previously computed D-ADV images: *LRF* and *UDF*. The classifier approach is open and allows the use of any CNN architecture (VGG, Resnet, Alexnet, LeNet, etc.). This type of networks typically uses a fully connected layer on the output with a softmax activation function to decide to which class the image corresponds (e.g. objects, locations, poses, etc.). The architecture ignores the individual dense layers. However, the previous layers of the model are concatenated in a late merge with a fusion layer. Finally, we use a fully connected layer with sigmoid activation function to connect the fusion layer and predict multiple classes.

To overcome the challenge of training the architecture with small datasets such as BEHAVE and CAVIAR [7], [18], we perform transfer learning of the trained models with the ImageNet dataset [25]. As a result, we refined the CNN-based network three times. First, we replaced the fully connected layer of the ImageNet architecture with a new one that matches our classes. Second, we trained a subset of the lower layers because the *LRF* and *UDF* inputs are different from the *RGB* input images expected by ImageNet. Finally, we retrained a subset of the upper layers for fine-tuning.

For the training step, we use the binary cross-entropy as an objective function to consider each output class as an independent Bernoulli distribution. For classification, considering that more than one class can be present in a time frame of the sequence, different thresholds ϵ_i are considered for each output neuron i. Here the ϵ_i thresholds are calculated to be the value that maximises the true positive rate (TPR) and minimises the false positive rate (FPR), for each class, C_i.

B. ABNORMAL ACTIVITY CLASSIFICATION (D-ADV-OC)

This section specifies the instance of the proposed architecture for the detection of anomalous activities in a scene. This architecture uses three data streams: the two motion-related streams from the D-ADV with the *LRF* and *UDF* images, and the a third stream associated with the scene context. This architecture has been coined D-ADV-OC, and does not use the individual dense layers as it connects to layers upstream of them. Therefore, the upstream layers in the convnet are concatenated in a late merging manner using a concatenation layer of the two streams. After all, we use a fully connected layer with linear activation to connect the concatenation layer and predict abnormal activity in the cluster. This architecture design is based on recent work by Ruff *et al.* [44] providing a deep model for training a neural network by minimizing the volume of a hypersphere enclosing the data network representations. This proposal differs from the work of Chalapathy *et al.* [10] by combining the ability of CNN-based networks to progressively learn from a subset of images that are the representation of the input data along with the one-class target. Unlike the latter work, which uses auto-encoders to establish the representation of the input data by defining the center of the hypersphere, in this paper some layers of the CNN-based network are trainable, allowing the architecture to continue training

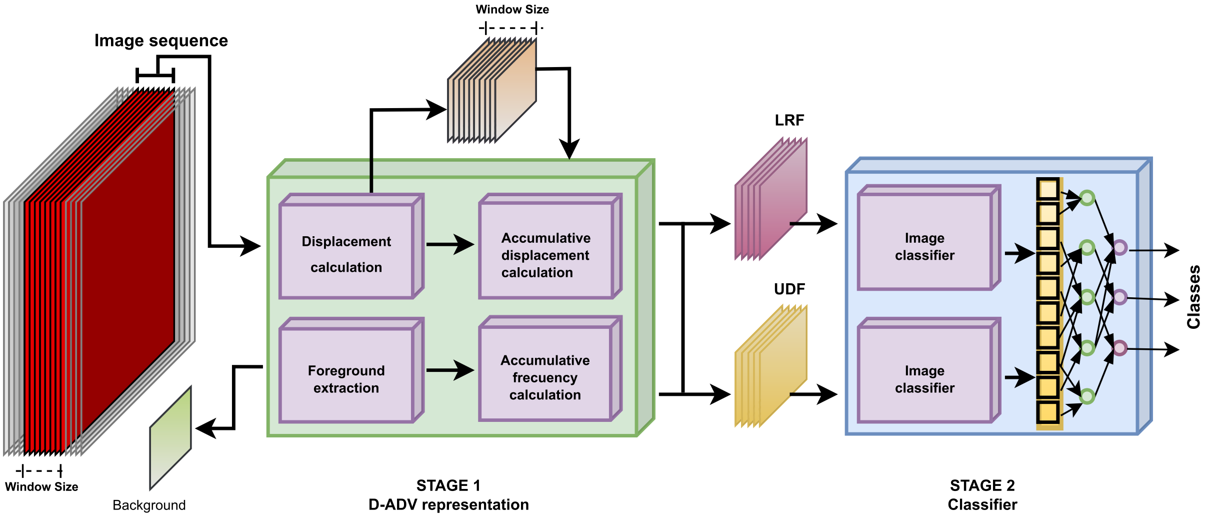


FIGURE 3. Data flow of the proposed D-ADV. The D-ADV architecture is mainly divided into two parts, the D-ADV-MC representation stage where the displacement is calculated using the ADV descriptor from a sequence of images and their optical flow motion. The second stage defines the classifier using CNN classifiers and a fully connected layer to predict the class.

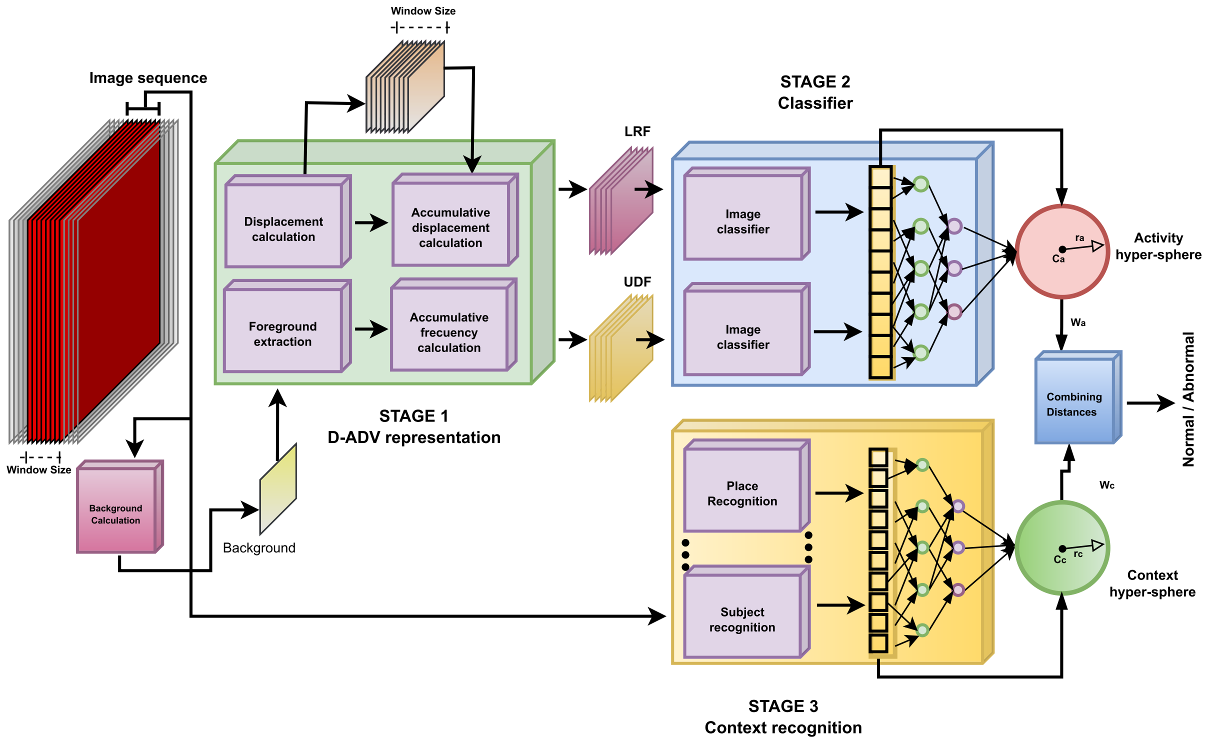


FIGURE 4. The data flow of the D-ADV-OC method. The D-ADV-OC architecture is mainly divided into two parts, the D-ADV-OC representation stage where the displacement is calculated using the ADV descriptor from a sequence of data and its optical flow movement. The second stage defines the classifier using CNN classifiers and a fully connected layer to identify the class whether it is normal or abnormal.

both the center and adjusting the radius of the hypersphere. To avoid the problems of large datasets to train our model and with the goal that it can be used for small datasets, transfer learning of the models trained with ImageNet is used.

