Deep Learning Architecture for Group Activity Recognition using Description of Local Motions*

Luis Felipe Borja-Borja
Fac. de Ing., C. Físicas y Matemática.
Universidad Central del Ecuador
Quito, Ecuador
lborja@uce.edu.ec

Jorge Azorin-Lopez
Computer Technology Dep.
University of Alicante
Alicante, Spain
jazorin@ua.es

Marcelo Saval-Calvo
Computer Technology Dep.
University of Alicante
Alicante, Spain
m.saval@ua.es

Andres Fuster-Guillo
Computer Technology Dep.
University of Alicante
Alicante, Spain
fuster@dtic.ua.es

***Abstract*—Nowadays, the recognition of group activities is a significant problem, specially in video surveillance. It is increasingly important to have vision architectures that automatically allow timely recognition of group activities and predictions about them in order to make decisions. This paper proposes a computer vision architecture able to learn and recognise group activities using the movements of it in the scene. It is based on the Activity Description Vector (ADV), a descriptor able to represent the trajectory information of an image sequence as a collection of the local movements that occur in specific regions of the scene. The proposal evolves this descriptor towards the generation of images able to be the input queue of a two-stream convolutional neural network capable of robustly classifying group activities. Hence, this proposal, besides the use of trajectory analysis that allows a simple high level understanding of complex groups activities, takes advantage of the deep learning characteristics providing a robust architecture for multi-class recognition. The architecture has been evaluated and compared to other approaches using BEHAVE and INRIA dataset sequences obtaining great performance in the recognition of group activities.**

***Index Terms*—D-ADV, group activity recognition, deep learning, convolutional neural networks, video surveillance.**

I. Introduction

The analysis of human behaviour has taken a lot of effort in the area of artificial intelligence and still a very significant problem. Numerous applications are related to this problem, highlighting Video Surveillance [23] and Ambient-Assisted Living [14]. There, computer vision techniques along with machine learning methods were used to cope with the different aspects of the problem. Currently, most real cases involve multiple individuals conforming a group or even a crowd in the scene, and it is of high relevance to study the behaviour in this situations [12]. Regardless the level of application, the traditional pipeline of a machine learning method able to deal with the analysis of human behaviour is divided into two main stages: feature extraction and data analysis (each of these stages can be sub-divided into parts).

The feature extraction stage is usually the fundamental key of the method as it is close related to the performance of it. Moreover, the design process sometimes requires domain knowledge about the problem to be solved. Hence, this stage is heterogeneous and very different for the proposed methods in the literature due to it is the main contribution. Hence, it includes from pre-processing techniques, which includes cleaning, aligning images if we have multiple sources, etc., to the proper methods able to describe the expected output class. It includes methods to detect and track the Region-of-Interest (ROI) by segmenting the images and estimating the motion of it in the sequence. Image segmentation has been widely tackled using traditional techniques [39], [49] and, recently, Deep Learning (DL) approaches [21]. After the ROI is segmented, several works track it to estimate the motion in the scene. Ojha et al. reviewed some tracking technique in [40], and more recently Yazdi and Bouwmans presented in [55] a review of new tracking methods including deep learning approaches.

Regarding the data analysis, supervised and unsupervised learning are the two main approaches. They are fed with the descriptors calculated previously (trajectories, descriptors, ROI blob, etc.) in order to estimate the predefined behaviour or action that are performed in the group (supervised) or to help finding previously unknown patterns in the data set without preexisting classes. Nowadays, the emergence of the deep learning approaches has dramatically improved the state-of-the-art providing a significant improvement in machine vision problems. Deep learning increases the number of hidden layers in neural networks and potential layer-to-layer transformations, allowing multiple levels of abstraction and learning complex functions [31].Deep learning approaches tend to perform the whole pipeline in one single network architecture, where the raw images are fed and the result is the actual behaviour occurred [2], [29]. Traditional and DL-based proposals are reviewed [9], [15], [46].

Despite the large effort made by the science community to improve the human behaviour analysis, there is still room for improvement, mainly in multi-class classification where various activities are considered [30]. Moreover, the lack of generality in current proposals, in terms of number of individuals in the scene (i.e, from group of two people to crowds), makes it difficult establish a reference architecture to define how to approach different cases using similar proposals. This motivates us to define a deep learning solution, as proved to outperform classic machine learning ones, to, regardless the

*This work was supported by the Spanish State Research Agency (AEI) and the European Regional Development Fund (FEDER) under project TIN2017-89069-R

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

number of individuals, classify the behaviour using motion information. It has been demonstrated that using trajectory descriptors improves the quality of the actual behaviour estimation as it provides a simple high level of understanding of complex group activities. The Activity Description Vector (ADV) [4], [6] showed a very good performance, regardless the use of different classifiers, in the description of activities related to individuals. It also, demonstrated its predictive capabilities not only the capabilities of it to papers behaviour from new inputs but also to detect behaviour using a portion of the input, to early detect the behaviour performed by a person in a scene [5], [8]. Finally, an ADV variant was also specified to analyse group behaviour (GADV) in [3] showing also excellent results. The GADV is calculated from the trajectory described by the group and by the individuals who form it. Specifically, it uses three different components: the trajectory followed by the group, the coherence of the individual trajectories in the group and, finally, the movement relationships among different groups in the scene.

Hence, the main objective of this work is to combine the advantages of the ADV to represent activities based on the trajectories of the subjects, and the deep learning approach by introducing the motion description as an ADV variant in a Convolutional Neural Network architecture. From this objective, the contribution of the paper is the improvement of the generality and the performance in multi-class classification of group activities. Moreover, the use of the ADV variant allows to train the model using small sets of labelled data as opposed to using large volumes of data for training it. Instead of learning from raw image sequences, the use of the ADV variant allows the network to learn features from that descriptor, reducing the solution space.

