Noname manuscript No.
(will be inserted by the editor)

Constrained Self-Organizing Feature Map to Preserve Feature Extraction Topology

Jorge Azorin-Lopez · Marcelo Saval-Calvo · Andres Fuster-Guillo · Jose Garcia-Rodriguez · Higinio Mora-Mora

Received: date / Accepted: date

Abstract In many classification problems, it is necessary to consider the specific location of an n-dimensional space from which features have been calculated. For example, considering the location of features extracted from specific areas of a two-dimensional space, as an image, could improve the understanding of a scene for a video surveillance system. In the same way, the same features extracted from different locations could mean different actions for a 3D HCI system. In this paper, we present a self-organizing feature map able to preserve the topology of locations of an n-dimensional space in which the vector of features have been extracted. The main contribution is to implicitly preserving the topology of the original space because considering the locations of the extracted features and their topology could ease the solution to certain problems. Specifically, the paper proposes the n-Dimensional constrained Self-Organizing Map Preserving the Input Topology (nD-SOM-PINT). Features in adjacent areas of the n-dimensional space, used to extract the feature vectors, are explicitly in adjacent areas of the nD-SOM-PINT constraining the neural network structure and learning. As a study case, the neural network has been instantiate to represent and classify features as trajectories extracted from a sequence of images into a high level of semantic understanding. Experiments have been thoroughly carried out using the CAVIAR datasets (Corridor, Frontal and Inria) taken into account the global behaviour of an individual in order to validate the ability to preserve the topology of the two-dimensional space to obtain high performance classification for trajectory classification in contrast of non-considering the location of features. Moreover, a brief example

This study was supported in part by the University of Alicante, Valencian Governmen and Spanish government under grants GRE11-01, GV/2013/005 and DPI2013-40534-R.

Department of Computer Technology
University of Alicante
03080 Alicante, Spain
E-mail: {jazorin,msaval,fuster,jgarcia,hmora}@dtic.ua.es

has been included to focus on validate the nD-SOM-PINT proposal in other domain than the individual trajectory. Results confirm the high accuracy of the nD-SOM-PINT outperforming previous methods aimed to classify the same datasets.

Keywords Self-organizing feature map · Topology preservation · human behaviour analysis

1 Introduction

For most pattern recognition and machine learning applications, the original input data is transformed to a different space of variables in order to easily solve the problem. Applications process the original input data, calculating features from them, in order to extract valuable information that will be used by the clustering or classification step. Usually, this feature extraction step considers a transformation from the input data space to another one considering the input space as a whole. For example in computer vision, considering the recognition of animals from colour images, features (corners, edges, colour histograms, etc.) are calculated from the two-dimensional space (the image) in order to transform it to a n-dimensional space of features (SIFT - Scale-invariant feature transform, SURF - Speeded Up Robust Features, etc.) that could categorize easier the animals . In this example of image classification, the location of the features in the input space, the image, could be irrelevant for the main objective of the computer vision system that is to properly identify animals.

However depending on the application purpose, sometimes it is important to consider the location from which the features have been calculated. Using the above same example, if the application requires not only to know if there exist an animal but also the position of it in the image, the location of the extracted features from the image become to be part of the transformed space used by the computer vision method in order to provide the correct solution. Moreover, regarding the location of a feature in the input space, the meaning of it could be different. For example, consider now a surveillance system of a metropolitan railway network and *walking* as a feature extracted from the video system to describe the movement of a person. The meaning of the action of a person *walking* close or crossing a safety line at the edge of a platform is very different that the meaning of the action of a person *walking* in the hall. In order to detect the intrusion of the first person, the location again could become to be part of the transformed space of variables.

Finally, the relationships between locations of features as the topology of the input space could be very important to solve a pattern recognition problem. For example in computer vision, relationships among keypoints (points of interest in the image) are a very important mechanism to solve different problems including identification, recognition and image registration [17], [9], [36]. Usually, these methods detect and extract some features in the image (e.g. SIFT or SURF) considering the specific location (keypoint) of them in the

image. The features and the locations are used by pattern recognition methods (e.g. k-NN, SVM, SOM, etc.) to find matches between them (e.g. from an image in a database and an image query for image retrieval or from two images for image registration). Finally, a model based method (e.g. RANSAC) is used to reject inconsistent matches and provide a good solution. In this case, the relationships between topology, location, distances, etc. in the input space in which the features has been extracted are the key aspect to solve the problem.

In this paper, we are interested in considering the location of the extracted features from the input space and its topology as an important feature. Hence, we are interested in representing features extracted from a n-dimensional space for a posterior classification but preserving the topological relations of the input space in which the features have been extracted. Although, there exist different definitions and topology preserving measures, in general terms, a network preserves the topology if it preserves the input neighbourhood. It means that adjacent variables in the feature vector space are adjacent in the network as well [41]. In consequence, we propose the n-Dimensional constrained Self-Organizing Map Preserving the Input Topology (nD-SOM-PINT) map. It is a variant of the self-organizing feature map that is able to preserve the topological information of the original space from which the feature vector has been extracted. Features in adjacent areas of the n-dimensional original space are explicitly in adjacent areas of the self-organizing map preserving the input topology. The structure and the learning algorithm is constrained to the topology of the input space. The nD-SOM-PINT is evaluated in this paper by a case of study aimed to represent and classify trajectories of people from video sequences into high level of semantic understanding: human behaviour analysis. The neural network is able to deal with the big gap between human trajectories in a scene and the global behaviour associated to them preserving the spatial information about trajectories.

A Self-Organizing Map (SOM) is one of the most popular and used artificial neural network model. It was introduced by Kohonen in 1982 [20] and nowadays it remains been used and applied in many areas. It mainly converts a high dimensional input into a low dimensional map of features, being the two-dimensional map the most used representation. The map structure varies in some characteristics (lattice shape, neighbourhood connections, etc.) but generally it conforms a grid of interconnected neurons with a specific neighbourhood. This interconnection assures a topological preservation in the map space. The SOM uses a unsupervised learning algorithm based on a competition where nodes of the grid compete to become the winning neuron (the most similar neuron to the input vector) in order to adjust its vector components and those of the nearest neighbour neurons. This competition allows to project the input space into the map.

Different variants have been proposed to the classical SOM for many years [18]. For example, variants proposed by Fritzke as the Growing Cell Structure, GC [13] and Growing Grid, GG [14] are able to automatically find a suitable network structure and size by a controlled growth process, while maintain the network topology. Other networks as Neural Gas, NG [25] and Growing Neu-

ral Gas, GNG [15] are able to reconfigure the neighbourhood relationships to avoid a predefined network topology. They are able to make explicit the important topological relations in a given distribution of input data. Moreover, the Growing Neural Gas eliminates the need to predefine the network size in the similar way as the Growing Cell Structure. In consequence, the above neural networks are able to transform high-dimensional input data manifolds (the input space) onto elements of a low-dimensional array (the projection space). The SOM, GG and GC preserve the topological constraints in the space conformed by the map (the projected space)[40]. That is, they are always able to preserve the neighbourhood relations in the map but although nearby data vectors in the input space are mapped onto neighbouring locations in the output space, they are not always able to preserve the input space topology as GNG and NG due to the fixed predefined topology of the map. For example, if we consider in a 2D input space a distribution of points composed by two unconnected circles, the SOM, GG and GC are not able to preserve the topology of the unconnected points pertaining to the two different circles. However, since GNG and NG can adapt their topology, they are able to properly adapt to the unconnected circle distribution. Finally, although both GNG and NG can preserve the topology of the input space, they are not able to preserve the topology of the original space used to extract the feature vectors unless they become part of the input space of the network. Following the above examples about computer vision, consider that some SIFT descriptors have been extracted from an image to recognise animals using a GNG. The topology of the input space of the network, composed by the SIFT space, could be preserved by the GNG but it is not necessarily able to preserve the topology in the original space (the image) because it is not explicitly considered. We are interested in implicitly preserving the topology of the original space (the image in the previous example) in order to ease the solution to certain problems.

The remainder of the paper is organized as follows. Section 2 presents the novel Constrained Self-Organizing Map Preserving Topology (nD-SOM-PINT). In Section 3, the nD-SOM-PINT is instantiated and evaluated by a case of study aimed to represent and classify trajectories of people from video sequences into human behaviour analysis. Experimental results of the case of study are presented in Section 4. They are discussed and compared to other approaches in the same Section. Finally, conclusions about the research are presented in Section 5.

2 N-Dimensional Constrained Self-Organizing Map Preserving the Input Topology

The n-Dimensional constrained Self-Organizing Map Preserving the Input Topology (nD-SOM-PINT) is a novel neural network able to represent features and classify them preserving the input space topology. It is based on the classical SOM. Briefly, a SOM neural network involves two phases: a training/learning and a classification process. For the training phase using a se-

quential learning, in each step, one sample (a vector) is selected from the input data set to calculate a distance measure between the sample and each neuron (the corresponding reference vector for the neuron). The neuron whose reference vector is closer to the input sample is selected as the winning neuron. The neighbourhood of the winning neuron is adapted to the input sample. In this paper, self organizing basis are considered to represent features preserving the n-dimensional input space in a map with the same dimensions. The whole map is used in classification tasks. Some variants have been introduced in the learning and classification process to constrain them to the input topology preserving it in the network.

Specifically, the nD-SOM-PINT uses as input data a collection of features F extracted from an n-dimensional space S (the space in which the topology is preserved). As we are interested in preserving the topological information, an n-dimensional matrix, Feature Matrix (FM), is the specific collection of features (see Fig. 1). Note that not necessarily all elements of the matrix FM must have a value different to zero. The matrix suits the input space S. However, features could be calculated only for a subset of the space S. Moreover, we assume that the input space S could be discretized into cells C that represent a subset of S. Hence, each component of the FM, represents a region of the input space S, a specific cell of C, by means of the feature F calculated in that region of the space.

The network structure is completely related to the Feature Matrix (FM) and is specified as (see Fig. 2):

- A set N of neurons that represent cells of the grid C in which the input space S has been discretized. Each neuron $\nu \in N$ stores an associated reference vector $w_\nu \in R^n$. The reference vectors are related to the features extracted from the discretized space C of the region assigned to the neurons.
- A structure about connections between adjacent neurons. It defines the topological structure of the map that represents the geometry of the input space. For example, if the input space is an image, a two-dimensional regular grid of neurons.

The size of the map is related with the number of cells in C. Specifically, a nD-SOM-PINT establishes a subset of neurons N_i to represent a specific cell C_i. Each N_i contains a master neuron and a neighbourhood of neurons within a radius r_c. The number of neurons in N_i depends on the lattice shape, neighbourhood connections, etc. The connections between adjacent neurons pertaining to different N_i follow the same structure that the connections between adjacent neurons representing the same cell.