The third stream of this architecture is related to the context information in the scene, and at the training step, we calculate the maximum values of the input patterns to normalize the output data, which could be objects, locations, etc. The mean value of the normalization establishes initially the center of the hypersphere but it is jointly the length of the hypersphere radius optimized by means of a fully connected layer at the end of the network.

The distance combination module (see Figure 4) uses the weights w_a and w_c for the activity and context loss functions to train the network and calculates the distance, d, of an input pattern to the normal class according to the prediction stage using the following function:

$$d(i) = \frac{1}{n} w_a \sum_i ||i_a - c_a||^2 + w_c \; ||i_c - c_c||^2 \qquad (4)$$

where i_a is the computed representation for the activities using motion; i_c, the computed representation of the context in the scene; and, finally, c_a and c_c are the centers of the hyperspheres. Additionally, in this model, the combination of the weights w_a for behavior and w_b are taken into account.

For prediction, P, an input pattern i is classified as normal if the distance dist (Eq. 4) is less or equal than a threshold ϵ. Otherwise, the input is considered abnormal.

$$P(i) = \begin{cases} \text{normal} & \textit{if} \;\; d(i) <= \epsilon \\ \text{abnormal} & \text{otherwise}\ldots \end{cases}$$

The threshold ϵ of the length of the hypersphere is calculated at the end of the training stage as the optimal cut point of the ROC curve. It is the distance (Eq. 4) that classifies most of the individuals correctly and thus least of them incorrectly. Using the ROC curves, the cut point is calculated as follows:

$$\epsilon = \underset{d}{\text{argmin}} \, |\text{TPR}(d) - (1 - \text{FPR}(d))|,$$

being TPR and FPR, the true and false positive rates respectively.

V. EXPERIMENTAL RESULTS

In this section, the experimental results for the two proposed instances are presented. First, the Deep-ADV Multi-Class Classification (D-ADV-MC) instance has been validated with the INRIA, CAVIAR and BEHAVE datasets [7], [18]. On the other hand, the Deep-ADV One-Class Classification (D-ADV-OC) has been tested with the Ped 1, Ped 2 [34] and Avenue [29] datasets.

The general experimentation configuration for both instances are as follows:

- For the calculation of the optical flow, Gunnar Farneback's algorithm [17] has been used, which allows a dense calculation.
- Size of the cells that conform the images *LRF* and *UDF* is 224 x 224.
- The CNN based image classifier is the ResNet50.
- The images *LRF* and *UDF* provided by the D-ADV representation stage have been normalized as input of the ResNet50 to the range between 0 and 1 by dividing each cell (pixel) by the maximum value of each component (L,R,U,D and F).

The CNN based image classifier has been trained using transfer learning from the Imagenet dataset in three steps. First, only the last layer has been trained with the specific labels of the corresponding dataset. Second, the first 139 layers of the activity recognition module has been trained as a domain adaptation solution from the RGB to the *LRF* and *UDF* domain. Finally, from the top layer to the layer 249 is finally trained.

A. MULTI-CLASS CLASSIFICATION RESULTS

A 10-fold cross validation has been used to calculate the performance of the architecture for the different datasets. Moreover, 25% of the training data in each fold has been used for the validation set. Specifically, the performance of the D-ADV has been evaluated using the sensitivity, specificity (see Table 1) and the AUC and ROC curves (see Figure 5, Figure 6 and Figure 7). These values have been calculated for frames and for sequences. That is, the performance per frame is calculated according to the prediction made on each individual frame independently of the sequence. The performance per sequence is calculated according to the prediction made for all the frames in the sequence. For this, the prediction of the sequence is the one corresponding to the one corresponding to at least 80% of the frames of the sequence. Finally, in order to evaluate the ability of the representation to synthesize the information extracted from the scene, two different values for the windowsize (*ws*) parameter have been tested: 10 and 40 (i.e. 0.5 sec. and 2 sec.).

The per-frame performance results with a window size (*ws*) of 10 reach according to the sensitivity and specificity, for the INRIA dataset, 71.70% and 84.85% on average, 91.47% and 94.51% for BEHAVE dataset, while 78.18% and 87.12% respectively for the CAVIAR (corridor) dataset. Using a larger window size (i.e. *ws* of 40), the results are improved in the three datasets, obtaining 89.93% and 95.65% (senditivity and specificity) for INRIA, 92.55% and 94.79% for BEHAVE dataset, and 79.00% and 88.88% for the CAVIAR dataset (see 1).

In terms of performance per sequence, D-ADV achieves very high results. Considering a window size of 10, for the INRIA dataset, a total of 91.67% of sensitivity and 95.83% of specificity is achieved, while 95.07% and 95.52% for the BEHAVE dataset, 80.00% and 93,06 respectively for CAVIAR dataset. Again, the results considering a larger value for the window size ($ws = 40$) are improved achieving the best ones. On average, 95.83% of sensitivity and 97.92% of specificity for the INRIA dataset, 95.52% and 95.70% respectively for the BEHAVE dataset and, finally, 80.58% and 94,27 for CAVIAR dataset.

Finally, the D-ADV architecture is compared with the methods proposed in [1], [12], [36], [67], [64] considering the seven classes of the BEHAVE dataset. Only [12] and our previous work (GADV) consider the seven classes as well. The rest of the works use a subset of four classes. Table 2 shows the comparison of the sensitivity results. As we can see, our proposal, D-ADV, achieves in average the best results outperforming all compared methods.

TABLE 1. Comparison of results with datasets (INRIA, BEHAVE, CAVIAR) calculated for frame and sequence with windowsize (*ws*) values 10 and 40.