The remaining of the paper is structured as follows: Section II where a state-of-the-art review is presented; Section III introduces the Deep ADV (D-ADV) proposal with a detailed explanation of the different components in the architecture for group action recognition; Section IV shows a set of experiments that prove the performance of the proposal; and finally, Section V concludes the paper summarizing the main contributions and achievements, as well as future works.

II. Related works

Human Behaviour Analysis (HBA), also known as Human Behaviour Recognition, consists in detect the action/activity/behaviour of people using Artificial Intelligence (AI) techniques. Different approaches tackle this problem from different perspectives, either on a low level of understanding as a single action (e.g. move a hand), or on more complicated behaviour (e.g. shopping) [12], [14]. On the other hand, the interest could be focused on a single person or groups and crowds [18].

Human behaviour analysis was initially studied with traditional Machine Learning techniques, and in the last years, Deep Learning (DL) solutions have shown an improvement in the accuracy of the results. Classic HBA proposals used several of the well-known AI methods, such as, Markov models (HMMs) [47], SVM [48] and AdaBoost [37] classifiers, Intrusion Detection System (IDS) [54], probabilistic classification methods [36], [44] and stochastic sampling [41], [50] , shape model analysis from 2D and 3D data [22], Human computer Interactions [1], [42], [42], behaviour semantic classification [1], [52], [57]. Also, methods such as Self-Organizing Map (SOM) [28], Supervised Self-Organizing Map (SSOM) [43], Neural GAS (NGAS) [35],Linear Discriminant Analysis (LDA) [10], Markov Chain Monte Carlo (MCMC) [32], Gaussian Mixture Model (GMM) [32],Histogram of Optical Flow (HOF) [24] have been used.

However, all those methods needed a pre-processing of the images to segment and track the person or people, and after to feed the classifier with that information. This made the classification highly dependant on the quality of each of the steps (segmentation, tracking, classifier). Using a trajectory descriptors has been proved to improve the results as it minimizes tracking noise and provides a homogeneous representation of the motion in the scene. Activity Description Vector (ADV) [4], for instance, was evaluated with various ML methods outperforming the previous approaches. Moreover, it was extended to Group HBA (GADV) [3] with similar improvement, and tested in prediction [5], [8] problems.

In computer vision research, deep neural networks have evolved to be used consistently because of their good results. Deep learning methods have gained superiority over others in the field of image recognition and classification in single images and sequences, as LSTM-based action recognition [19], [33], [51], [51], multi-streams based architectures to behaviour recognition [25], [26], [53],skeleton-based to human behaviour recognition [17], [34], [45].

III. Learning architecture for group activity recognition

The Activity Description Vector (ADV) [4] proposal consists of a representation method that takes as a reference the scene or terrain where a person moves as a basic geometric model to describe his or her trajectory. Using the ADV descriptor to extract movement characteristics, and based on it to recognize human behaviour, some extensions have been provided [6], [7]. In this paper, the deep variant of the ADV, coined D-ADV, is presented. The main stages of the pipeline are presented in Fig. 1. The D-ADV is able to classify several classes from a sequence of images. It can be divided into two parts. First, two images called LRF and UDF are calculated. After that, a CNN based classifier with two streams is used to determine the classes presented in the image sequence.

A. Activity Description Vector

The first stage of the pipeline calculates a representation of the image sequence as a deep variant of the original Activity Description Vector (ADV). It is a trajectory-based feature initially presented in ADV [4] for representing trajectory data with classification purposes. For the sake of completeness, a brief summary of the ADV is shown but we refer you to [4], [6] for further details about its calculation.