It is important to highlight that the nD-SOM-PINT is based on SOM because the network requires to adapt to the specific characteristics of the input space using a fixed number of neurons, related to the number of cells in C, and a fixed topology, closely related to the space S used to extract features. However, other self-organizing networks as GG or GC could be considered in order to reduce the number of neurons in N_i used to represent a specific cell C_i.

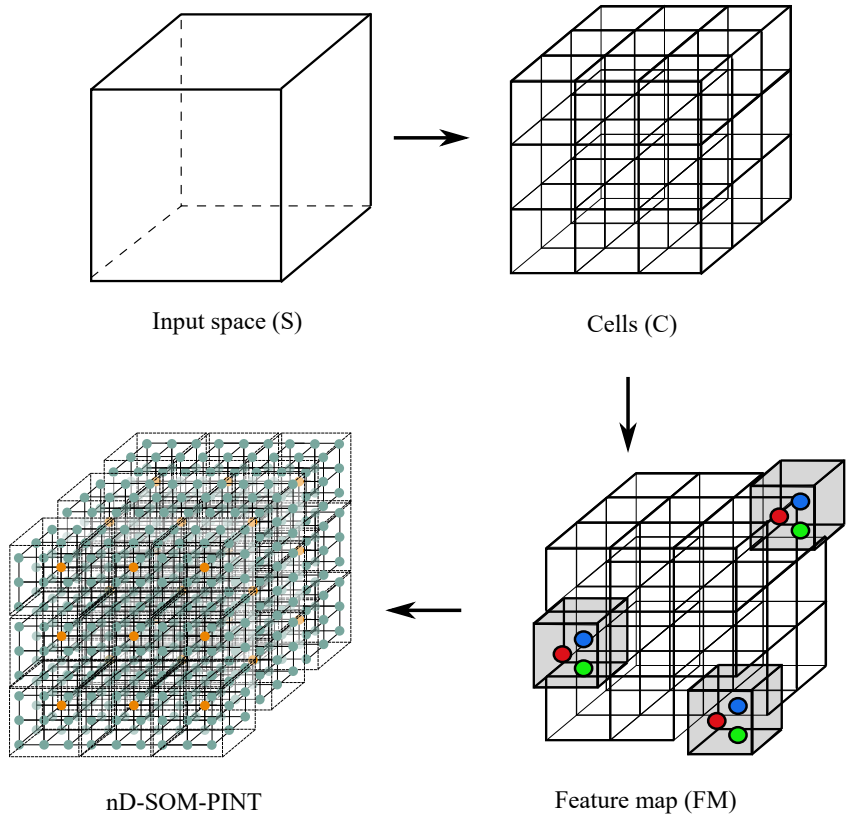


Fig. 1: The input space S is discretized into a set of cells C. For each cell in C, features are extracted to calculate the feature map (FM). The FM is the input data for the nD-SOM-PINT.

2.1 Learning process

Although we are interested in an unsupervised learning, nD-SOM-PINT considers specific labels for classification purposes. The label is the higher level of semantic that a FM represents. Hence, a different nD-SOM-PINT$_b$ is trained for a specific label b. In order to train the maps, the input dataset $X = \{FM_1, FM_2, \ldots, FM_n, \}$ is divided into different groups according to samples pertaining to a single label $X_b \in X$. The training process for each nD-SOM-PINT$_b$ considers a set of winning neurons as opposite of a winner for an input FM$_i$ (as in the original SOM). Each set of features corresponding to a specific cell activate a winning neuron in the corresponding N_i associated to it. The neuron is not necessarily the master neuron. In consequence, the whole map is

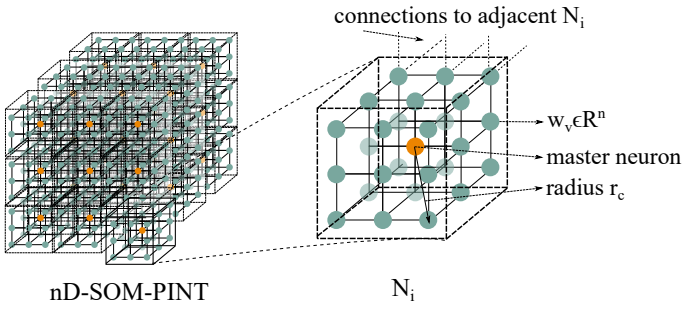


Fig. 2: nD-SOM-PINT structure. Each set of neurons N_i for each cell in C is composed by a master neuron and a neighbourhood of neurons within a radius r_c.

adapted considering all winning neurons (one for each N_i) in nD-SOM-PINT$_b$ to represent the FM for a specific label.

The learning algorithm for a single nD-SOM-PINT$_b$ having an input dataset $X_b = \{FM_1, FM_2, \ldots, FM_n, \}$ of a specific label is as follows (see Fig. 3):

1. Initialize each neuron $\nu \in$ nD-SOM-PINT$_b$ with random values w_ν in R^n.
2. Choose randomly a sample pattern FM_i from the input data set X.
3. For each element F_j in the FM_i matrix, find the nearest neuron, winning neuron, s_j in the corresponding N_j set of neurons associated to the cell C_j. In consequence, a set of winning neurons $S \in$ nD-SOM-PINT will be associated to the sample pattern FM_i. The size of the winning neuron is the same that the number of cells in which the space S was discretized.

$$\|w_{s_j} - F_j\|^2 = \min_{\nu \in N_i} \left\{ \|w_\nu - F_j\|^2 \right\} \tag{1}$$

4. Update the map
   - Determine the adapted neighbours and the strength of the adaptation by the neighbourhood function $h(d,t)$ for each winning neuron s_j. It depends on the distance d of each neuron in nD-SOM-PINT to the winning neuron and on the training time t.

$$a_{s_j} = h_{s_j}(d,t) * \|w_{s_j} - F_j\| \tag{2}$$

   - Update the reference vector w_ν for each $\nu \in$ nD-SOM-PINT according to all adapted neighbours a_{s_j}. A neuron ν could be affected by the adaption of different neighbourhoods. Finally, the strength of the update for each reference vector w is established by the learning rate α. It depends also on the training time t.

$$w_\nu(t+1) = w_\nu(t) - \alpha(t) * \sum_{\forall s_j \in S} a_{s_j} \tag{3}$$

5. If the training time t is not yet achieved, go to step 2.

2.2 Classification process

For classification purposes, each nD-SOM-PINT$_b$ is compared to a new input FM to establish the minimum distance for the whole map. The classification process (see Fig. 4) of a new input data FM, for a set of k labels represented by the set $B = \{\text{nD-SOM-PINT}_1, \text{nD-SOM-PINT}_2, \ldots, \text{nD-SOM-PINT}_k\}$ is as follows (see Fig. 4):

1. Determine the distance η from the FM to each nD-SOM-PINT$_i \in B$
   - For each element F_j in the FM matrix, find the nearest neuron, winning neuron, s_j in the corresponding N_i set of neurons associated to the cell C_i as in the step 3 of the training process.

$$\|w_{s_j} - F_j\|^2 = \min_{\nu \in N_i} \{\|w_\nu - F_j\|^2\} \tag{4}$$

   - Determine the sum of distances of the winning neurons in S

$$\eta_{\text{nD-SOM-PINT}_i} = \sum |w_{s_j} - F_j\|^2 \tag{5}$$

2. Select the nD-SOM-PINT$_i$ with minimum distance $\eta_{\text{nD-SOM-PINT}_i}$ as the label associated to the FM input

$$label = \min_{\forall \text{nD-SOM-PINT}_i \in B} (\eta_{\text{nD-SOM-PINT}_i}) \tag{6}$$

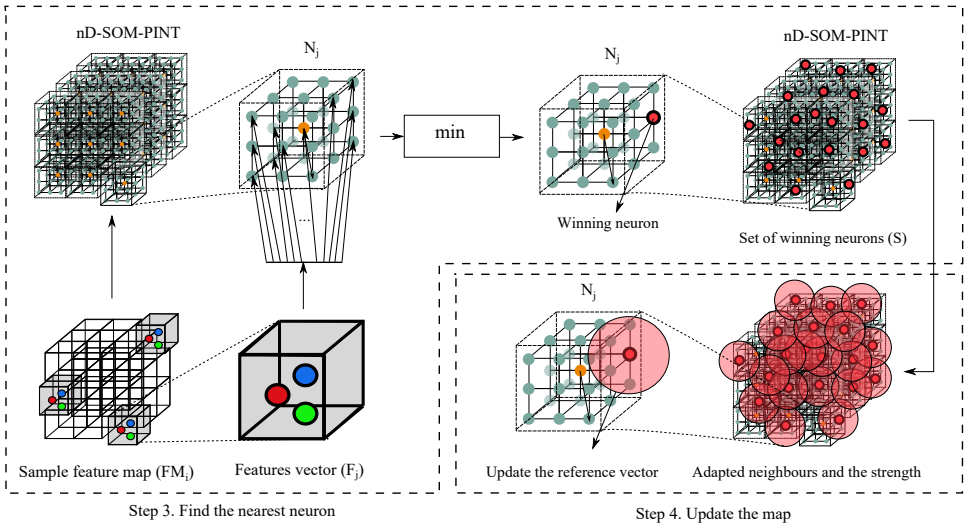


Fig. 3: Learning process. Each feature vector (F_i) in the Feature Map (FM) is compared in parallel with its corresponding set of neurons N_i to establish the set S of winning neurons (step 3). The whole map is adapted with respect to the set S (step 4)

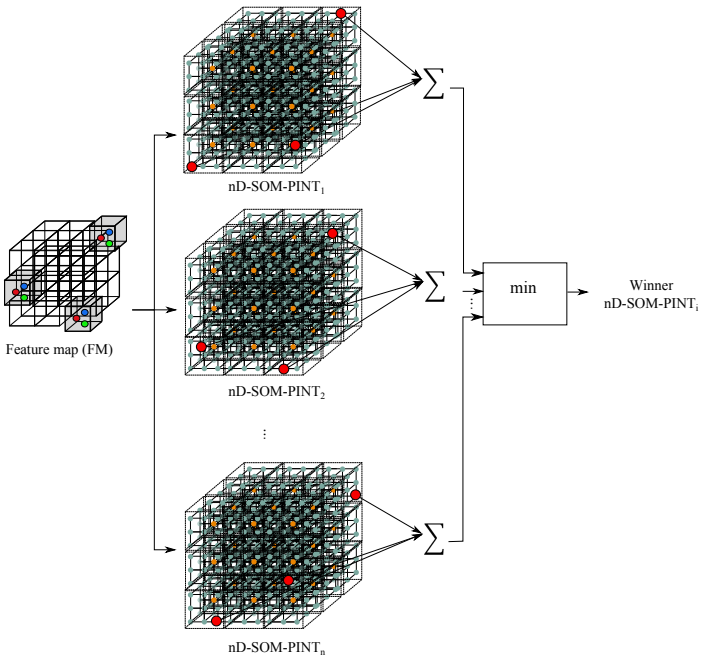


Fig. 4: Classification process. Each nD-SOM-PINT$_b$ is compared to a new input FM to establish the minimum distance for the whole map and to determine the winner nD-SOM-PINT.