		Frame				**Sequence**			
		$ws = 10$		$ws = 40$		$ws = 10$		$ws = 40$	
Dataset	**Class**	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Inria	Fighting	82.84%	76.46%	95.48%	94.00%	100.00%	100.00%	100.00%	100.00%
	Leaving	95.87%	94.79%	99.68%	99.74%	87.50%	93.75%	100.00%	100.00%
	Meeting	65.11%	75.10%	86.85%	87.58%	87.50%	93.75%	87.50%	93.75%
	Overall	**71.70%**	**84.85%**	**89.93%**	**95.65%**	**91.67%**	**95.83%**	**95.83%**	**97.92%**
Behave	Approach	90.88%	92.45%	92.02%	92.68%	93.94%	92.08%	93.94%	95.05%
	Split	92.58%	93.23%	95.18%	93.92%	97.14%	95.96%	97.14%	96.97%
	Fight	95.52%	96.35%	93.40%	95.27%	100.00%	98.28%	94.44%	93.10%
	InGroup	94.41%	94.32%	94.15%	93.75%	94.83%	94.74%	93.10%	93.42%
	RunTogheter	99.87%	99.95%	100.00%	99.99%	100.00%	100.00%	100.00%	100.00%
	WalkTogheter	84.12%	88.30%	87.92%	90.82%	92.31%	88.41%	96.92%	94.20%
	Overall	**91.47%**	**94.51%**	**92.55%**	**94.79%**	**95.07%**	**95.52%**	**95.52%**	**95.70%**
Caviar	Meeting	87.59%	86.25%	90.58%	90.47%	88.24%	93.18%	94.12%	100.00%
	Shop-Enter	91.28%	91.14%	85.80%	93.86%	100.00%	95.56%	87.50%	93.33%
	Shop-Exit	81.21%	87.10%	83.63%	85.82%	90.00%	95.12%	90.00%	90.24%
	Walking	72.16%	76.49%	73.57%	75.17%	65.96%	78.57%	68.09%	78.57%
	Overall	**78.18%**	**87.12%**	**79.00%**	**88.88%**	**80.00%**	**93.06%**	**80.00%**	**93.06%**

TABLE 2. Comparison according BEHAVE results.

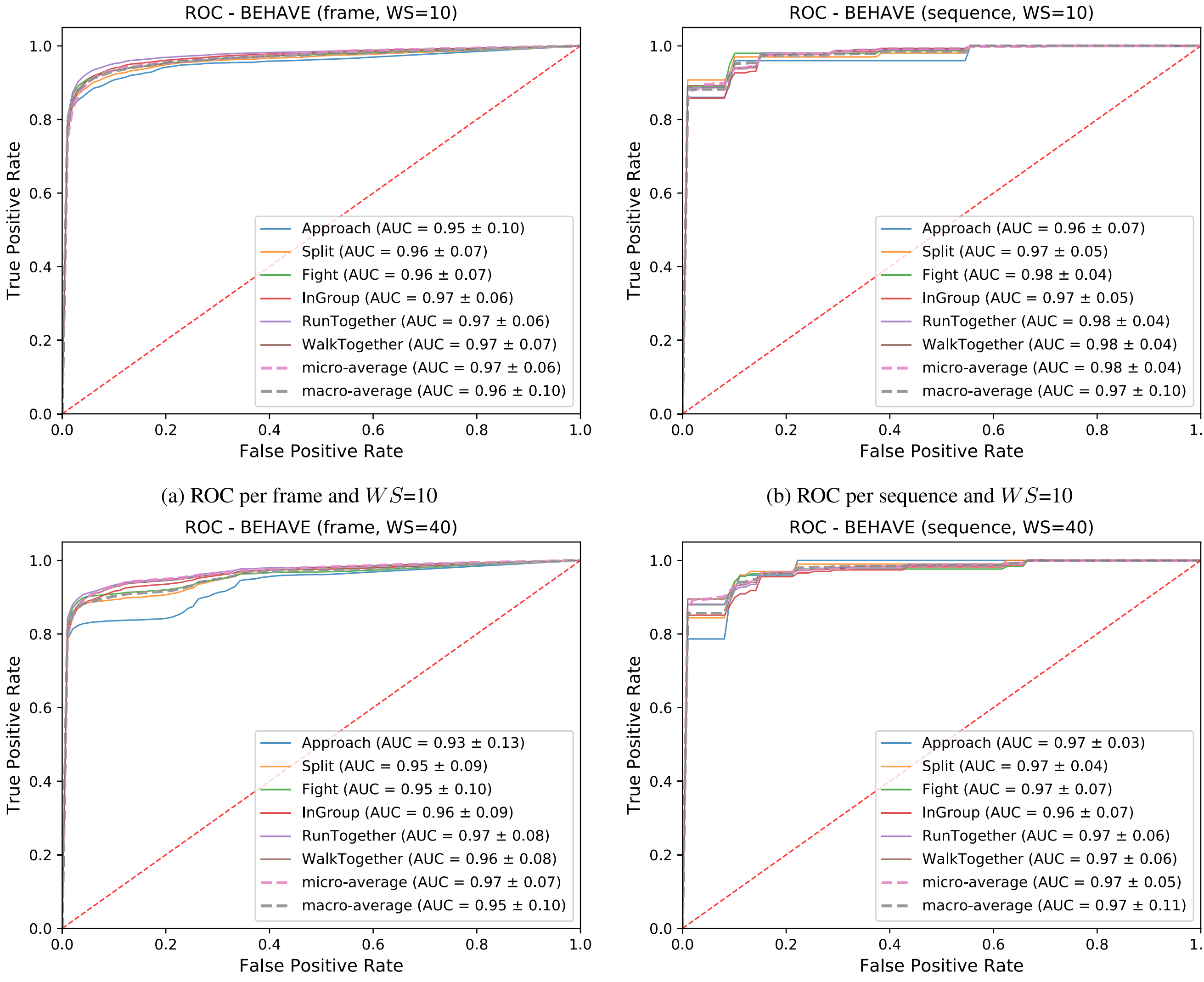
	D-ADV	GADV [1]	[12]	[67]	[36]	[64]
InGroup	93,10%	86,67%	100,00%	88,00%	90,00%	94,30%
Approach	93,94%	100,00%	83,33%	71,00%	60,00%	–
Fight	94,44%	90,00%	83,33%	–	–	95,10%
WalkTogether	96,92%	86,67%	91,66%	88,00%	45,00%	92,10%
Split	97,14%	100,00%	100,00%	79,00%	70,00%	93,10%
RunTogether	100,00%	100,00%	83,33%	–	–	–
Average	**95,93%**	93,89%	90,28%	81,50%	66,25%	93,65% t

TABLE 3. Results of the D-ADV-OC experiments for the PED 1, PED 2 and Avenue datasets. The OCC-SVDD and OCC-NN loss functions, with a combination of D-ADV, D-ADV+Context, D-ADV+CNN and D-ADV+CNN+Context classifiers are used to calculate sensitivity and specificity.

		Dataset							
		Ped1		Ped2		Avenue		All	
Instance	Loss func.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
D-ADV	OC-SVDD	73,84%	73,87%	82,89%	82,87%	75,62%	75,62%	77,45%	77,45%
	OC-NN	74,29%	74,20%	81,80%	81,77%	74,30%	74,29%	76,80%	76,75%
	Average	74,07%	74,04%	82,35%	82,32%	74,96%	74,96%	77,12%	77,10%
D-ADV+Context	OC-SVDD	73,84%	73,87%	**88,53%**	**88,40%**	**77,29%**	**77,30%**	**79,89%**	**79,86%**
	OC-NN	**75,43%**	**75,95%**	83,68%	83,15%	75,57%	75,17%	78,23%	78,09%
	Average	74,64%	74,91%	86,11%	85,78%	76,43%	76,24%	79,06%	78,97%
D-ADV+CNN	OC-SVDD	69,74%	69,73%	71,36%	71,27%	71,63%	71,61%	70,91%	70,87%
	OC-NN	68,90%	68,94%	73,24%	73,20%	73,36%	73,35%	71,83%	71,83%
	Average	69,32%	69,34%	72,30%	72,24%	72,50%	72,48%	71,37%	71,35%
D-ADV+CNN+Context	OC-SVDD	70,31%	70,30%	85,19%	85,64%	71,93%	71,87%	75,81%	75,94%
	OC-NN	69,84%	69,84%	74,27%	74,31%	74,35%	74,34%	72,82%	72,83%
	Average	70,08%	70,07%	79,73%	79,98%	73,14%	73,11%	74,32%	74,38%