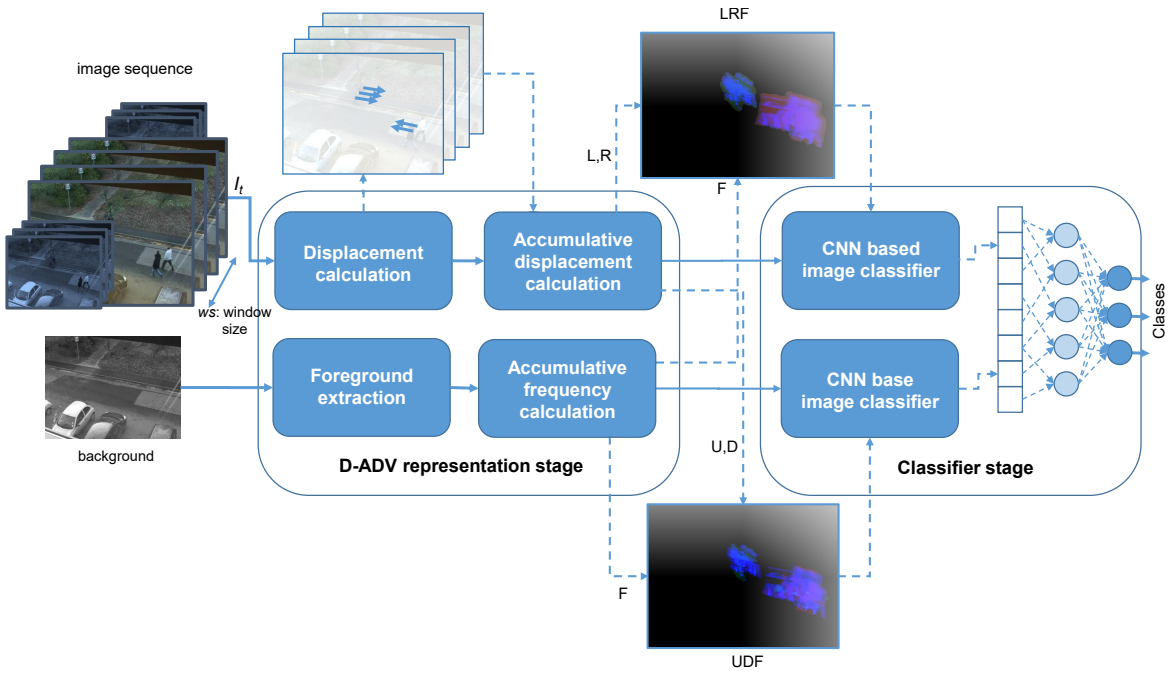


Fig. 1: Pipeline of the D-ADV method. The D-ADV architecture is mainly divided into two parts, D-ADV representation stage were the displacement is calculated using ADV descriptor from a sequence of data and its optical flow movement. The second stage defines the classifier using two ResNet50 classifiers, one for each image (LRF and UDF), and a fully connected layer using late fusion to decide the class.

ADV uses the number of occurrences of a person in a specific point of the scenario and its local movements in it. This method tessellates the ground scenario, G, in cellular regions as a grid, C, to discretize the environment. It is important to mention that, in order to have a more accurate result, G should be flatten using, for instance, homography. Each cell of the grid has information of the movements in the region including up (U), down (D), left (L), right (R) and frequency (F) data. The four former values are extracted from the single displacement between two consecutive points. If we focus on the U movement, it is calculated as follows:

$$U(p_i) = \begin{cases} (p_i - p_{i-1}) \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} & if \frac{(p_i - p_{i-1}) \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix}}{\|(p_i - p_{i-1})\|} > 0 \\ 0 & otherwise \end{cases} \quad (1)$$

where p_i and p_{i-1} are two consecutive locations of the trajectory of an individual in G, and knowing that U is assumed to be a displacement in the positive vertical y axis. This formula is similar in the other three displacements. On the other hand, frequency, F, is estimated as the number of occurrences of a person that is in a specific point.

Finally, the ground plane G is spatially sampled in a matrix C of $m \times n$ cells, so that the transformed points p_g and the functions of frequency and movements of it are in one of the cells of the matrix C. Each cell will describe the activity happened in that region of the scene considering the vector of relevant values, called *Activity Description Vector* (ADV_C). This vector will be composed by the frequency and the U, D, L and R movements of all points of the ground plane inside a cell:

$$ADV_C = \langle F, U, D, L, R \rangle \quad (2)$$

Therefore, within a particular cell, the accumulative histograms of the movements U, D, L, R and F for the points on G of the cell $C_{i,j}$ of C are calculated. Let $u \times v$ the actual size of the scenario, split in $m \times n$ cells, and $p_{k,l}$ the point located in the position k and l of the G space, each ADV in a cell is:

$$\begin{gathered} \forall c_{i,j} \in C \wedge \forall p_{k,l} \in G / i = \left\lfloor \frac{kxm}{u} \right\rfloor \wedge j = \left\lfloor \frac{kxn}{v} \right\rfloor \\ ADV_{i,j} = \begin{pmatrix} \sum F(p_{k,l}), \sum U(p_{k,l}), \sum D(p_{k,l}), \\ \sum L(p_{k,l}), \sum R(p_{k,l}) \end{pmatrix} \end{gathered} \quad (3)$$

With this feature, the trajectory is described by dividing the scene into regions and compressing the data in cumulative values. It is interesting to highlight that Activity Description Vector integrates the trajectory information without length and sequential constraints.

B. D-ADV: activity descriptor for deep learning purposes

The D-ADV uses a sequence of images as input. In contrast to ADV, the D-ADV is not based on the specific and individual

movements of a person in the scene and the occurrences in it (i.e. Frequency). It considers the apparent motion of the subjects in the visual scene and the appearance of them assuming a specific background. For the former, the optical flow calculation is the starting stage of the process. It calculates the optical flow between two consecutive frames $(t, t+\delta t)$ of the sequence by using the differential method as the most widely used method [27]. It is based on the assumption of image brightness constancy: given a video sequence, the intensity of the pixel (x, y) of the frame t, $I_t(x, y)$, remains the same despite small changes of position and time period. Let (δx,δy,δt) the small change of the movement, and assuming the brightness constancy and expanding as Taylor series, it can be expressed and approximated as (more details can be found in [13], [27]):

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

, solving and dividing the second term throughout 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 $U = \frac{\delta x}{\delta t}$ and $V = \frac{\delta y}{\delta t}$ are the two components of the optical flow in t.

In this case, the points p_i used to calculate the components of ADV as in Eq. 1 for the component Up (U) were those extracted from consecutive points in a trajectory on a plane. If we assume the image as a plane of the ground and a static camera (i.e. the apparent motion is only generated by the subjects in the scene, not for the observer – camera), the difference in the trajectory $(p_i - p_{i-1})$ could be approximated as the derivatives of pixels in x and y for the frame t as $(p_i - p_{i-1}) \approx (\frac{\delta x}{\delta t}, \frac{\delta y}{\delta t}) = (U, V)$. Moreover, as the movements are considered in each axis, the movements U and D are closely related to V component of the optical flow, and the components L and R related to the U. In consequence, the components could be calculated as:

$$\begin{aligned} U(I_t) &= \begin{cases} -V_t & if\ \ V_t < 0 \\ 0 & otherwise \end{cases} \\ D(I_t) &= \begin{cases} V_t & if\ \ V_t > 0 \\ 0 & otherwise \end{cases} \end{aligned} \tag{4}$$

With respect to the component F, it is estimated as:

$$F = |I - B| > 2 * std(I - B) \tag{5}$$

, where B is the background calculated from a sequence of images, and std is the standard deviation of the difference between a frame and the background. The foreground is extracted in order to obtain the subjects that appear in the scene independently if the are moving.