3 Case of study: individual human behaviour analysis

The research in analysing human behaviour from video sequences is not just a consolidated topic but it is also an increasing area according to the number of research works related with it. The behaviour can be studied from different levels of understanding, as single movements such as a step or a hand displacement in the lowest level, to complex activities or behaviours in the highest. A classification of those levels can be found in [27] where four levels are proposed: motion, action, activity and behaviour from lower to upper. The terms *activity* and *behaviour* can be found interchangeable depending on the authors or the purpose of the research in the literature. We refer to the following papers, [21, 28, 12], for general reviews of human behaviour studies.

In this paper as a case of study of the nD-SOM-PINT proposal, we are focused in the behaviour level using video sequence data. Interesting surveys could be found in [38] and [4]. The methods can be grouped in state models (e.g. Bayesian, HMM), pattern recognition (e.g. Neural Networks, SVM), and semantic models (e.g. Petri Nets, grammars) following the survey by Lavee et al. [21]. Pattern recognition methods use the actual data to cluster the space

of solutions, being more flexible than state and semantic models which have to predefine a model and rules. Regarding the previous aspects, this paper is focused in pattern recognition techniques, concretely in Self-Organizing neural networks, in order to analyse global people behaviour by means of the trajectory described by them in a scenario. The trajectory provides very interesting data for a wide rage of activities, and it is also easy to obtain due to the large number of cameras available nowadays, mainly CCTV.

Self-Organizing networks have been widely used in Computer Vision. Concretely, in trajectory analysis, various studies are present in the state-of-the-art. Morris and Trivedi [29] present a framework for live video analysis in which the behaviours are described using motion patterns, for real-time characterization and prediction of future activities, as well as the detection of abnormalities. Trajectories are utilized to automatically build activity models in a 3-stage hierarchical learning process. Owens et al. [33] use a flow vector to sample the track information to train a Self-Organizing Map (SOM). Martinez-Contreras et al. [26] use SOMs only for motion (trajectory) sampling. A SOM is trained with different motions, and then a new motion is classified and the template is used in a Hidden Markov Model to determinate the action. Schreck et al. [37] developed a framework to classify trajectories using SOMs, scaling the paths into unit square values and sampling them in a predefined number of parts. Saul et al. [35] compared map-based trajectory analysis with SOM to detect unusual behaviour of traffic objects, where the network achieves better results in abrupt avoidance detection. Andrew et al. [22] explore the use of a SOM to visualize patterns of urban social change. They were able to visualize geospatial patterns of socio-economic change and the magnitude and direction of change. Hu et al. [16] use the whole trajectory as an input to the fuzzy Self-Organizing network to learn the trajectory model of real world pedestrians and toy cars. Madokoro et al. [24] extracts typical behaviour patterns and specific behaviour patterns from human trajectories quantized using One-Dimensional Self-Organizing Maps (1D-SOMs). Subsequently, they apply Two-Dimensional SOMs (2D-SOMs) for unsupervised classification of behaviour patterns. A hierarchical SOM is presented in [34] to detect abnormal human activities based on trajectories, body features and directions, showing accurate results in different scenarios and sensors.

Normally, raw trajectories are not studied directly using pattern analysis due to the varying in length of data (same trajectory pattern can be done slower or making small variations of the path). Therefore, a normalization of data has to be done. Hu et al. [16] propose a normalization by using a maximal length component vectors, filling the empty data of shorter paths with no movement. In [1–3] PCA is used to sample trajectories. Meanwhile, Xi et al. [23] proposed a Trajectory Directional Histogram (TDH) to describe the statistic directional distribution of one trajectory. In [32], on the other side, they use a Discrete Fourier Transform to reduce the components of the trajectories. In our previous works [5, 7], we proposed a descriptor of the behaviour using the trajectory information. As the trajectory classification needs a normalization technique to equalize the length of the data, the the Activity Description

Vector (ADV) was proposed to equally divide the scene in cells, and estimates the up, down, left, right and frequency of the person in each cell. We evaluated this using different pattern recognition techniques obtaining high accuracy in classification. Moreover, this descriptor has been proved to obtain prediction capabilities for early recognition [7]. The ADV descriptor is able to represent the behaviour by means of trajectories in a very efficient way. It is capable of outperform state-of-the art methods as it summarizes the activity in a specific region of the scene without taking into account the neighbourhood. However, it lacks from topological relationships preservation between cells in which the ground plane is divided. The activity is constrained to a specific region of the scene and relations between adjacent areas of the scene are not considered. However, it is more likely to have similar behaviours in adjacent areas of the scene than in not connected areas. For example, if a person is walking in a specific region of the scene, it is more likely that he or she be walking in an adjacent areas of the scene. Otherwise, the division in cells of the scene could be more critical.

This paper proposes the use of the nD-SOM-PINT to represent and classify high level of semantic understanding from video sequences. The neural network is able to deal with the big gap between human trajectories in a scene and the global behaviour associated to them. This map is able to preserve the topological information about the scene which is very important when the spatial information is treated. In this case of study, the nD-SOM-PINT proposal uses the ADV descriptor as inputs of the network, the Feature Map (FM), in order to incorporate the trajectory information and to preserve topological information about trajectories of the people according to adjacent cells.

3.1 Feature map: Activity description vector

As we stated before, the Activity description vector (ADV) has been selected for this case of study as the Feature Map (FM). The ADV is a trajectory-based feature presented in previous works to describe global human behaviour [5] and was used as the input of an early prediction method [7,6]. For the sake of completeness, the ADV is presented but we refer you to [5] in order to obtain further details about its calculation.

The ADV descriptor is invariant to the point of view of the camera due to the trajectory is represented using the ground where people are moving as the basic geometric model. Therefore, the space of values has to be perpendicular to the point of view of the camera. If the camera is not on the roof, any information contained on the image plane captured from a static camera has to be transformed to the corresponding plane that fits the ground by means of a Homography, H (7). The projective transformation allows us to consider the whole space of movements of the people in the Euclidean space (see Figure 5). Then, any point p_i on the image is transformed to a point p_g on the ground plane G.

$$p_g = H \cdot p_i \tag{7}$$

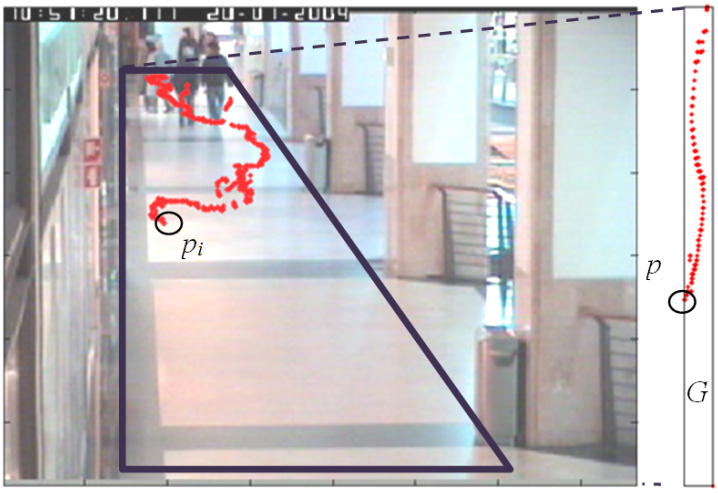


Fig. 5: Projective transformation to obtain the basic geometric model able to represent the trajectory of a person in the scenario.

Trajectories are described dividing the scene in regions and compressing the data in cumulative values. It is interesting to highlight that ADV integrates the trajectory information without length and sequential constraints, what makes it appropriate for predictive purposes. Specifically, ADV uses the number of occurrences of a person in a specific point of the scenario and the local movements performed in it. This method divides the scenario, G, in cells, C, to discretize it. Each cell of the grid has information about the movements performed it including up (U), down (D), left (L), right (R) and frequency (F) data. The four former values are extracted from the local displacement between two consecutive points of a trajectory. It assumes a local Cartesian coordinate system with origin in one of the points. Points of the trajectory are extracted according a person is tracked in the scene. In consequence, U and R are the movements in the positive axis y and x respectively. D and L are the corresponding coordinates for the negative axis. The displacement is calculated as the dot product of the displacement vector between two consecutive tracked points on, p_{g_i} and $p_{g_{i-1}}$, and the corresponding normal vector for each axis. For example, if we focus on the U movement, Eq. 8 explains how this value is extracted:

$$U_i = \begin{cases} (z_i - z_{i-1}) \cdot (0,1)^T & if \frac{(z_i - z_{i-1}) \cdot (0,1)^T}{\|(z_i - z_{i-1})\|} > 0 \\ 0 & otherwise \end{cases} \tag{8}$$

U is assumed to be a displacement in the positive vertical y axis. This formula is similar for the other three displacements. These four particular movements have information about the direction of the trajectory and the velocity of the group in a specific point on the scenario G. Additionally, we consider the frequency, F, as the number of occurrences of the centroid is in a specific point of G. That is, the number of frames that the centroid of the group has been in a specific location. F contains information about the spatial trajectory of a group but not considering the movements itself.

For each cell $C_{i,j}$, the specific $ADV_{i,j}$ is calculated as the accumulative histograms of the movements U, D, L, R and frequency F for the points on G discretized in the cell. Let u x v the actual size of the scenario, mxn the cells it has been split and $p_{k,l}$ the point located in the position k and l of the G space, each ADV in a cell is:

$$\begin{array}{l} \forall c_{i,j} \in C \wedge \forall p_{k,l} \in G / i = \lfloor \frac{kxm}{u} \rfloor \wedge j = \lfloor \frac{kxn}{v} \rfloor \\ ADV_{i,j} = \begin{pmatrix} \sum F(p_{k,l}), \sum U(p_{k,l}), \sum D(p_{k,l}) \\ \sum L(p_{k,l}), \sum R(p_{k,l}) \end{pmatrix} \end{array} \tag{9}$$

Finally, the ADV that describes a person trajectory uses a collection of the $ADV_{i,j}$ for each cell. In Fig. 6, we can see an example of different behaviours from the CAVIAR database [12]. The trajectories of different people in the image is transformed to the ground plane G of the corridor. The components of the ADV for a single trajectory in a discretized cell of 5x7 is represented as well.