B. ONE-CLASS

The previous instance of the architecture to solve the multi-class problem has been tested with small datasets to validate the architecture is able to be trained with a small amount of data. Now, larger datasets have been used to validate the use of context in improving the performance of the model. Here, the D-ADV-OC architecture has been evaluated using different instances. The simplest one is the *D-ADV* instance that uses the original ADV descriptor [4] with 15 x 15 cells followed by a fully connected layer of 15 x 15 x 5 neurons corresponding to the space of the ADV. The *D-ADV+CNN* uses two Resnet50 neural networks to learn the



(a) ROC per frame and WS=10

(b) ROC per sequence and WS=10

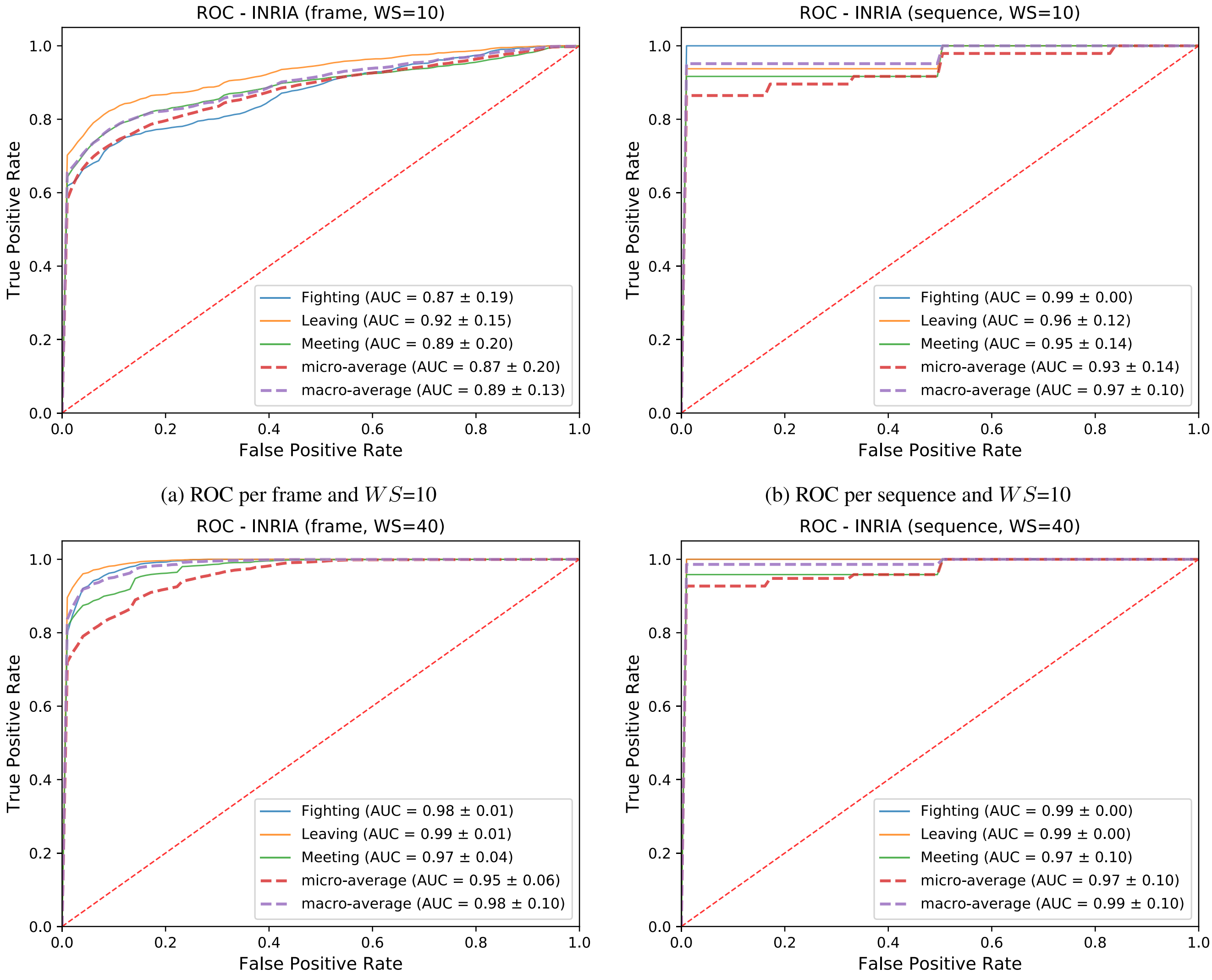
(c) ROC per frame and WS=40

(d) ROC per sequence and WS=40

FIGURE 5. ROC curves for BEHAVE dataset.

LRF and UDF images without the top and the flatten layer connected to two concatenated 2D global average pooling layer (one per image stream) followed by a fully connected layer of 4096 neurons. Moreover, *D-ADV+Context* and *D-ADV+CNN+Context* uses the previous configurations with a third stream: the context recognition. In this case, a YOLO neural network trained with VOC has been used. The weights used for training are $w_a = 0.9$ (activity) and $w_c = 0.1$ (context). For all instances, two loss functions has been used: the OC-SVDD [44] and the OC-NN [10]. The window size (ws) as the number of consecutive frames considered in the accumulative process (see Figure 3) is 5 for the Ped1 and Avenue dataset and 10 for Ped2. Finally, the instances have been tested using the splits of training and tests predefined in the datasets.

The experimental results considering the sensitivity and the specificity as performance measure can be seen in Table 3. As can be seen, high performances are obtained for all combinations in the different datasets. Having an average value in all cases higher than 70% in all cases for both sensitivity and specificity, and reaching values on average close to 80% for the D-ADV-Context instance with OC-SVDD. This configuration is the one that reaches the highest values for the Ped2 and Avenue datasets. Reaching close to 90% in both parameters for Ped2. Although the configuration with OC-NN loss function has a very similar performance, it is in the Ped1 dataset where it achieves the best results. As for the use of the third stream with the Yolo object detection based context, it is shown that in all cases it improves the accuracy rates, avoiding a higher number of false alarms.



(a) ROC per frame and WS=10

(b) ROC per sequence and WS=10

(c) ROC per frame and WS=40

(d) ROC per sequence and WS=40

FIGURE 6. ROC curves for INRIA dataset.