This accumulative stage is responsible for calculating the ADV in a cell as presented in Eq. 2. On the one hand, accumulative displacement is responsible for the L, R, U and D parameters and the accumulative foreground is for the F component. The accumulation is considered for a set of consecutive frames of size, ws (see Fig. 1). In this case, the components are not concatenated all together, they are separated conforming two images LRF compose by the components L, R and F, and, similarly UDF combines the U, D and F components. Figure 1 shows an example of the UDF and LRF images where the accumulated data is shown in cyan and magenta.

C. Two-stream image classifier based on CNN

The last stage of the proposal is the image classifier based on deep neural networks. The proposed architecture of the D-ADV for multi-class problem considers a two-stream Convolutional Neural Network able to classify the previously calculated single images: LRF and UDF. The proposal for the CNN based classifier is open and any architecture could be used (VGG, ResNet, AlexNet, LeNet, etc.). This kind of networks usually uses a fully connected layer at the output with softmax activation in order to decide the class to which the image corresponds (e.g. objects, places, poses, etc.). The D-ADV architecture does not take into account the individual dense layers. However, the previous layers in the convnet are concatenated into a late fusion in a concatenation layer. Finally, a fully connected layer with sigmoid activation is used to connect the concatenation layer to predict multiple classes.

In order to avoid the problems of large datasets to train our model and with the objective that the model could be used for small datasets, we propose transfer learning from models trained with ImageNet. In consequence, the CNN based network is fine tuned three times. First of all, the fully connected layer of the ImageNet architecture is replaced by a new one that is fine tuned with the new classes. After that, a subset of the bottom layers are trained because of the input of LRF and UDF are different to the RGB images of ImageNet. Finally, a subset of the top layers is fine tuned again.

The learning phase uses binary cross entropy as the loss function in order to consider each output class as an independent Bernoulli distribution. Regarding the classification phase, and taking into account that more than one class could be presented in a frame of the sequence, different thresholds ϵ are considered for each output neuron. They are calculated as the value that maximizes the true positive rate tpr and minimize the false positive rate, fpr, for each class, C_i.

IV. Experiments

A. Experimental setup

The experiments have been carried out using different datasets in order to evaluate the capabilities of the proposal and its generality.

1) Datasets: We evaluate the effectiveness of our proposed architecture on two benchmark datasets, including the BEHAVE [11], a dataset with several annotated video sequences of two views of various scenarios with groups of people acting out various interactions. Concretely, the used classes are

Approach, Split, Fight, InGroup, RunTogether, WalkTogether. And the INRIA dataset, part of the CAVIAR project [20] that has images of people/groups meeting, walking together and splitting up; and Two people fighting scenarios. The specific classes used in this work are *Fighting, Leaving, Meeting*

2) Architecture parameters: The tests performed to the previous datasets use the same parameters except for the window size as the number of consecutive frames considered in the accumulative process (see Fig. 1). The size of the cells that conform the images LRF and UDF is 224 x 224. The CNN based image classifier is the ResNet50. Finally, the fine tuning has been performed to the 139 bottom layers at the first step and, finally, from the top to the layer 249.

B. Results and discussion

Experiments have been performed for two different window sizes (10 and 40) in order to evaluate the ability of the representation to synthesize the information extracted from the scene. Additionally, the images LRF and UDF have been normalized to the range (0,1) dividing each pixel (cell) by the maximum value for each component. In order to obtain results that can be generalized to an independent dataset, a 10-fold cross validation has been performed. For the train folds, the 25% of the data has been used for the validation set. Sensitivity, Specificity (see Table I), AUC and ROC curves (see Fig. 2 and Fig. 3) have been calculated in order to analyse the performance of the D-ADV for frames and for sequences.

The performance results obtained by frames with a window size (ws) of 10 achieve, for the INRIA dataset, a 71,70% of Sensitivity and 84,85% in Specifity as average, whereas for the BEHAVE dataset, a total of 91,47% of Sensitivity and 94,51% in Specifity. Using a window size larger, 40 in this case, the results improve in both datasets. We obtain in total 89,93% of Sensitivity and 95,65% Specifity for INRIA and 92,55% of Sensitivity and 94,79% Specifity for BEHAVE.

Regarding the performance by sequence, the D-ADV obtains high results. Considering a window size of 10, for the INRIA dataset, a total of 91,67% of Sensitivity and 95,83% Specifity is achieved. Moreover, D-ADV obtains a total of 95,07% of Sensitivity and 95,52% Specifity for the BEHAVE. Again, the results considering a value of 40 for the window size improve achieving the best ones. In average, 95,83% of Sensitivity and 97,92% Specifity for the INRIA dataset and 95,52% of Sensitivity and 95,70% Specifity for BEHAVE.

Finally, we compare our D-ADV descriptor with the methods proposed in [3], [16], [58], [38], [56] considering the seven classes of the BEHAVE dataset. Only [16] and our previous work (GADV) consider the seven classes as well. The rest of the works use a subset of four classes. Table II shows the a comparison of the Sensitivity results. As we can see, our proposal, D-ADV, achieves in average the best results outperforming all compared methods.