3.2 nD-SOM-PINT to describe and classify human activity

The generic nD-SOM-PINT explained in Sect. 2 will be particularized to describe and classify human activity in a sequence of images. In this case, the nD-SOM-PINT uses as input an ADV collection extracted from a sequence of two-dimensional spaces, a sequence of images. In previous works [5,7,6], a vector concatenating all $ADV_{i,j}$ was used as the collection of ADVs describing a person trajectory. Hence, for a scenario space of uxv, split in mxn cells, the ADV will contain mxnx5 elements as:

$$ADV = (ADV_{1,1}, ADV_{1,2}, ADV_{1,3}, ..., ADV_{m,n}) \tag{10}$$

However, as we are interested in preserving the topological information, a two-dimensional matrix is the collection of ADVs that conform the Feature Map (FM). The two-dimensional matrix contains exactly the same elements as the cells in which the scenario G has been divided. For coherence, the Activity Descriptor Vector would be named as the Activity Descriptor Map (ADM):

$$ADM = \begin{pmatrix} ADV_{1,1} & ADV_{1,2} & \dots & ADV_{1,n} \\ ADV_{2,1} & ADV_{2,2} & \dots & ADV_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ ADV_{m,1} & ADV_{m,2} & \dots & ADV_{m,n} \end{pmatrix} \tag{11}$$

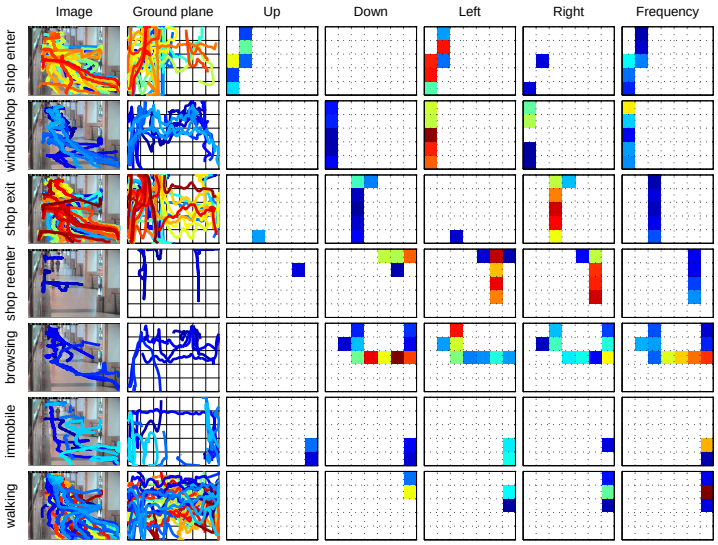


Fig. 6: Examples of CAVIAR trajectories in image space and ground plane G. U, D, L and R movements and frequencies F of the ADV representation for a specific sample.

The network structure is completely related to the two-dimensional ADM, instantiating the nD-SOM-PINT as the 2D-SOM-PINT to describe and classify human activity. In consequence, the set N of neurons represent cells of the grid C in which the ground plane G has been discretized. Each neuron $\nu \in N$ stores an associated reference vector $w_\nu \in R^5$. The reference vectors are related to ADVs (movements Up, Down, Left, Right and Frequency) of the region assigned to the neurons. The topological structure of the 2D-SOM-PINT represents the geometry of the ground plane G. Hence, a two-dimensional regular grid of neurons (sheet shape) with a rectangular lattice is proposed for the 2D-SOM-PINT.

The size of the map is determined by the number of cells in C and a subset of neurons $N_{i,j}$ to represent a specific cell $C_{i,j}$. As the generic nD-SOM-PINT, each $N_{i,j}$ in the 2D-SOM-PINT contains a neuron and a neighbourhood of neurons within a radius r_c. For example, Fig. 7 represents the structure of a 2D-SOM-PINT with $r_c = 1$ and 35 (5x7) subsets $N_{i,j}$ of 9 neurons for a ground plane divided into a 5x7 grid cells.

For learning process, a different 2D-SOM-PINT$_b$ is trained for a specific behaviour b. In consequence, the input dataset $X = ADM_1, ADM_2, \ldots, ADM_n$, is divided into different groups according to samples pertaining to a single behaviour $X_b \in X$. As we stated before, the training process for each 2D-$_b$ considers a set of winning neurons as opposite of a winner for an input ADM$_i$ (as in the original SOM). Each ADV in ADM$_i$ activate a neuron. In conse-

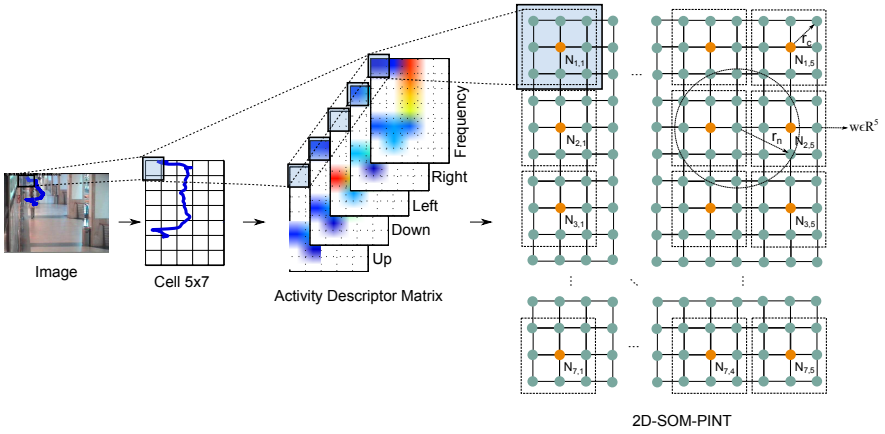


Fig. 7: Example of the structure of the nD-SOM-PINT for a 5x7 grid cell with a $r_c = 1$ for a trajectory extracted from a sequence of images.

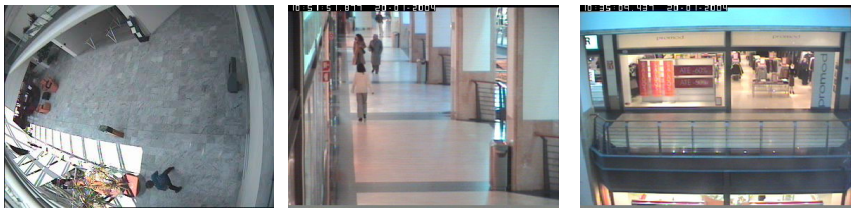
quence, the whole map is adapted considering all winning neurons to represent the ADM for a specific behaviour. For classification purposes, each 2D-SOM-PINT$_b$ is compared to a new input ADM to establish the minimum distance for the whole map.

4 Experiments

Experiments have been carried out using the CAVIAR database [12]. It contains two datasets: Inria and Shopping Centre. The first dataset was recorded in the entrance lobby of the INRIA Labs at Grenoble, France (see Fig. 8a) a image sequence of 384x288 pixels at 25 frames per second. The Shopping Centre dataset contains different clips (at the same resolution and frame rate as before) from a shopping centre in Portugal recorded from two points of view: *Corridor* view (see Fig 8b) and *Frontal view* (see Fig 8c).

The Inria dataset contains 28 clips of people. In total, it has 26.419 frames capturing 139 individuals at 25 frames per second. Although the Shopping Centre dataset was recorded at the same time from 2 different views, the total number of people and the labelled behaviours are different. The *Corridor* dataset contains information about behaviours and trajectories performed in a long corridor with different stores. In total, 235 persons were labelled in the 26 labelled clips performing 255 different trajectories. However, the *Frontal view* dataset contains information about just a specific part of the corridor (a store) having, in consequence, less people and trajectories for the same behaviours: 144 samples.

Each sequence was labelled frame-by-frame by hand and each individual is tracked using a unique identifier in the sequence. Therefore, each frame has a set of tracked individuals visible in that frame that are surrounded by a



(a) Inria frame (b) Corridor view frame (c) Frontal view frame

Fig. 8: Frames from the different image datasets

bounding box and labelled according to the situation in which the individual is involved.

Each tracked individual has a set of labels that describe different levels of understanding (from bottom level as active, inactive or walking to high level containing human behaviours). As we are interested in the high semantic level, experiments just take into account the *context* label of the CAVIAR sequences. This information is subjective and depends on the observer. Moreover, the labelled individuals in the clips could have different contexts during a sequence. The *Inria* dataset contains four different behaviours: *Walking* (73 samples), *Browsing* (11 samples), *Inmobile* (51 samples) and *Drop down* (4 samples). The *Corridor* dataset perform 255 different trajectories: shop enter (55 samples), windowshop (18 samples), shop exit (63 samples), shop reenter (5 samples), browsing (10 samples), immobile (22 samples) and walking (82 samples). Finally, the *Frontal* dataset contains the same contexts but with different samples. The samples are imbalanced. Thus, the Synthetic Minority Over-Sampling Technique (SMOTE) [31] has been applied to obtain the same number of samples for each context. Also, for the contexts with more samples, a subset of samples is randomly taken. The objective was to use exactly the same number of samples per behaviour. All these samples have been used for training and classification steps.

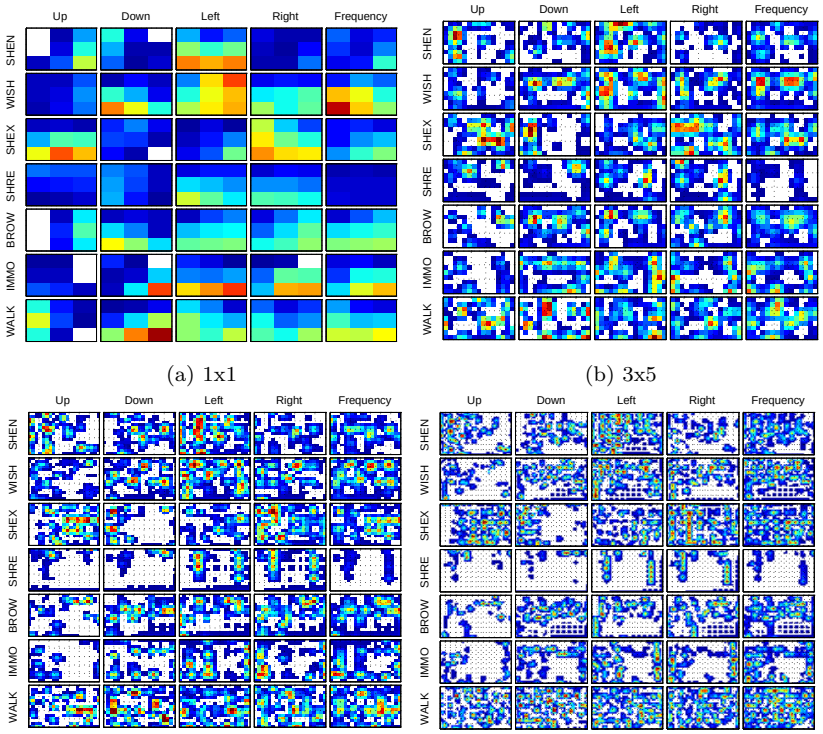
Additionally, we use the bounding box positions to calculate the ADV of each region in the scene. These positions have some errors due to the labelling was done by humans. In order to avoid the errors, a data sampling has been carried out at a sampling frequency of 1 Hz (i.e. we take into account the position data each 25 frames). Finally, a SPLINE curve is calculated from the sampled data to obtain the trajectories included in each context.