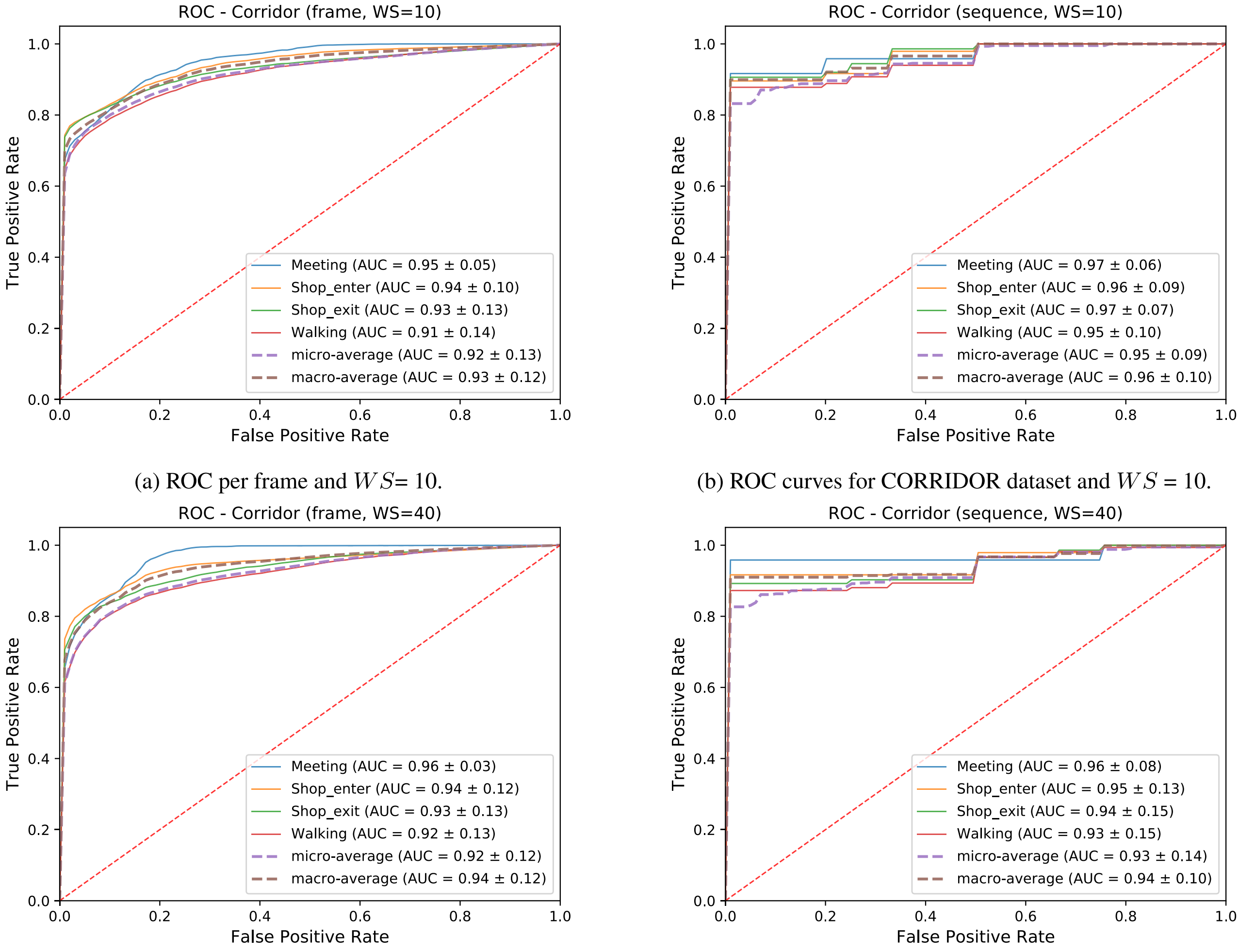
TABLE 4. Average performance results for the considered datasets considering sensitivity, specificity, AUC and EER for the different loss functions and components.

Loss		Components		Performance			
OC-SVDD	**OC-NN**	**with CNN**	**with Context**	**Sensitivity**	**Specificity**	**AUC**	**EER**
✓	✗	✗	✗	77,45%	77,46%	84,57%	22,55%
✔	✖	✖	✔	**79,89%**	**79,86%**	**86,12%**	**20,11%**
✓	✗	✓	✗	70,91%	70,87%	78,27%	29,10%
✓	✗	✓	✓	75,81%	75,94%	82,77%	24,18%
✗	✓	✗	✗	76,79%	76,75%	84,97%	23,22%
✗	✓	✗	✓	78,22%	78,09%	86,23%	21,84%
✗	✓	✓	✗	71,83%	71,83%	79,05%	28,17%
✗	✓	✓	✓	72,82%	72,83%	79,67%	27,18%

1) ABLATION STUDY

In this section, an ablation study on the architecture to investigate the effect of each component of has been carried out.

First of all, the average performance results for the PED1, PED2 and Avenue datasets have been calculated. As can be seen in 4, the architecture trained with a OC-SVDD loss,



(a) ROC per frame and WS= 10.

(b) ROC curves for CORRIDOR dataset and WS = 10.

(c) ROC per frame and $WS = 40$.

(d) ROC per sequence and $WS = 40$.

FIGURE 7. **ROC curves for CORRIDOR dataset for frame and sequence with windosize parameter value of 10 and 40.**

TABLE 5. **Ablation study performance for different losses and components.**

Loss		Components		Performance			
OC-SVDD	OC-NN	with CNN	with Context	Sensitivity	Specificity	AUC	EER
✓				**76,02%**	**76,03%**	**82,93%**	**23,98%**
	✓			74,92%	74,88%	82,48%	25,10%
		✓		72,84%	72,87%	79,94%	27,16%
		✕		**78,09%**	**78,04%**	**85,47%**	**21,93%**
			✓	**76,69%**	**76,68%**	**83,70%**	**23,33%**
			✕	74,25%	74,23%	81,72%	25,76%

without the CNN network, and using the context information has the best results achieving an AUC of 86.12%. It will be the reference architecture for comparison.

a: EFFECTIVENESS OF THE LOSS FUNCTION

The selection of a particular loss function is not significant in view of the results obtained in Table 5. Differences in performance is slightly about 1% per sensitivity, specificity and EER. Even considering AUC, the difference is only 0.45%. The average results for both loss functions is again very similar, obtaining a difference of less than 4% compared to the reference architecture w.r.t. AUC, for example (82.93 % → 86.32% and 82.48 % → 86.32 % for OC-SVDD and OC-NN respectively).

TABLE 6. Average performance details considering the CNN module for the considered datasets.

Datas.	Samples	Perf.	CNN	w/o CNN	Diff.
Ped 1	6800	Sensitivity	69,70%	74,35%	4,65%
		Specificity	69,70%	74,47%	4,77%
		AUC	77,41%	82,47%	5,06%
		EER	30,30%	25,60%	-4,70%
Ped 2	2550	Sensitivity	76,02%	84,22%	8,21%
		Specificity	76,10%	84,05%	7,94%
		AUC	83,18%	91,85%	8,66%
		EER	23,97%	15,81%	-8,16%
Avenue	15328	Sensitivity	72,82%	75,69%	2,88%
		Specificity	72,79%	75,60%	2,81%
		AUC	79,23%	82,11%	2,87%
		EER	27,20%	24,38%	-2,82%

Datas.: Dataset; Samples: training samples; Perf.: Performance variables; Diff.:Difference

b: EFFECTIVENESS OF THE CNN MODULE

This parameter has an important role in the performance. As can be seen in Table 5 differences in performance are above 5% for all parameters, being the AUC 5.53% (85.47 % → 79.94%) less if the CNN network is used. The architecture without the CNN block uses a fully connected layer of 1125 (15 × 15x5) neurons, as we said before, while with the CNN block, two Resnet50 networks that end up in a fully connected layer of 4096 neurons. Both the 1125 or the 4096 neurons are used as input parameters of the hyperspace where the hypersphere that serves to decide normal or abnormal behaviour is calculated. In other words, the number of parameters to learn is much higher in the latter case (including the training neurons of the Resnet50 models). In order to explain it, it is important to see Table 6. The number of frames of the training set for PED2 is the smallest (2550 frames) having a difference in AUC parameter of 8.66% (91.85 % w/o CNN → 83.18 % with it). Considering the Avenue dataset, the number of samples is 6 times higher (15329) what reduces drastically the difference in performance to an AUC of 2.87% (82,11 % w/o CNN → 79,23% with it). Therefore, although the architecture allows training with small datasets, the larger the input, the higher the performance achieved.

c: EFFECTIVENESS OF THE CONTEXT MODULE

The use of more streams in the architecture is affected by this parameter. The experiments have used a Yolo network to identify objects in the scene. Differences in performance are important, achieving about a 2.5% per sensitivity, specificity and EER. Considering AUC, the difference is about 2% (81,72% → 83,70%). Hence, the use of context information improves the overall performance of the architecture.