V. Conclusions

In this paper a novel group activity recognition method based on trajectory descriptor and deep learning, D-ADV, has been proposed. The trajectory descriptor is a variant of the Activity Description Vector proposed in previous works serving as input of a CNN architecture. The variant considers any motion in the image instead of making use of specific trajectories of individual or the group providing generality at the input, allowing its use in many different situations and scenes. The apparent motion is calculated by optical flow, transformed and accumulated in cells spatially distributed according to the input image of the sequence. It allows us to generate two images containing the description of the motion and the occurrence of subjects in the scene. The CNN architecture is fed by the previous images using two streams and using late fusion with a dense layer. In this paper, transfer learning has been used.

Experiments have been carried out using the BEHAVE and INRIA datasets. The experimental results show the capacity of the architecture to classify the activities of the groups presented in the datasets. Moreover, it is shown that the architecture is able to have good results using small datasets due to the use of the representation as the input allow to the network to develop a hierarchy of higher understanding concepts from simpler ones. In this case, not from the image but from the motion representation.

As future lines, we propose the use of other CNN based classifier. We are currently exploring the feasibility of this architecture to represent and analyse abnormal behaviour moving from a multi-class to a one-class problem.

References

[1] Jake K Aggarwal and Michael S Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011.

[2] Roberto Arroyo, J Javier Yebes, Luis M Bergasa, Iván G Daza, and Javier Almazán. Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls. *Expert systems with Applications*, 42(21):7991–8005, 2015.

[3] 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 *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 1585–1592, July 2016.

[4] Jorge Azorín-López, Marcelo Saval-Calvo, Andrés Fuster-Guilló, and José García-Rodríguez. Human behaviour recognition based on trajectory analysis using neural networks. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2013.

[5] Jorge Azorin-Lopez, Marcelo Saval-Calvo, Andres Fuster-Guillo, and Jose Garcia-Rodriguez. A novel prediction method for early recognition of global human behaviour in image sequences. *Neural Processing Letters*, 43(2):363–387, 2016.

[6] Jorge Azorin-Lopez, Marcelo Saval-Calvo, Andres Fuster-Guillo, Jose Garcia-Rodriguez, and Sergio Orts-Escolano. Self-organizing activity description map to represent and classify human behaviour. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2015.

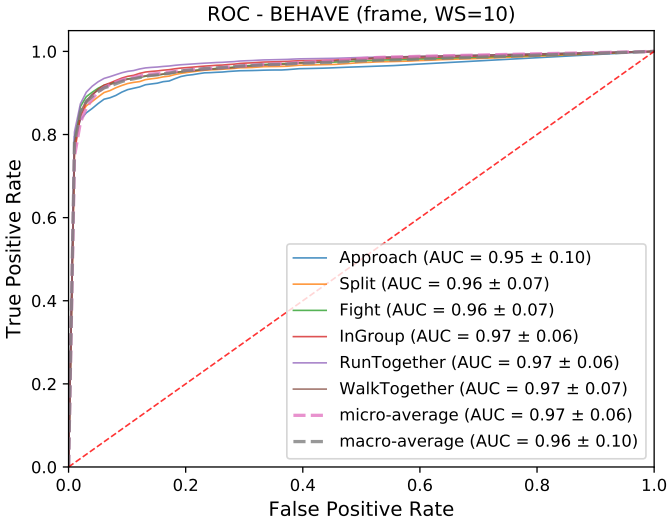
[7] Jorge Azorin-Lopez, Marcelo Saval-Calvo, Andres Fuster-Guillo, Jose Garcia-Rodriguez, and Sergio Orts-Escolano. Self-organizing activity description map to represent and classify human behaviour. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2015.

[8] Jorge Azorin-López, Marcelo Saval-Calvo, Andrés Fuster-Guilló, and Antonio Oliver-Albert. A predictive model for recognizing human behaviour based on trajectory representation. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 1494–1501. IEEE, 2014.

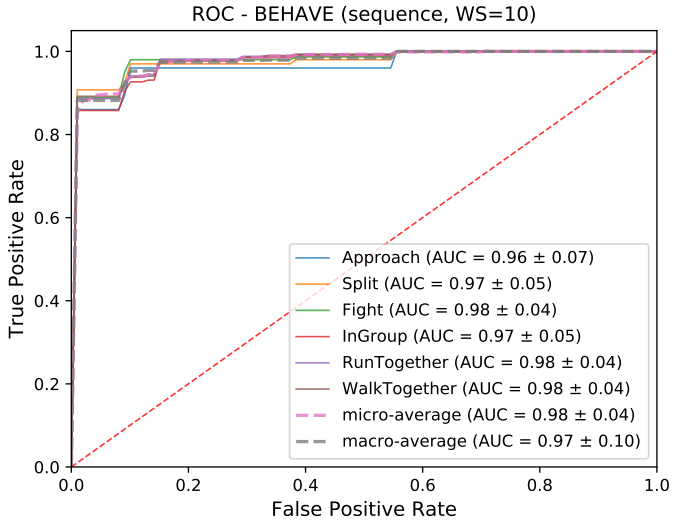
TABLE I: Comparison of results with different value of parameter Window Size (WS) for sequence and frame.