According to the 2D-SOM-PINT training, all samples in X have been normalized to the range (0,1) dividing each component of the ADV vector by the maximum value for each component. All experiments uses the same parameters that were selected experimentally for the map: the radius r_c of neighbourhood in N_i, j is 1 conforming 9 neurons for each cell C_i, j being the minimum number of neurons in the neighboorhood for a specific cell. The method will be validated with the minimum number of neurons in the experiments. The learning rate is established in range (0.5, 0.1). The neighbourhood function $h_{s_i,j}(d,t)$ in Eq. 12 to determine the strength of the adaptation radius is a

Gaussian function, being $\sigma = r_n = 2$, and d_{ci} the distance from the neuron w_c to w_i on the map grid:

$$h_{s_i,j}(d,t) = \exp^{-d_{wi}^2/2\sigma^2} \tag{12}$$

4.1 Results and discussion



(a) 1x1 (b) 3x5

(c) 5x7 (d) 9x13

Fig. 9: 2D-SOM-PINTs calculated for the *Corridor* dataset considering the different behaviours and grid sizes

Experiments have been performed for different 2D-SOM-PINTs (see Figure 9) calculated from ADMs with different grid sizes: 1x1, 3x5, 5x7 and 7x11 and 9x13. The objective is to evaluate the ability of the 2D-SOM-PINT to represent information extracted from the scene and classify human behaviour.

For each grid selection, a 10-fold cross validation has been performed to analyse the classification performance of the 2D-SOM-PINT proposal. Table 1 shows the results of classification accuracy of the 2D-SOM-PINT with different grid sizes for the different datasets considered for human behaviour analysis.

Columns present the values of Sensitivity (correctly classified positive samples / true positive samples), Specificity (correctly classified negative samples / true negative samples), and Accuracy (correctly classified samples / total samples). The performance of the 2D-SOM-PINT increases with the number of neurons and grid sizes. The 2D-SOM-PINT gets the best performance using a grid cell of 9x13 and, in consequence, 27x39 neurons for the different datasets. Moreover, the best performance for Inria is achieved using a 5x7 grid. These results show that, for human behaviour analysis, even a 1x1 (no sampling) grid size, the 2D-SOM-PINT provide good results. For the Frontal dataset, the neural network is able to detect the 91% of the action performed in the scene. Hence, our representation is able to recognize the behaviour of the people in the CAVIAR database (Corridor, Frontal and Inria) with great accuracy.

Table 1: Classification performance for different grid sizes and datasets

Dataset	Grid	Neurons	Sens.	Spec.	Acc.
Corridor	1x1	3x3	0.7333	0.9556	0.9238
	3x5	9x25	0.8333	0.9722	0.9524
	5x7	15x21	0.8548	0.9758	0.9585
	7x11	21x33	0.8619	0.9770	0.9605
	9x13	27x39	**0.8786**	**0.9798**	**0.9653**
Frontal	1x1	3x3	0.9167	0.9873	0.9782
	3x5	9x25	0.9381	0.9897	0.9823
	5x7	15x21	0.9333	0.9889	0.9810
	7x11	21x33	0.9333	0.9849	0.9741
	9x13	27x39	**0.9476**	**0.9913**	**0.9850**
Inria	1x1	3x3	0.7375	0.9125	0.8688
	3x5	9x25	0.7688	0.9229	0.8844
	5x7	15x21	**0.7813**	**0.9271**	**0.8906**
	7x11	21x33	**0.7813**	**0.9271**	**0.8906**
	9x13	27x39	**0.7813**	**0.9271**	**0.8906**
Average	1x1	3x3	0.7958	0.9518	0.9236
	3x5	9x25	0.8467	0.9616	0.9397
	5x7	15x21	0.8565	0.9639	0.9434
	7x11	21x33	0.8588	0.9630	0.9418
	9x13	27x39	**0.8692**	**0.9661**	**0.9470**

In order to study in depth the classification accuracy according to the behaviour, confusion matrices are presented in Tables 2, 3 and 4 for the Corridor, Frontal and Inria datasets respectively and for the different number of studied neurons. Matrix columns represent the true classes, and rows represent the classifier prediction. The ideal classifiers will have only non-zero numbers in the main diagonal. In general, as shown in the confusion matrices, 2D-SOM-PINT has a high accuracy in classifying for each behaviour and dataset. Specifically, the 2D-SOM-PINT results for the Corridor dataset (Table 2) show that SHRE (shop reenter) is the best classified irrespectively the grid size because it is the most different trajectory among the whole possible tested paths. WISH (window shop) and BROW (browsing) have a high classification rate using a small grid size (from 3x5). Using a grid size greater or equal 5x7 assures a high

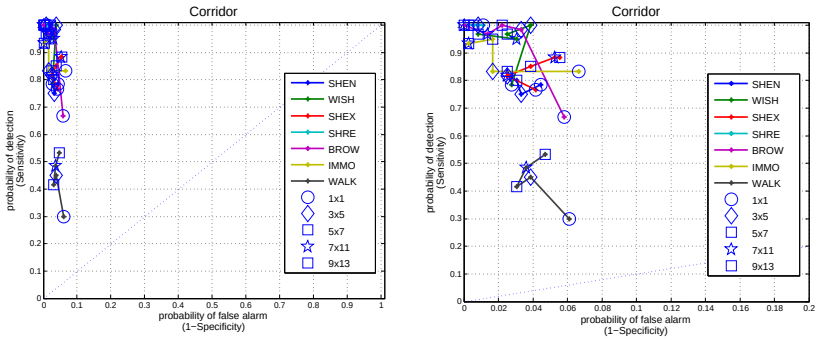
accuracy for all behaviour except for WALK (walking). Walking is the worst behaviour classified for all grid sizes. Although, using a grid size of 9x13, in which the 2D-SOM-PINT is able to classify more than 50% of samples, the problem is that all trajectories, except immobile, have walking component. Then, the classifier is not able to distinguish properly between the generic walk and a specific walk for another action. Moreover, WISH and BROW are a priori similar behaviours but 2D-SOM-PINT is able to correctly represent both.

The results of the 2D-SOM-PINT for the Frontal dataset (Table 3) are very similar to those for the Corridor dataset (both belong to the Shopping Centre dataset). As in the Corridor, SHRE (shop reenter) is the best classified irrespectively the grid size because it is the most different trajectory. WISH (window shop) has the highest probability of detection regardless the grid size. However in this dataset, BROW (browsing) has a classification rate slightly less than for the Corridor (in average, it is 7% below). In the same way, the SHEX (shop exit) is classified better using this dataset achieving the best results for 3x5, 5x7 and 9x13 grid sizes, being in average a 13% better than in the Corridor dataset. Regardless the grid size, the 2D-SOM-PINT assures a high accuracy for all behaviours (close to 90%) except for WALK (walking). Againg, Walking is the worst behaviour classified for all grid sizes but is classified about 30% better in average than in the Corridor dataset. For grid size of 9x13, the 2D-SOM-PINT is able to classify more than 77% of samples. The problem for this behaviour is the same as detailed before: the classifier is not able to distinguish properly between the generic walk and a specific walk for another action. Finally, the INMO (Inmobile) behaviour achieves the best classification rate for this dataset, outperforming about 10% in average the Corridor dataset. In general, the performance results of the 2D-SOM-PINT obtained for Corridor and Frontal datasets, recognising specific behaviour, are very similar but not the same. The use of the ground plane, in which people are moving, to extract descriptors minimizes the effect of the point of view of the camera. However, the ground plane used in the experiments (and in the datasets) are different. The Frontal dataset is a subset of the space and people considered in the Corridor dataset.

Finally, the 2D-SOM-PINT results for the Inria dataset (Table 4) show that DRDO (Drop down) is the best classified irrespectively the grid size because it is the most different trajectory compared to the others. For a 5x7 grid size, BROW achieves the high classification rate. However the classification rate decreases from this point. Similar performance results are obtained for the WALK (walking) behaviour having the high classification rate for a 1x1 grid size (73%) and decreases for bigger sizes. Finally, INMO (inmobile) is the worst classified behaviour, even below 50% for 1x1 and 3x5 grid sizes. In average, the 2D-SOM-PINT achieves only the 53% for this behaviour. These results are determined for the characteristics of this dataset: it only contains 4 behaviours and they are very unbalanced.

Additonally, the results are presented using the ROC space in Figures 10, 11 and 12 to analyze the effect of the grid size on the performance for

the different behaviours for the Corridor, Frontal and Inria datasets. The 2D-SOM-PINT is able to improve the performance on classifying the behaviours according the increasing the number of cells of the grid. The results in the ROC space confirm visually this fact. The greater the size, the better the classification performance: jointly increasing the probability of detection and decreasing the probability of false alarm. For the Shopping Center datasets, the probability of false alarms are very low, less than 7% for the Corridor and less than about 4% for the Frontal dataset. According to the specific behaviours, for SHEX and WALK in the Corridor and Frontal and also for SHEX in Frontal dataset the trend is to increase the sensitivity but increasing the probability of false alarm. Increasing the grid cell, the WALK behaviour is labelled as SHEX increasing the false alarm of this. In the same way, mainly SHEN and SHEX behaviours are labelled as WALK incresing the fase alarm of the latter. This occurs due to the SHEN and SHEX is close to the WALK behaviour differing is just a part of the trajectory close to the doors of the shop. Fig. 12b shows the trends for the Inria dataset. BROW and DRDO reflect the general performance (the greater the size, the best the classification performance). However, the trends for INMO and WALK is to increase the sensitivity but increasing the probability of false alarm (achieves about 16% for WALK).