2) COMPARSION WITH OTHER METHODS

Finally, the experimental results are showed and compared with other state-of-the-art methods in Table 7 at frame-level.

TABLE 7. Comparison of D-ADV-OC results at frame level with other methods with PED 1, PED 2 and Avenue datasets. Cells containing the letters N/A indicate that no data is available.

	Frame level					
	Ped 1		Ped 2		Avenue	
Reference	AUC %	EER %	AUC %	EER %	AUC %	EER %
[53]	82.34	23.50	97.52	4.68	71.54	36.38
[21]	80.90	26.30	95.90	10.50	84.20	23.00
[69]	83.50	25.20	94.90	10.30	86.10	22.00
[28]	83.80	22.30	96.50	8.70	84.50	21.50
[55]	77.80	29.20	96.40	8.90	85.30	23.90
[43]	53.50	48.00	81.40	26.00	73.80	32.80
[26]	82.10	N/A	96.50	N/A	87.20	N/A
[33]	86.26	N/A	96.06	N/A	85.78	N/A
[22]	N/A	N/A	97.80	N/A	90.40	N/A
[38]	N/A	N/A	96.20	N/A	86.90	N/A
[63]	N/A	N/A	96.80	N/A	86.20	N/A
[39]	N/A	N/A	82.80	N/A.	84.30	N/A
Average	**78.78**	**29.08**	**94.07**	**11.51**	**83.85**	**26.60**
D-ADV-OC	**84.41**	**24.36**	**95.04**	**11.49**	**82.29**	**22.70**

Results for the UCSD Ped 1 dataset show that the lowest value of EER is 23.50% provided in the work by Vu *et al.* [53] and the highest AUC value is 86.26% in the work by Lu *et al.* [33]. According to EER for the UCSD Ped2 dataset, the best results (4.68%) are obtained again in the work by Vu *et al.* However in AUC, the maximum value is achieved by Ionescu *et al.* [22] (97.80%). For the Avenue, the best AUC is again for the work by Ionescu *et al.* (90.40%) and the best EER for other work proposed by Li *et al.* [28] (21.50%). Our proposed architecture with the different instances is not the best for any dataset but the performance is in accordance with those obtained in the other works. As can be seen, if we compare our work with the average results obtained by the state-of-the-art works, itimproves in all cases except for Avenue's AUC (our work achieves 82% compared to an average of 83%). The best configurations for AUC and EER are with the *D-ADV+Context* configuration and with a loss function OC-NN for Ped1 and Avenue, and OC-SVDD for Ped2.

VI. CONCLUSION

In this paper a generic deep learning architecture based on the analysis of local movements has been provided which allows the classification of group activities independently of the number of people and activities. The architecture is composed by multi-streams, being the two main streams the deep learning variant of the Activity Description Vector (D-ADV). The D-ADV consists on a transformation of a sequence of images into two sequences of local movements that occur in specific regions of the scene. The other streams correspond to context information information (e.g. location, time, etc.) used to strengthen the classification of activities. The use of the D-ADV as input queues of deep learning classifiers allows to learn from the characteristics of the descriptor, reducing the space of solutions. The proposed architecture has been urged to address multi-class classification (D-ADV-MC) and one-class classification (D-ADV-OC) in a robust way. The starting hypothesis has been validated with the D-ADV-MC experimentation using small datasets in which the number of people in the scene varies improving the results of the state of the art. Regarding the experiments for the D-ADV-OC classification, large datasets have been used to validate the use of context in improving the performance of the model, outperforming the state-of-the-art approaches.

In terms of future lines, it is proposed to include new streams in order to determine which ones can have a greater impact on performance. In addition, in the short term, it is proposed to replace the calculation of the displacement of the D-ADV with a deep learning-based optical flow module to speed up the calculations. Similarly, in the medium term, it is proposed to study the inclusion of the calculation of the D-ADV directly in the network architecture. This would imply the design of a deep network that would allow the calculation of the displacements (based on optical flow), the background and the accumulations over time. In addition, in the long term, we plan to study higher levels of semantics for groups or crowds. In this case, given the difficulty of having a larger time window, the generation of new datasets is proposed. Finally, we plan to transfer the study to other areas such as the study of groups of vehicles, animals, etc.

REFERENCES

[1] J. Azorin-Lopez, M. Saval-Calvo, A. Fuster-Guillo, J. Garcia-Rodriguez, M. Cazorla, and M. T. Signes-Pont, "Group activity description and recognition based on trajectory analysis and neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 1585–1592.

[2] J. Azorín-López, M. Saval-Calvo, A. Fuster-Guilló, and J. García-Rodríguez, "Human behaviour recognition based on trajectory analysis using neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Aug. 2013, pp. 1–7.

[3] J. Azorin-Lopez, M. Saval-Calvo, A. Fuster-Guillo, J. Garcia-Rodriguez, and S. Orts-Escolano, "Self-organizing activity description map to represent and classify human behaviour," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2015, pp. 1–7.

[4] J. Azorin-Lopez, M. Saval-Calvo, A. Fuster-Guillo, and A. Oliver-Albert, "A predictive model for recognizing human behaviour based on trajectory representation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2014, pp. 1494–1501.

[5] A. Bailly, C. Blanc, É. Francis, T. Guillotin, F. Jamal, B. Wakim, and P. Roy, "Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models," *Comput. Methods Programs Biomed.*, vol. 213, Jan. 2022, Art. no. 106504.

[6] M. Bendali-Braham, J. Weber, G. Forestier, L. Idoumghar, and P.-A. Müller, "Recent trends in crowd analysis: A review," *Mach. Learn. Appl.*, vol. 4, Jun. 2021, Art. no. 100023.

[7] S. Blunsden and R. Fisher, "The BEHAVE video dataset: Ground truthed video for multi-person behavior classification," *Ann. BMVA*, vol. 4, nos. 1–12, p. 4, 2010.

[8] L. F. Borja-Borja, M. Saval-Calvo, and J. Azorin-Lopez, "A short review of deep learning methods for understanding group and crowd activities," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[9] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," 2019, *arXiv:1901.03407*.

[10] R. Chalapathy, A. Krishna Menon, and S. Chawla, "Anomaly detection using one-class neural networks," 2018, *arXiv:1802.06360*.

[11] J. Chen, R. D. J. Samuel, and P. Poovendran, "LSTM with bio inspired algorithm for action recognition in sports videos," *Image Vis. Comput.*, vol. 112, Aug. 2021, Art. no. 104214.

[12] N.-G. Cho, Y.-J. Kim, U. Park, J.-S. Park, and S.-W. Lee, "Group activity recognition with group interaction zone based on relative distance between human objects," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 29, no. 5, Aug. 2015, Art. no. 1555007.