		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%

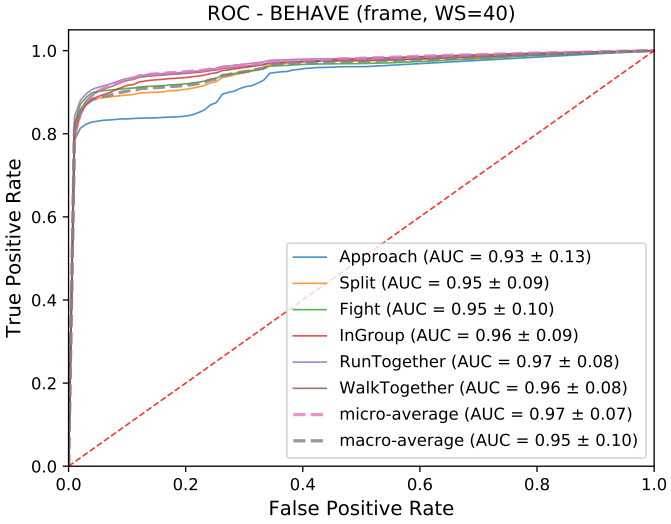
* WS = windosize



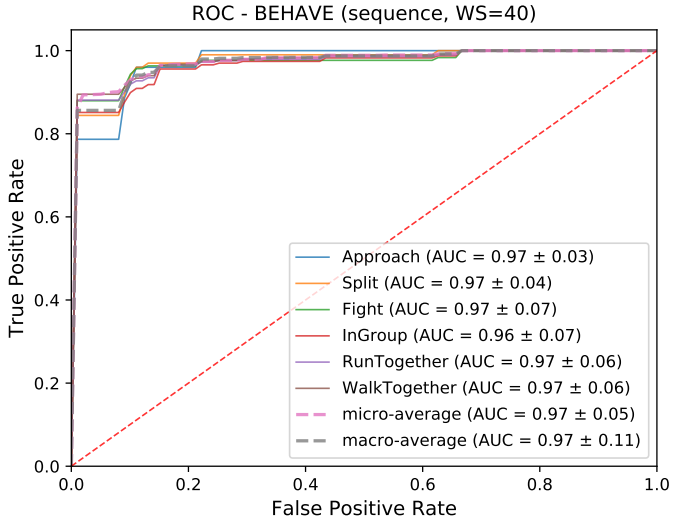
(a) ROC per frame and WS=10



(b) ROC per sequence and WS=10

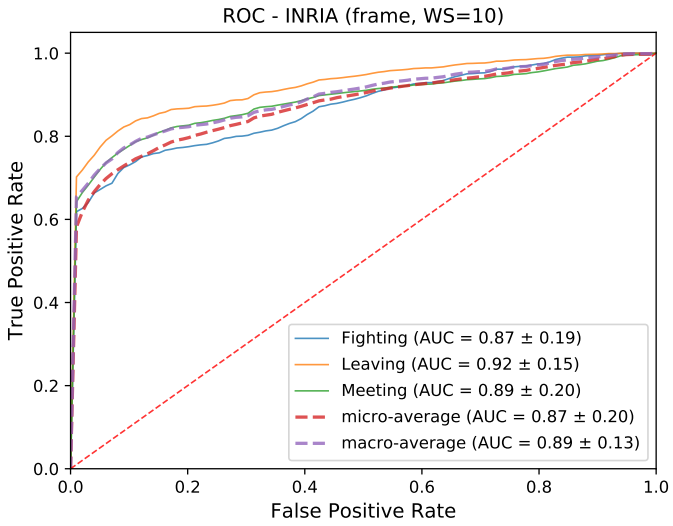


(c) ROC per frame and WS=40

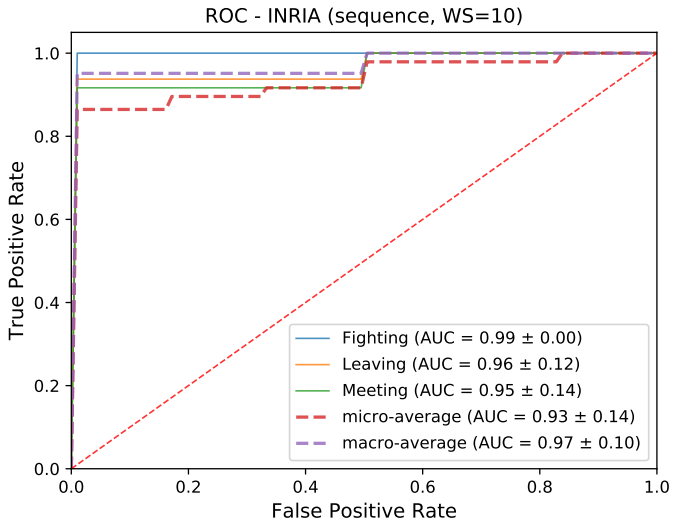


(d) ROC per sequence and WS=40

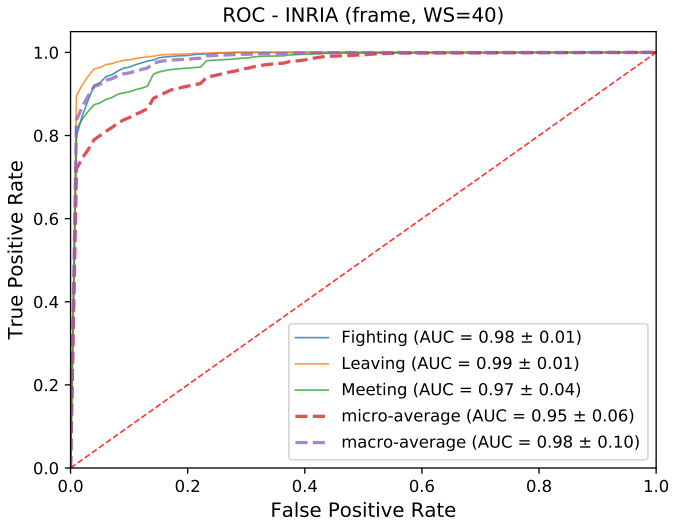
Fig. 2: ROC curves for BEHAVE dataset



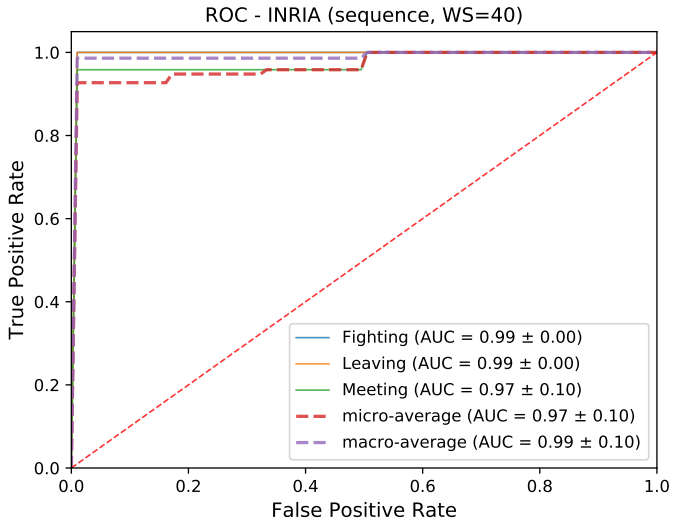
(a) ROC per frame and WS=10



(b) ROC per sequence and WS=10



(c) ROC per frame and WS=40



(d) ROC per sequence and WS=40

Fig. 3: ROC curves for INRIA dataset

TABLE II: Comparison according BEHAVE results

	D-ADV	GADV [3]	[16]	[58]	[38]	[56]
Approach	93,94%	100,00%	83,33%	71,00%	60,00%	
Split	97,14%	100,00%	100,00%	79,00%	70,00%	93,10%
WalkTogether	96,92%	86,67%	91,66%	88,00%	45,00%	92,10%
InGroup	93,10%	86,67%	100,00%	88,00%	90,00%	94,30%
Fight	94,44%	90,00%	83,33%			95,10%
RunTogether	100,00%	100,00%	83,33%			
Average	**95,93%**	93,89%	90,28%	81,50%	66,25%	93,65% t

[9] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential deep learning for human action recognition. In Albert Ali Salah and Bruno Lepri, editors, *Human Behavior Understanding*, pages 29–39, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[10] Suresh Balakrishnama and Aravind Ganapathiraju. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18:1–8, 1998.

[11] Scott Blunsden and RB Fisher. The behave video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA*, 4(1-12):4, 2010.

[12] Luis Felipe Borja, Jorge Azorin-Lopez, and Marcelo Saval-Calvo. A compilation of methods and datasets for group and crowd action recognition. *International Journal of Computer Vision and Image Processing (IJCVIP)*, 7(3):40–53, 2017.

[13] Pierre Bour, Emile Cribelier, and Vasileios Argyriou. Crowd behavior analysis from fixed and moving cameras. In Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe, editors, *Multimodal Behavior Analysis in the Wild*, Computer Vision and Pattern Recognition, pages 289 – 322. Academic Press, 2019.

[14] Alexandros André Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, 39(12):10873–10888, 2012.