(a) (b)

Fig. 10: Performance of the 2D-SOM-PINT in the ROC space for each behaviour according to different grid sizes using the Corridor dataset. The range for the probability of false alarm is represented in (0,1) (a) and in (0,0.2) (b)

In order to compare the advantages of preserving the topology for the extracted features, the 2D-SOM-PINT proposal has been compared with the results obtained by a SOM and a NG approach (see Table 5). The original SOM is the basis of the 2D-SOM-PINT proposal. It is able to preserve the topology in the map but it is not able to preserve the input topology. The NG is able to preserve the space defined by the feature vectors. Specifically, the parameters for the SOM and NG are the same, having 225 neurons with a

Table 2: Confusion matrices for each trained 2D-SOM-PINT for Corridor dataset. Columns represent the actual class and rows the predicted class

Grid size 1x1

SHEN	WISH	SHEX	SHRE	BROW	INMO	WALK
78%	3%	2%		5%	3%	13%
7%	78%			2%	5%	3%
		77%		10%		15%
			100%			7%
3%	3%	5%		67%	7%	17%
3%	5%	5%		12%	83%	15%
8%	10%	12%		5%	2%	30%

Grid size 3x5

SHEN	WISH	SHEX	SHRE	BROW	INMO	WALK
75%		2%				18%
8%	100%	2%			10%	3%
		82%		2%		13%
		2%	100%		2%	2%
3%				98%	2%	15%
5%		2%			83%	3%
8%		12%			3%	45%

Grid size 5x7

SHEN	WISH	SHEX	SHRE	BROW	INMO	WALK
80%	2%	2%			2%	13%
5%	97%					10%
2%		85%			2%	20%
		3%	100%			
3%	2%	2%		100%		7%
2%					95%	8%
8%		8%			2%	42%

Grid size 7x11

SHEN	WISH	SHEX	SHRE	BROW	INMO	WALK
82%	2%	2%			2%	10%
5%	95%			2%		12%
2%		88%		2%	3%	25%
			100%			
5%				97%		3%
					93%	2%
7%	3%	10%			2%	48%

Grid size 9x13

SHEN	WISH	SHEX	SHRE	BROW	INMO	WALK
83%					2%	13%
3%	97%					2%
2%		88%			2%	30%
			100%			2%
				100%		
2%					93%	
10%	3%	12%			3%	53%

Average

SHEN	WISH	SHEX	SHRE	BROW	INMO	WALK
80%	2%	2%		5%	2%	14%
6%	93%	2%		2%	8%	6%
2%		84%		4%	2%	21%
		3%	100%		2%	3%
4%	3%	3%		92%	4%	10%
3%	5%	3%		12%	90%	7%
8%	6%	11%		5%	2%	44%

Table 3: Confusion matrices for each trained 2D-SOM-PINT for Frontal dataset. Columns represent the actual class and rows the predicted class

Grid size 1x1

SHEN	WISH	SHEX	SHRE	BROW	INMO	WALK
90%				3%		
3%	100%			3%		
		87%				
		3%	100%			
				77%		7%
3%		7%		3%	100%	
3%		3%		13%		93%

Grid size 3x5

SHEN	WISH	SHEX	SHRE	BROW	INMO	WALK
90%						13%
7%	100%					
		100%		7%		10%
			100%			
				87%		3%
				3%	100%	
3%				3%		73%

Grid size 5x7

SHEN	WISH	SHEX	SHRE	BROW	INMO	WALK
97%						17%
3%	100%			3%		7%
		100%		7%		3%
			100%			
				87%		3%
					100%	
				3%		70%

Grid size 7x11

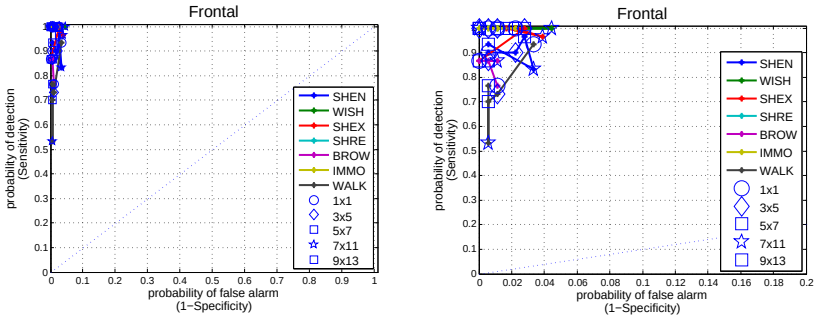
SHEN	WISH	SHEX	SHRE	BROW	INMO	WALK
83%				3%		17%
17%	100%					10%
		97%		7%		17%
			100%			
		3%		87%		3%
					100%	
				3%		53%

Grid size 9x13

SHEN	WISH	SHEX	SHRE	BROW	INMO	WALK
93%						3%
7%	100%			3%		7%
		100%		7%		7%
			100%			
				87%		
					100%	7%
				3%		77%

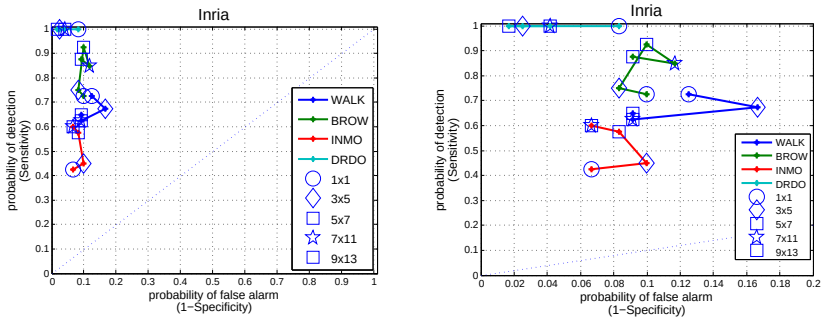
Average

SHEN	WISH	SHEX	SHRE	BROW	INMO	WALK
91%				3%		13%
7%	100%			3%		8%
		97%		7%		9%
		3%	100%			
		3%		85%		4%
3%		7%		3%	100%	7%
3%		3%		5%		73%



(a) (b)

Fig. 11: Performance of the 2D-SOM-PINT in the ROC space for each behaviour according to different grid sizes using the Frontal dataset. The range for the probability of false alarm is represented in (0,1) (a) and in (0,0.2) (b)



(a) (b)

Fig. 12: Performance of the 2D-SOM-PINT in the ROC space for each behaviour according to different grid sizes using the Inria dataset. The range for the probability of false alarm is represented in (0,1) (a) and in (0,0.2) (b)

Gaussian neighbourhood function to preserve the topological properties. The SOM has a map grid size of 1515 using a toroidal shape. They are trained for 50 epochs. Moreover, SOM has an additional fine-tuning for 500 epochs. The features used to train the maps are the same: the ADV. In this case, a vector concatenating all $ADV_{i,j}$ was used as the collection of ADVs describing a person trajectory (Eq. 10).

Table 5 shows the sensitivity and specificity of the compared artificial neural neworks for the different grid sizes and datasets used in the experiments. Moreover, in Fig. 13, the average sensitivity and specificity for the different datasets are graphically presented. It is important to highlight that the minimum probability of detection of the 2D-SOM-PINT is about 73% using a 1x1 grid (this means the whole ground plane was sampled in 1 ADV) for the

Table 4: Confusion matrices for each trained 2D-SOM-PINT for Inria dataset. Columns represent the actual class and rows the predicted class

Grid size 1x1			
WALK	BROW	INMO	DRDO
73%	8%	30%	
13%	73%	18%	
13%	8%	43%	
3%	13%	10%	100%

Grid size 3x5			
WALK	BROW	INMO	DRDO
68%	10%	40%	
13%	75%	13%	
18%	13%	45%	
3%	3%	3%	100%

Grid size 5x7			
WALK	BROW	INMO	DRDO
63%	3%	25%	
13%	93%	18%	
23%	3%	58%	
3%	3%		100%

Grid size 7x11			
WALK	BROW	INMO	DRDO
58%	8%	23%	
15%	85%	20%	
23%		58%	
5%	8%		100%

Grid size 9x13			
WALK	BROW	INMO	DRDO
65%	3%	25%	
13%	88%	15%	
20%		60%	
3%	10%		100%

Average			
WALK	BROW	INMO	DRDO
65%	6%	29%	
13%	83%	17%	
19%	8%	53%	
3%	7%	6%	100%

Table 5: Classification performance of 2D-SOM-PINT compared to SOM and NG

		Dataset							
		Corridor		Frontal		Inria		Average	
ANN	Grid	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
2D-SOM-PINT	1x1	0.7333	0.9556	0.9167	0.9873	0.7375	0.9125	0.7958	0.9518
	3x5	0.8333	0.9722	0.9381	0.9897	0.7688	0.9229	0.8467	0.9616
	5x7	0.8548	0.9758	0.9333	0.9889	**0.7813**	**0.9271**	0.8565	0.9639
	7x11	0.8619	0.9770	0.9333	0.9849	**0.7813**	**0.9271**	0.8588	0.9630
	9x13	**0.8786**	**0.9798**	**0.9476**	**0.9913**	**0.7813**	**0.9271**	**0.8692**	**0.9661**
SOM	1x1	0.6714	0.9429	0.8762	0.9794	0.7167	0.9056	0.7548	0.9426
	3x5	0.6714	**0.9670**	**0.9286**	**0.9881**	0.7167	0.9056	0.7722	0.9536
	5x7	0.7143	0.9571	0.8905	0.9817	0.7750	0.9250	0.7933	0.9546
	7x11	0.6571	0.9381	0.9238	0.9873	**0.8167**	**0.9389**	**0.7992**	**0.9548**
	9x13	0.6300	0.91	**0.9286**	**0.9881**	0.7667	0.9222	0.7751	0.9401
NG	1x1	0.7429	0.9643	0.8952	0.9825	0.7417	0.9139	0.7933	0.9536
	3x5	**0.7857**	0.9643	**0.9571**	**0.9929**	0.7417	0.9139	**0.8282**	0.9570
	5x7	0.7714	**0.981**	0.9095	0.9849	0.7750	0.9250	0.8186	**0.9636**
	7x11	0.7000	0.9667	0.9190	0.9865	**0.8000**	**0.9333**	0.8063	0.9622
	9x13	0.6700	0.9445	0.9190	0.9865	0.7667	0.9222	0.7852	0.9511

Corridor dataset. This case uses just 3x3 neurons (see Figure 9a) to represent each behaviour. In total, the 2D-SOM-PINT scheme uses 63 neurons (9 by the 7 behaviours) and 84 connections (12 by the 7 behaviours) to achieve these results. Comparatively, 2D-SOM-PINT is 6% better than SOM in sensitivity and 1% in specificity for this dataset. However, for this structure, the 2D-SOM-PINT is about 1% worst than NG in sensitivity and in specificity. To

achieve these results the NG is using 162 neurons and 523 synapses more than 2D-SOM-PINT. The minimum probability of detection of the 2D-SOM-PINT for the Frontal and Inria datasets are about 91% and 73% using a 1x1 grid. This means 4% better compared to SOM and 2% to NG using the Frontal dataset and 4% better compared to SOM and less than 1% worse compared to NG using the Inria dataset.

The SOM and NG classifiers are able to achieve the best results for a 5x7 (71.43% sensitivity and 95.71% specificity) and 3x5 (78.57% sensitivity and 96.43% specificity) grid size respectively for the Corridor dataset. For the Frontal dataset, a grid size of 3x5 provide the best results for SOM and NG, being the 7x11 the best size for the Inria dataset However, the 2D-SOM-PINT achieve the best results using the biggest grid size (9x13).