[13] Z. Deng, A. Vahdat, H. Hu, and G. Mori, "Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4772–4781.

[14] M. Devanne, P. Papadakis, and S. M. Nguyen, "Recognition of activities of daily living via hierarchical long-short term memory networks," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 3318–3324.

[15] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers Comput. Sci.*, vol. 14, no. 2, pp. 241–258, 2020.

[16] Z. Fan, H. Zhang, Z. Zhang, G. Lu, Y. Zhang, and Y. Wang, "A survey of crowd counting and density estimation based on convolutional neural network," *Neurocomputing*, vol. 472, pp. 224–251, Feb. 2022.

[17] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*, J. Bigun and T. Gustavsson, Eds. Berlin, Germany: Springer, 2003, pp. 363–370.

[18] R. B. Fisher, "The PETS04 surveillance ground-truth data sets," in *Proc. 6th IEEE Int. Workshop Perform. Eval. Tracking Surveill.*, 2004, pp. 1–5.

[19] Y. Hao, Z.-J. Xu, Y. Liu, J. Wang, and J.-L. Fan, "Effective crowd anomaly detection through spatio-temporal texture analysis," *Int. J. Autom. Comput.*, vol. 16, no. 1, pp. 27–39, Feb. 2019.

[20] R. Hinami, T. Mei, and S. Satoh, "Joint detection and recounting of abnormal events by learning deep generic knowledge," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3619–3627.

[21] J. Hu, E. Zhu, S. Wang, X. Liu, X. Guo, and J. Yin, "An efficient and robust unsupervised anomaly detection method using ensemble random projection in surveillance videos," *Sensors*, vol. 19, no. 19, p. 4145, Sep. 2019.

[22] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7842–7851.

[23] W. Jingying, "A survey on crowd counting methods and datasets," in *Advances in Computer, Communication and Computational Sciences* (Advances in Intelligent Systems and Computing), vol. 1158, S. K. Bhatia, S. Tiwari, S. Ruidan, M. C. Trivedi, and K. K. Mishra, Eds. Singapore: Springer, 2021, doi: 10.1007/978-981-15-4409-5_76.

[24] Q. Ke, J. Liu, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Computer vision for human–machine interaction," in *Computer Vision for Assistive Healthcare* (Computer Vision and Pattern Recognition), M. Leo and G. M. Farinella, Eds. New York, NY, USA: Academic, 2018, pp. 127–145.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[26] S. Lee, H. G. Kim, and Y. M. Ro, "STAN: Spatio-temporal adversarial networks for abnormal event detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1323–1327.

[27] X. Li, W. Li, B. Liu, and N. Yu, "Object and patch based anomaly detection and localization in crowded scenes," *Multimedia Tools Appl.*, vol. 78, no. 15, pp. 21375–21390, Aug. 2019.

[28] Y. Li, Y. Cai, J. Liu, S. Lang, and X. Zhang, "Spatio-temporal unity networking for video anomaly detection," *IEEE Access*, vol. 7, pp. 172425–172432, 2019.

[29] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2720–2727.

[30] X. Lu, C. Ma, J. Shen, X. Yang, I. Reid, and M.-H. Yang, "Deep object tracking with shrinkage loss," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2386–2401, May 2022.

[31] X. Lu, W. Wang, J. Shen, D. Crandall, and J. Luo, "Zero-shot video object segmentation with co-attention Siamese networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2228–2242, Apr. 2022.

[32] X. Lu, W. Wang, J. Shen, D. Crandall, and L. Van Gool, "Segmenting objects from relational visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 28, 2021, doi: 10.1109/TPAMI.2021.3115815.

[33] Y. Lu, K. M. Kumar, S. S. Nabavi, and Y. Wang, "Future frame prediction using convolutional VRNN for anomaly detection," in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2019, pp. 1–8.

[34] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1975–1981.

[35] K. Muhammad, A. Ullah, A. S. Imran, M. Sajjad, M. S. Kiran, G. Sannino, and V. H. C. de Albuquerque, "Human action recognition using attention based LSTM network with dilated CNN features," *Future Gener. Comput. Syst.*, vol. 125, pp. 820–830, Dec. 2021.

[36] D. Münch, E. Michaelsen, and M. Arens, "Supporting fuzzy metric temporal logic based situation recognition by mean shift clustering," in *Advances in Artificial Intelligence* (Lecture Notes in Computer Science), vol. 7526, B. Glimm and A. Krüger, Eds. Berlin, Germany: Springer, 2012, doi: 10.1007/978-3-642-33347-7_21.

[37] N. Nayan, S. S. Sahu, and S. Kumar, "Detecting anomalous crowd behavior using correlation analysis of optical flow," *Signal, Image Video Process.*, vol. 13, pp. 1233–1241, Apr. 2019.

[38] T. N. Nguyen and J. Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1273–1283.

[39] T. Nguyen Nguyen and J. Meunier, "Hybrid deep network for anomaly detection," 2019, *arXiv:1908.06347*.

[40] N. A. Rahmad, M. A. As'Ari, N. F. Ghazali, N. Shahar, and N. A. J. Sufri, "A survey of video based action recognition in sports," *Indonesian J. Elect. Eng. Comput. Sci.*, vol. 11, no. 3, pp. 987–993, 2018.

[41] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1577–1581.

[42] M. Ravanbakhsh, E. Sangineto, M. Nabi, and N. Sebe, "Training adversarial discriminators for cross-channel abnormal event detection in crowds," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1896–1904.

[43] M. Ribeiro, A. E. Lazzaretti, and H. S. Lopes, "A study of deep convolutional auto-encoders for anomaly detection in videos," *Pattern Recognit. Lett.*, vol. 105, pp. 13–22, Apr. 2018.

[44] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *Proc. 35th Int. Conf. Mach. Learn.* (Proceedings of Machine Learning Research), vol. 80, D. Jennifer and A. Krause, Eds. Stockholm, Sweden: PMLR, Jul. 2018, pp. 4393–4402.

[45] M. Sabokrou, M. Fathy, Z. Moayed, and R. Klette, "Fast and accurate detection and localization of abnormal behavior in crowded scenes," *Mach. Vis. Appl.*, vol. 28, no. 8, pp. 965–985, Nov. 2017.

[46] D. Shehab and H. Ammar, "Statistical detection of a panic behavior in crowded scenes," *Mach. Vis. Appl.*, vol. 30, no. 5, pp. 919–931, Jul. 2019.

[47] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.

[48] M. Su, C. Zhang, Y. Tong, B. Liang, S. Ma, and J. Wang, "Deep learning in video violence detection," in *Proc. Int. Conf. Comput. Technol. Media Converg. Design (CTMCD)*, Apr. 2021, pp. 268–272.

[49] J. Sun, J. Shao, and C. He, "Abnormal event detection for video surveillance using deep one-class learning," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3633–3647, 2017.

[50] M. Ullah, M. M. Yamin, A. Mohammed, S. D. Khan, H. Ullah, and F. A. Cheikh, "Attention-based LSTM network for action recognition in sports," *Electron. Imag.*, vol. 2021, no. 6, pp. 1–302, 2021.

[51] S. A. Vahora and N. C. Chauhan, "Deep neural network model for group activity recognition using contextual relationship," *Eng. Sci. Technol., Int. J.*, vol. 22, no. 1, pp. 47–54, Feb. 2019.