[15] Guangchun Cheng, Yiwen Wan, Abdullah N Saudagar, Kamesh Namuduri, and Bill P Buckles. Advances in human action recognition: A survey. *arXiv preprint arXiv:1501.05964*, 2015.

[16] Nam-Gyu Cho, Young-Ji Kim, Unsang Park, Jeong-Seon Park, and Seong-Whan Lee. Group activity recognition with group interaction zone based on relative distance between human objects. *International Journal of Pattern Recognition and Artificial Intelligence*,

29(05):1555007, 2015.

[17] Wenwen Ding, Kai Liu, Xujia Fu, and Fei Cheng. Profile hmms for skeleton-based human action recognition. *Signal Processing: Image Communication*, 42:109–119, 2016.

[18] Chairani Fauzi, Selo Sulistyo, et al. A survey of group activity recognition in smart building. In *2018 International Conference on Signals and Systems (ICSigSys)*, pages 13–19. IEEE, 2018.

[19] Jiageng Feng, Songyang Zhang, and Jun Xiao. Explorations of skeleton features for lstm-based action recognition. *Multimedia Tools and Applications*, 78(1):591–603, 2019.

[20] Robert B Fisher. The pets04 surveillance ground-truth data sets. In *Proc. 6th IEEE international workshop on performance evaluation of tracking and surveillance*, pages 1–5, 2004.

[21] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and Jose Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70:41 – 65, 2018.

[22] Dariu M Gavrila. The visual analysis of human movement: A survey. *Computer vision and image understanding*, 73(1):82–98, 1999.

[23] D. Gowsikhaa, S. Abirami, and R. Baskaran. Automated human behavior analysis from surveillance videos: a survey. *Artificial Intelligence Review*, 42(4):747–765, Dec 2014.

[24] Hossein Mousavi Hamidreza Rabiee, Javad Haddadnia. Emotion-Based Crowd Representation for Abnormality Detection Hamidreza. *International Journal on Artificial Intelligence Tools*, 2016.

[25] Yamin Han, Peng Zhang, Tao Zhuo, Wei Huang, and Yanning Zhang. Going deeper with two-stream convnets for action recognition in video surveillance. *Pattern Recognition Letters*, 107:83–90, 2018.

[26] Longlong Jing, Yuancheng Ye, Xiaodong Yang, and Yingli Tian. 3d convolutional neural network with multi-model framework for action recognition. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1837–1841. IEEE, 2017.