In average (see Fig. 13) SOM performance increases up to a 7x11 grid size. However, the performance decreases for the biggest grid size 9x13. Comparatively to 2D-SOM-PINT, the performance gap is about 6% and 1% for sensitivity and specificity respectively up to 7x11, being about 10% and 2.5% for sensitivity and specificity respectively using the 9x13 grid size. In the NG case, the performance decreases as grid size increases even before (from 3x5 grid size), being the gap for a 9x13 grid about 8% in sensitivity and 1.5% in specificity. The problem is that the SOM and NG are not able to preserve the topological information about cell grids as they are converted to a vector for training and classification. The neighbourhood for a cell is not preserved due to each component of the vector reference associated to each neuron is compared to corresponding component of the input vector. Learning process for the SOM and NG takes into account the neighbourhood close in the space R^{nxmx5} but does not take into account the neighbourhood in the ground space. However, the structure of the 2D-SOM-PINT assure the neighbours for a specific neuron are close to adjacent cells in the ground plane.

The 2D-SOM-PINT proposal has been compared to other contemporary methods in order to show the accuracy of the proposed representation and classification method to include behaviour information. Sensitivity and specificity results of context classification have been calculated from reported success rates in [11], [39] and [21] of comparable experiments on the Corridor dataset (results for the other datasets are not available). These methods are grouped as state and semantic models using predefined models and rules to evaluate behaviours. Additionally, 2D-SOM-PINT has been compared to our previous work [5] in order to show the advantages of the topology preservation using the self-organizing proposal.

In [11], a rule-based approach, used semantic rules on both the role and movement classifications to evaluate the context from video sequences. The work in [39], used an extension of the HMM, specifically, to interpret the context, hidden semi-Markov model (HSMM). HSMMs extend the standard Hidden Markov model with an explicit duration model for each state [12]. Finally, in [21] Lavee et al. proposed the use of Petri Nets (PN) for recognition of event occurrences in video. The Petri Net was used to express semantic knowledge about the event domain as well as for recognizing events as they occur in a

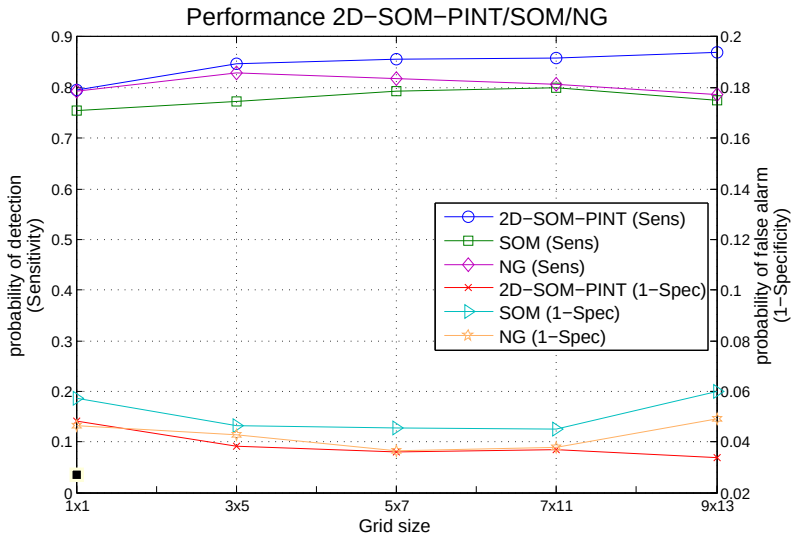


Fig. 13: Comparative performance between 2D-SOM-PINT, SOM and GNG for different grid sizes

particular video sequence. In our previous work, the ADV was used as a vector input for different classic classifiers including SOM, Supervised SOM, Neural Gas, Linear Discriminant Analysis and K-nearest neighbour and multiclassfier (MC) taking into account the output behaviour of each individual classifier. Table 6 shows results for the above four methods (Rule-based, HSMM, PN, and the previous multiclassifier (MC)) for the ADV representation using a 5x7 grid. Results for the proposed 2D-SOM-PINT with 5x7 and 9x13 grid size has been included in the comparison.

Table 6: Classification performance comparison for the Corridor dataset

Method	Sensitivity	Specificity	Dif. Sens	Dif. Spe
Rule-based [12]	0,570	N/A	31%	N/A
HSMM [39]	0,651	0,987	23%	-1%
PN [21]	0,809	0,968	7%	1%
MC 5x7 [5]	0,814	**0,988**	6%	-1%
2D-SOM-PINT 5x7	0,855	0,976	2%	0%
2D-SOM-PINT 9x13	**0,879**	0,980		

As it is shown in Table 6, the 2D-SOM-PINT approach achieves a significant improvement over both the Rule-based and the HSMM results for sensitivity. Also, for PN the improvement is close to 8%. Additionally, compared to the MC our novel proposal can achieve an extra 6% more due to preserving input topology. The 2D-SOM-PINT representation and classification

outperform the results without having semantic knowledge about behaviour and preserving input topology.

Finally, the last experiment is focused on validate the 2D-SOM-PINT proposal in other domain than the individual trajectory interpretation and to conduct experiments with descriptors different to the ADV. In this case, recognising group activities has been selected as another domain of study. Understanding small group behaviour, interacting with the environment (moving to a specific place), with other people (fighting), or group changes (split) are a very challenging problem. It shares with the previous domain the activity recognition and the trajectories as basic input to calculate descriptors but in this case is based on groups of interacting people. In consequence, we would like to show how the 2D-SOM-PINT is able not only in the learning and classification process to include topological relations between the points or regions where features were extracted but also the relationships among them using other descriptors than those extracted from single trajectories.

In this experiment, we make use of the trajectory described by the group and by the individuals who form it as basic input to describe the group activity. Specifically, the features describes the group activity by using three different components: the trajectory described by the centroids of the group over time (calculated as the ADV), the coherence of the movements of each person with respect to the movement of the centroid in a specific group ($IntraGD$) and, finally, the movement relationships in terms of directions among different groups in the scene ($InterGD$). This descriptor consider the regions where features are extracted in order to incorporate the topological relations among them. Specifically, for this experiment, the grid sizes of the cell C considered are 1x1, 3x3, 5x5, 7x7 and 9x9.

Experiments have been carried out using the public BEHAVE [8] and the group characteristics of the previous used CAVIAR dataset. For the INRIA sequences, three different behaviours have been labelled: Walking, Browsing, Inmobile and DropDown. In the case of Shopping Center, the trajectories have been used as samples classified into 4 contexts or activities: ShopEnter, ShopExit, Meeting and Walking. The BEHAVE public dataset provides the group information in images of 640x480 pixels at a frame rate of 25 fps. Kim and Cho et al. [19,10] have demonstrated a very good performance using this dataset. In order to be able to compare with them, we have used the same activities they do. Specifically, six activities: Approach, Split, WalkTogether, RunTogether, Fighting, and InGroup.

Table 7 presents the average Sensitivity and Specificity according to the classifier (columns) and each dataset (rows) for the different grid sizes and classifiers. In this case, the 2D-SOM-PINT, which preserves the topological information of the input data, gets the best results. It is important to highlight that the NG also offer a good performance rate.

Finally, we compare the 2D-SOM-PINT results with the methods proposed in [10,43,30,42] considering the seven classes of the Behave. Only [10] considers the seven classes as well. The rest of the works consider a subset of four classes. Table 8 shows the Sensitivity of the comparison. As we can see,

Table 7: Classification performance for different classifiers

	2D-SOM-PINT		**SOM**		**NG**	
Dataset	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
Behave	**0.837**	**0.967**	0.836	0.966	**0.837**	**0.967**
Corridor	**0.790**	**0.930**	0.783	0.928	0.788	0.929
Frontal	**0.783**	**0.928**	0.783	**0.928**	0.782	0.927
Inria	**0.843**	**0.922**	0.839	0.919	0.840	0.920
TOTAL	**0.813**	**0.937**	0.810	0.935	0.812	0.936

our proposal achieves a performance better than [30]. For the comparison of [43] and [42], the 2D-SOM-PINT shows slightly lower performance for WalkTogether and InGroup and better for the others. The comparison with the state-of-the-art method [10] shows the better performance of our method for all activities except for InGroup (6% lower). In average, the 2D-SOM-PINT classifier outperforms all compared methods.

Table 8: Comparison with other state-of-the-art methods for the Behave dataset

	2D-SOM-PINT	[10]	[43]	[30]	[42]
Approach	**100**	83.33	71	60	
Split	**100**	**100**	79	70	93.1
WalkTogether	86.67	91.66	88	45	**92.1**
InGroup	86.67	**100**	88	90	94.3
Fight	90	83.33			**95.1**
RunTogether	**100**	83.33			
Average	**93.89**	90.275	81.5	66.25	93.65

5 Conclusion

In this paper a constrained self-organizing neural network nD-SOM-PINT has been proposed to preserve the location of the extracted features from an n-dimensional space and its topological information. It is able to represent and classify features preserving the topological relations of the input space in which the features have been extracted. The nD-SOM-PINT is a variant of the self-organizing feature map in which the learning and the training process is constrained by the topology of the n-dimensional space used to extract the input features. Extracted features in adjacent areas of the n-dimensional space are explicitly in adjacent areas of the nD-SOM-PINT map. This self-organizing scheme contains specific neurons for specific regions of the input space. Each input to the map is able to activate a neuron of each area associated to the cell at the same time. It allow us to adapt the whole map according to an input sample. After training, a specific nD-SOM-PINT represents a higher level

of understanding, a label. In classification, a sample is compared in parallel to each nD-SOM-PINT to establish the winner map and, consequently, the associated high-level of understanding.

The nD-SOM-PINT is evaluated in this paper by a case of study aimed to represent and classify trajectories of people from video sequences into high level of semantic understanding: human behaviour analysis. The instantiated neural network, 2D-SOM-PINT, is able to deal with the big gap between human trajectories in a scene and the global behaviour associated to them preserving the spatial information about trajectories. It uses our previous Activity Description Vector (ADV) [5] as features to describe the activity happened in each region of the scene (cells). The network is able to learn the behaviour by means of learning activities happening in each specific area of the scene. The 2D-SOM-PINT contains specific neurons for each cell preserving the topology of the scene. Experimental results show how 2D-SOM-PINT proposal is able to classify input ADVs into human behaviour in complex situations with great accuracy outperforming the original SOM and the NG (as network preserving input space topology). Moreover, it is able to outperform previous methods that uses the same dataset (Corridor in CAVIAR). Moreover, a brief example of use of the 2D-SOM-PINT for another domain than the individual trajectory interpretation has been presented. Specifically, a experiment to broadly show how the 2D-SOM-PINT is able not only in the learning and classification process to include topological relations between the points or regions where features were extracted but also the relationships among them using other descriptors than those extracted from single trajectories has been included. All experiments show that the 2D-SOM-PINT is able to improve the results obtained with other self-organizing neural networks using the same descriptor and, also, outperforms the results obtained with other descriptors and state-of-the-art methods using the same datesets.