[52] D. K. Vishwakarma, P. Rawat, and R. Kapoor, "Human activity recognition using Gabor wavelet transform and ridgelet transform," *Proc. Comput. Sci.*, vol. 57, pp. 630–636, Jan. 2015.

[53] H. Vu, T. D. Nguyen, T. Le, W. Luo, and D. Phung, "Robust anomaly detection in videos using multilevel representations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 5216–5223.

[54] M. Wang, B. Ni, and X. Yang, "Recurrent modeling of interaction context for collective activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3048–3056.

[55] S. Wang, Y. Zeng, Q. Liu, C. Zhu, E. Zhu, and J. Yin, "Detecting abnormality without knowing normality: A two-stage approach for unsupervised video abnormal event detection," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 636–644.

[56] C. Wu, J. Zhang, O. Sener, B. Selman, S. Savarese, and A. Saxena, "Watch-n-patch: Unsupervised learning of actions and relations," 2016, *arXiv:1603.03541*.

[57] L.-F. Wu, Q. Wang, M. Jian, Y. Qiao, and B.-X. Zhao, "A comprehensive review of group activity recognition in videos," *Int. J. Autom. Comput.*, vol. 18, no. 3, pp. 334–350, Jun. 2021.

[58] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," 2015, *arXiv:1510.01553*.

[59] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Comput. Vis. Image Understand.*, vol. 156, pp. 117–127, Mar. 2017.

[60] Y. Yang, H. Lv, and N. Chen, "A survey on ensemble learning under the era of deep learning," 2021, *arXiv:2101.08387*.

[61] H. Yao and X. Hu, "A survey of video violence detection," *Cyber-Phys. Syst.*, pp. 1–24, 2021, doi: 10.1080/23335777.2021.1940303.

[62] L. Ye, T. Liu, T. Han, H. Ferdinando, T. Seppänen, and E. Alasaarela, "Campus violence detection based on artificial intelligent interpretation of surveillance video sequences," *Remote Sens.*, vol. 13, no. 4, p. 628, Feb. 2021.

[63] M. Ye, X. Peng, W. Gan, W. Wu, and Y. Qiao, "AnoPCN: Video anomaly detection via deep predictive coding network," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1805–1813.

[64] Y. Yin, G. Yang, and H. Man, "Small human group detection and event representation based on cognitive semantics," in *Proc. IEEE 7th Int. Conf. Semantic Comput.*, Sep. 2013, pp. 64–69.

[65] Y. Ying, Z. Huilin, Q. Jin, P. Cheng, and M. Duoqian, "Survey on deep learning based crowd counting," *J. Comput. Res. Develop.*, vol. 58, no. 12, p. 2724, 2021.

[66] C. Zalluhoglu and N. Ikizler-Cinbis, "Region based multi-stream convolutional neural networks for collective activity recognition," *J. Vis. Commun. Image Represent.*, vol. 60, pp. 170–179, Apr. 2019.

[67] C. Zhang, X. Yang, W. Lin, and J. Zhu, "Recognizing human group behaviors with multi-group causalities," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, vol. 3, Dec. 2012, pp. 44–48.

[68] X. Zhang, D. Lin, J. Zheng, X. Tang, Y. Fang, and H. Yu, "Detection of salient crowd motion based on repulsive force network and direction entropy," *Entropy*, vol. 21, no. 6, p. 608, Jun. 2019.

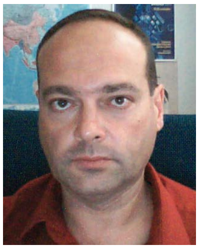
[69] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, "AnomalyNet: An anomaly detection network for video surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2537–2550, Oct. 2019.

[70] N. Zhuang, T. Yusufu, J. Ye, and K. A. Hua, "Group activity recognition with differential recurrent convolutional neural networks," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 526–531.

LUIS FELIPE BORJA-BORJA received the degree in computer engineering from Central University, Ecuador, in 2000, and the Ph.D. degree in computer science from the University of Alicante, Spain, in 2020. Since 2011, he has been a Faculty Member at the Faculty of Engineering and Applied Sciences, Universidad Central, where he is currently a Professor and the Director of the Computer Science degree. In addition, he worked at the Graduate Institute of the Faculty of Engineering and Applied Sciences as a Tutor of some theses. His current research interests include computer vision, computational intelligence, machine learning, deep learning, and the analysis of human activity. In these lines of research, he worked on a number of articles.

JORGE AZORÍN-LÓPEZ received the degree in computer engineering and the Ph.D. degree in computer science from the University of Alicante, Spain, in 2001 and 2007, respectively. Since 2001, he has been a Faculty Member of the Department of Computer Technology, University of Alicante, where he is currently an Associate Professor and the Academic Secretary. He has authored more than 100 contributions in several journals, conferences, and book chapters. His current research interests include 3D computer vision, computational intelligence, machine learning, deep learning, ambient intelligence, human activity analysis, and visual inspection. In these lines of research, he worked in 20 research projects (five of them as a co-ordinator) funded by national, regional, and local public and private entities.

MARCELO SAVAL-CALVO received the degree and master's degrees in computer engineering from the University of Alicante, in 2010 and 2011, respectively, and the Ph.D. degree in computer technology from the University of Alicante, in 2015, funded with a Public Grant from the Regional Government of Valencian Community. He is currently an Associate Professor at the University of Alicante. He has a collaboration with the University of Edinburgh that started in 2014 as part of his thesis research, later as a Postdoctoral Researcher for a year, and has continued until the present. He has participated in six research projects funded in competitive calls, being principal investigator of two of them. He has several publications in journals, most of them indexed in the JCR; three book chapters; and 18 prestigious international conferences. He has supervised several undergrad and master's thesis in both the University of Alicante and the University of Edinburgh, and has been the Ph.D. co-supervisor of a student under an agreement between the University of Alicante and Central de Ecuador. He has been a reviewer for several international journals and conferences, and has been a member of the board of examiners for Ph.D., undergraduate, and master's projects. His research interests include 3D registration and reconstruction of deformable elements, sensorization for autonomous vehicles, and human behavior analysis using artificial intelligence techniques.



ANDRÉS FUSTER-GUILLÓ received the B.S. degree in computer science engineering from the Polytechnic University of Valencia, Spain, in 1995, and the Ph.D. degree in computer science from the University of Alicante, Spain, in 2003. Since 1997, he has been a member of the Faculty of the Department of Computer Science and Technology, University of Alicante, where he is currently an Associate Professor. He was a Deputy Co-ordinator of the Polytechnic School and the Director of the Secretariat for Information Technology, University of Alicante. During this period, he has coordinated and participated in several strategic technological projects, including Open University (transparency portal and open data), UACloud, and Smart University, among others. He has published over 80 articles in different areas of research, including computer vision, 3D vision, machine learning, artificial neural networks, and open data.

MARC SEBBAN received the Ph.D. degree in machine learning from the Universite´ of Lyon 1, in 1996. After four years spent at French West Indies and Guyana University as an Assistant Professor, he got a position of a Professor at the University of Saint-Etienne, France, in 2002. Since 2010, he has been the Head of the Machine Learning Group and the Director of the Hubert Curien Laboratory, Computer Science, Cryptography and Imaging Department. He is currently the Deputy Director of the UMR CNRS, Hubert Curien Laboratory. His research interests include statistical learning theory, metric learning, representation learning, transfer learning, optimal transport, and theory of boosting and learning from highly imbalanced data.