[27] Qiuhong Ke, Jun Liu, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. Computer vision for human–machine interaction. In Marco Leo and Giovanni Maria Farinella, editors, *Computer Vision for Assistive Healthcare*, Computer Vision and Pattern Recognition, pages 127 – 145. Academic Press, 2018.

[28] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[29] Priyanka Kokate and Surendra Gupta. Behavior analysis from videos using motion based feature extraction. 2019.

[30] Thomas Kopinski, Stéphane Magand, Uwe Handmann, and Alexander Gepperth. A pragmatic approach to multi-class classification. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2015.

[31] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[32] Weiyao Lin, Ming-Ting Sun, Radha Poovandran, and Zhengyou Zhang. Human activity recognition for video surveillance. *IEEE International Symposium on Circuits and Systems*, (June):2737–2740, 2008.

[33] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C Kot. Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing*, 27(4):1586–1599, 2017.

[34] Diogo Carbonera Luvizon, Hedi Tabia, and David Picard. Learning features combination for human action recognition from skeleton sequences. *Pattern Recognition Letters*, 99:13–20, 2017.

[35] Thomas M Martinetz, Stanislav G Berkovich, and Klaus J Schulten. 'neural-gas' network for vector quantization and its application to time-series prediction. *IEEE transactions on neural networks*, 4(4):558–569, 1993.

[36] Stephen J McKenna, Sumer Jabri, Zoran Duric, and Harry Wechsler. Tracking interacting people. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 348–353. IEEE, 2000.

[37] Antonio S Micilotta, Eng-Jon Ong, and Richard Bowden. Detection and tracking of humans by probabilistic body part assembly. In *BMVC*, number 1, pages 429–438, 2005.

[38] David Münch, Eckart Michaelsen, and Michael Arens. Supporting fuzzy metric temporal logic based situation recognition by mean shift clustering. In *Annual Conference on Artificial Intelligence*, pages 233–236. Springer, 2012.

[39] Jacinto C Nascimento and Jorge S Marques. Performance evaluation of object detection algorithms for video surveillance. *IEEE Transactions on Multimedia*, 8(4):761–774, 2006.

[40] Shipra Ojha and Sachin Sakhare. Image processing techniques for object tracking in video surveillance-a survey. In *2015 International Conference on Pervasive Computing (ICPC)*, pages 1–6. IEEE, 2015.

[41] Kenji Okuma, Ali Taleghani, Nando De Freitas, James J Little, and David G Lowe. A boosted particle filter: Multitarget detection and tracking. In *European conference on computer vision*, pages 28–39. Springer, 2004.

[42] Maja Pantic and Ioannis Patras. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(2):433–449, 2006.

[43] Stergios Papadimitriou, Seferina Mavroudi, Liviu Vladutu, Georgios Pavlides, and Anastasios Bezerianos. The supervised network self-organizing map for classification of large data sets. *Applied intelligence*, 16(3):185–203, 2002.

[44] Arthur EC Pece. From cluster tracking to people counting. In *IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, pages 9–17, 2002.

[45] Huy-Hieu Pham, Louahdi Khoudour, Alain Crouzil, Pablo Zegers, and Sergio A Velastin. Learning to recognise 3d human action from a new skeleton-based representation using deep convolutional neural networks. *IET Computer Vision*, 13(3):319–328, 2018.

[46] Zirui Qiu, Jun Sun, Mingyue Guo, Mantao Wang, and Dejun Zhang. Survey on deep learning for human action recognition. In *International Conference of Pioneering Computer Scientists, Engineers and Educators*, pages 3–21. Springer, 2019.

[47] Lawrence R Rabiner and Biing-Hwang Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

[48] Rémi Ronfard, Cordelia Schmid, and Bill Triggs. Learning to parse pictures of people. In *European Conference on Computer Vision*, pages 700–714. Springer, 2002.

[49] Marcelo Saval-Calvo, Jorge Azorín-López, and Andrés Fuster-Guilló. Comparative analysis of temporal segmentation methods of video sequences. In *Robotic Vision: Technologies for Machine Learning and Vision Applications*, pages 43–58. IGI Global, 2013.

[50] Kevin Smith, Daniel Gatica-Perez, and J-M Odobez. Using particles to track varying numbers of interacting people. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 962–969. IEEE, 2005.

[51] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[52] Pavan Turaga, Rama Chellappa, Venkatramana S Subrahmanian, and Octavian Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video technology*, 18(11):1473, 2008.

[53] Liangliang Wang, Lianzheng Ge, Ruifeng Li, and Yajun Fang. Three-stream cnns for action recognition. *Pattern Recognition Letters*, 92:33–40, 2017.

[54] Xianpei Wang, Hua Xu, Sheng Zheng, and Anyu Cheng. Evidential reasoning research on intrusion detection. In *Fifth International Symposium on Instrumentation and Control Technology*, volume 5253, pages 930–934. International Society for Optics and Photonics, 2003.

[55] Mehran Yazdi and Thierry Bouwmans. New trends on moving object detection in video images captured by a moving camera: A survey. *Computer Science Review*, 28:157–177, 2018.

[56] Yafeng Yin, Guang Yang, and Hong Man. Small human group detection and event representation based on cognitive semantics. In *2013 IEEE seventh international conference on semantic computing*, pages 64–69. IEEE, 2013.

[57] Qing-song ZENG, Ming-hui YU, Wei-guo HE, and Ling LI. A new algorithm of action recognition. *Journal of Kunming University of Science and Technology (Science and Technology)*, (6):13, 2009.

[58] Cong Zhang, Xiaokang Yang, Weiyao Lin, and Jun Zhu. Recognizing human group behaviors with multi-group causalities. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 44–48. IEEE, 2012.