As future research lines, we plan to establish a hierarchical scheme for classification purposes in order to improve the performance of the nD-SOM-PINT. Instead of compare each input with the maps of the same size corresponding to each label and selecting the map of minimum distance, we propose comparing the input through different levels increasing the size of the grid cell until the difference of distances between the winner map and the second one be enough (could be a threshold). For example in the case of study of this paper, a map for a grid size of 1x1 is enough to detect a shop re-enter (SHRE) behaviour, having to increase the size to properly detect other behaviours. In the same way in order to decrease the number of neurons associated to each region of the n-dimensional space, we propose a growing scheme (as GG or GC) to associate different neurons depending on the information contained in each region of the space. For example in the case of study, if the space (image) is discretized using many cells, there are regions with the same characteristics than the neigbourhood, even areas with no-activity. Finally, we propose to extend the experiments in order to explore the feasibility of nD-SOM-PINT to represent and recognize the 3D space. We plan to use the map to represent and

analyse 3D trajectories extracted from RGB-D cameras in order to classfify different actions in a 3D human-computer interaction system.

References

1. Anjum, N., Cavallaro, A.: Single camera calibration for trajectory-based behavior analysis 2 Ground-plane calibration pp. 147–152 (2007)
2. Anjum, N., Cavallaro, A.: Multifeature Object Trajectory Clustering for Video Analysis **18**(11), 1555–1564 (2008)
3. Anjum, N., Cavallaro, A.: Trajectory Clustering for Scene Context Learning and Outlier Detection. In: D. Schonfeld, C. Shan, D. Tao, L. Wang (eds.) Video Search and Mining, vol. 287, pp. 33–51. Springer Berlin Heidelberg (2010). DOI 10.1007/978-3-642-12900-1_2
4. Antonakaki, P., Kosmopoulos, D., Perantonis, S.J.: Detecting abnormal human behaviour using multiple cameras. Signal Processing **89**(9), 1723–1738 (2009)
5. Azorin-Lopez, J., Saval-Calvo, M., Fuster-Guillo, A., Garcia-Rodriguez, J.: Human Behaviour Recognition based on Trajectory Analysis using Neural Networks. In: International joint conference in neural networks, 2013 (2013)
6. Azorin-Lopez, J., Saval-Calvo, M., Fuster-Guillo, A., Garcia-Rodriguez, J.: A novel prediction method for early recognition of global human behaviour in image sequences. Neural Processing Letters pp. 1–25 (2015). DOI 10.1007/s11063-015-9412-y. URL http://dx.doi.org/10.1007/s11063-015-9412-y
7. Azorin-Lopez, J., Saval-Calvo, M., Fuster-Guillo, A., Oliver-Albert, A.: A predictive model for recognizing human behaviour based on trajectory representation. In: 2014 International Joint Conference on Neural Networks, IJCNN 2014, Beijing, China, July 6-11, 2014, pp. 1494–1501 (2014)
8. Blunsden, S., Fisher, R.B.: The BEHAVE video dataset: ground truthed video for multi-person behavior classification. Annals of the BMVA **4** (2010)
9. Brown, M., Lowe, D.G.: Unsupervised 3d object recognition and reconstruction in unordered datasets. In: 3-D Digital Imaging and Modeling, 2005. 3DIM 2005. Fifth International Conference on, pp. 56–63. IEEE (2005)
10. Cho, N.G., Kim, Y.J., Park, U., Park, J.S., Lee, S.W.: Group Activity Recognition with Group Interaction Zone Based on Relative Distance Between Human Objects. International Journal of Pattern Recognition and Artificial Intelligence **29**, 1555,007 (2015). DOI 10.1142/S0218001415550071. URL http://www.worldscientific.com/doi/10.1142/S0218001415550071
11. Fisher, R., Santos-Victor, J., Crowley, J.: CAVIAR Hidden Semi-Markov Model Behaviour Recognition (2005). URL http://homepages.inf.ed.ac.uk/rbf/CAVIAR/hsmm.htm
12. Fisher, R.B.: The PETS04 Surveillance Ground-Truth Data Sets. In: Sixth IEEE Int. Work. on Performance Evaluation of Tracking and Surveillance (PETS04), pp. 1 – 5 (2004)
13. Fritzke, B.: Growing cell structuresa self-organizing network for unsupervised and supervised learning. Neural networks **7**(9), 1441–1460 (1994)
14. Fritzke, B.: Growing grida self-organizing network with constant neighborhood range and adaptation strength. Neural Processing Letters **2**(5), 9–13 (1995)
15. Fritzke, B., et al.: A growing neural gas network learns topologies. Advances in neural information processing systems **7**, 625–632 (1995)
16. Hu, W., Xie, D., Tan, T., Maybank, S.: Learning activity patterns using fuzzy self-organizing neural network. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on **34**(3), 1618–1626 (2004). DOI 10.1109/TSMCB.2004.826829
17. Juan, L., Gwun, O.: A comparison of sift, pca-sift and surf. International Journal of Image Processing (IJIP) **3**(4), 143–152 (2009)
18. Kangas, J.A., Kohonen, T.K., Laaksonen, J.T.: Variants of self-organizing maps. Neural Networks, IEEE Transactions on **1**(1), 93–99 (1990)

19. Kim, Y.j., Cho, N.g., Lee, S.w.: Group Activity Recognition with Group Interaction Zone. Icpr 2014 pp. 3517–3521 (2014). DOI 10.1109/ICPR.2014.605
20. Kohonen, T.: Clustering, taxonomy, and topological maps of patterns. In: Proceedings of the 6th International Conference on Pattern Recognition, pp. 114–128. IEEE (1982)
21. Lavee, G., Rivlin, E., Rudzsky, M.: Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in video. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on **39**(5), 489–504 (2009)
22. Lee, A.C.D., Rinner, C.: Visualizing urban social change with self-organizing maps: Toronto neighbourhoods, 19962006. Habitat International **45, Part 2**(0), 92 – 98 (2015). DOI http://dx.doi.org/10.1016/j.habitatint.2014.06.027. URL http://www.sciencedirect.com/science/article/pii/S0197397514001039. Special Issue: Exploratory Spatial Analysis of Urban Habitats
23. Li, X., Hu, W., Hu, W.: A Coarse-to-Fine Strategy for Vehicle Motion Trajectory Clustering pp. 18–21 (2006)
24. Madokoro, H., Honma, K., Sato, K.: Classification of behavior patterns with trajectory analysis used for event site. In: Neural Networks (IJCNN), The 2012 International Joint Conference on, pp. 1–8 (2012). DOI 10.1109/IJCNN.2012.6252565
25. Martinetz, T., Schulten, K.: A "Neural-Gas" Network Learns Topologies. Artificial Neural Networks **I**, 397–402 (1991)
26. Martinez-Contreras, F., Orrite-Urunuela, C., Herrero-Jaraba, E., Ragheb, H., Velastin, S.a.: Recognizing Human Actions Using Silhouette-based HMM. 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance pp. 43–48 (2009). DOI 10.1109/AVSS.2009.46
27. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding **104**(2-3), 90–126 (2006)
28. Morris, B., Trivedi, M.: A survey of vision-based trajectory learning and analysis for surveillance. Circuits and Systems for Video Technology, IEEE Transactions on **18**(8), 1114–1127 (2008). DOI 10.1109/TCSVT.2008.927109
29. Morris, B., Trivedi, M.: Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. Pattern Analysis and Machine Intelligence, IEEE Transactions on **33**(11), 2287–2301 (2011). DOI 10.1109/TPAMI.2011.64
30. Münch, D., Michaelsen, E., Arens, M.: Supporting fuzzy metric temporal logic based situation recognition by mean shift clustering. In: KI 2012: Advances in Artificial Intelligence, pp. 233–236. Springer (2012)
31. N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer: Smote : Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002)
32. Naftel, A., Khalid, S.: Classifying spatiotemporal object trajectories using unsupervised learning in the coefficient feature space. Multimedia Systems **12**(3), 227–238 (2006). DOI 10.1007/s00530-006-0058-5
33. Owens, J., Hunter, A.: Application of the Self-Organising Map to Trajectory Classification pp. 1–7 (2000)
34. Parisi, G., Wermter, S.: Hierarchical som-based detection of novel behavior for 3d human tracking. In: Neural Networks (IJCNN), The 2013 International Joint Conference on, pp. 1–8 (2013). DOI 10.1109/IJCNN.2013.6706727
35. Saul, H., Kozempel, K., Haberjahn, M.: A comparison of methods for detecting atypical trajectories. Urban Transport XX **138**, 393 (2014)
36. Saval-Calvo, M., Azorin-Lopez, J., Fuster-Guillo, A., Mora-Mora, H.: μ-mar: Multiplane 3d marker based registration for depth-sensing cameras. Expert Systems with Applications **42**(23), 9353–9365 (2015)
37. Schreck, T., Bernard, J., von Landesberger, T., Kohlhammer, J.: Visual cluster analysis of trajectory data with interactive Kohonen maps. Information Visualization **8**(1), 14–29 (2009). DOI 10.1057/ivs.2008.29
38. Turaga, P., Chellappa, R., Subrahmanian, V., Udrea, O.: Machine Recognition of Human Activities: A Survey. IEEE Transactions on Circuits and Systems for Video Technology **18**(11), 1473–1488 (2008)

39. Tweed, D., Fisher, R., Bins, J., List, T.: Efficient hidden semi-markov model inference for structured video sequences. In: Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on, pp. 247–254 (2005)
40. Uriarte, E.A., Martín, F.D.: Topology preservation in som. International journal of applied mathematics and computer sciences **1**(1), 19–22 (2005)
41. Villmann, T., Der, R., Herrmann, M., Martinetz, T.M.: Topology preservation in self-organizing feature maps: exact definition and measurement. Neural Networks, IEEE Transactions on **8**(2), 256–266 (1997)
42. Yin, Y., Yang, G., Man, H.: Small Human Group Detection and Event Representation Based on Cognitive Semantics. 2013 IEEE Seventh International Conference on Semantic Computing pp. 64–69 (2013). DOI 10.1109/ICSC.2013.20. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6693495
43. Zhang, C., Yang, X., Lin, W., Zhu, J.: Recognizing Human Group Behaviors with Multi-group Causalities. 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology pp. 44–48 (2012). DOI 10.1109/WI-IAT.2012.162. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6511646