A bio-inspired knowledge representation method for anomaly detection in cognitive video surveillance systems

Simone Chiappino, Pietro Morerio, Lucio Marcenaro, and Carlo S. Regazzoni
DITEN
University of Genoa
Via Opera Pia 11A 16145 Genoa - Italy
Email: {s.chiappino, p.morerio}@ginevra.dibe.unige.it
{lucio.marcenaro, carlo.regazzoni}@unige.it

Abstract—Human behaviour analysis has important applications in the field of anomaly management, such as Intelligent Video Surveillance (IVS). As the number of individuals in a scene increases, however, new macroscopic complex behaviours emerge from the underlying interaction network among multiple agents. This phenomenon has lately been investigated by modelling such interaction through Social Forces.

In most recent Intelligent Video Surveillance systems, mechanisms to support human decisions are integrated in cognitive artificial processes. These algorithms mainly address the problem of modelling behaviours to allow for inference and prediction over the environment. A bio-inspired structure is here proposed, which is able to encode and synthesize signals, not only for the description of single entities behaviours, but also for modelling cause-effect relationships between user actions and changes in environment configurations (i.e. the crowd). Such models are stored within a memory during a learning phase. Here the system operates an effective knowledge transfer from a human operator towards an automatic systems called Cognitive Surveillance Node (CSN), which is part of a complex cognitive JDL-based and bio-inspired architecture. After such a knowledge-transfer phase, learned representations can be used, at different levels, either to support human decisions by detecting anomalous interaction models and thus compensating for human shortcomings, or, in an automatic decision scenario, to identify anomalous patterns and choose the best strategy to preserve stability of the entire system.

Results are presented, where crowd behaviour is modelled by means of Social Forces and can interact with a human operator within a visual 3D simulator. The way anomalies are detected and consequently handled is demonstrated on synthetic data and also on a real video sequence, in both the user-support and automatic modes.

Index Terms—Cognitive systems, Bio-inspired learning, Anomalous interactions, Crowd monitoring

I. INTRODUCTION

Several works have been devoted in the last decade to link traditional computer vision tasks to high-level context aware functionalities such as scene understanding, behaviour analysis, interaction classification or recognition of possible threats or dangerous situations [1], [2].

Among the several disciplines which are involved in the design of next generation security and safety systems, cognitive sciences represent one of the most promising in terms of capability of provoking improvements with respect to state of the art, [3]. As a matter of fact, several recent studies have proposed the application of smart functionalities to camera and sensor networks in order to move from object recognition paradigm to event/situation recognition one [4]. Such a trend change has substantial implications for what concerns the processing of signals, as it will be shown throughout this work. The application of bio-inspired models to safety and security tasks represents a relevant added value. In fact, such models enhance the capability not only of detecting the presence of an intruder in a forbidden area or recognizing the trajectory of an object in an urban scenario (e.g. a baggage in a station or a car on the road) but also of interpreting the behaviour of the entity, or properly selecting events of interest with respect to normal situations, or even to automatically take decision and perform action on the environment.

The application of neurobiological sciences to the field of cognitive radar and cognitive radios lately led to the rise of a new broad discipline which was formalized in some works by S. Haykin [5] under the name of Cognitive Dynamic Systems, These works eventually gather and synthesize some of the main intuition of the last decades in this field. A working definition of Cognitive Dynamic Systems is given:

Cognitive dynamic systems build up rules of behaviour over time through learning from continuous experiential interactions with the environment, and thereby deal with environmental uncertainties.

The underlying hidden assumption behind the formalization of this discipline is that animal and human brains are the best cognitive systems on the market and are thus to be emulated.

In a previous work [6], the features of a cognitive architecture, motivated by the work of neurophysiologist Damasio [7] and based on the JDL model [8], were described. The application of the proposed framework to crowd analysis is here presented. In a video surveillance scenario, the proposed
Cognitive Node can be applied to the crowd analysis domain in order to identify patterns that deviate from expected behaviour: an abnormal behaviour is defined as any kind of deviation from central tendencies defined as normality condition. The cognitive node operating mode is made up of a learning and a detection phases. During the learning period the cognitive node stores the observed interactions between human operator actions and the resulting crowd state changes. It is important to note that the human actions acquired are devoted to avoid abnormal situation, e.g. overcrowding or abnormal flow directions. The automatic system is able to effectively learn representations of normal user-environment relationships for standard crowd behaviour maintenance through the aforementioned data structure and architecture. After such a knowledge acquisition phase, learned representations can be used at two different levels: first, to support human decisions by detecting anomalous crowd-operator interactions and compensating for human shortcomings; secondly, in an automatic decision scenario, to autonomously identify anomalous crowd-environment configurations and choose the best strategy to preserve stability of the entire system (i.e. a proper security level in the monitored area) by putting in action effective countermeasures. Many video analysis algorithms have been developed in order to identify crowd behaviours and classify the interactions among different entities. For instance, in [9] a method for crowd behaviour analysis based on social forces and optical flow is proposed. In [10] a video analysis technique for tracking of multiple interacting objects is presented. More recently, in [11] the authors present an innovative method based on people flow estimation. A new abstract viscous fluid field is proposed in [12] for detecting crowd events.

The main contribution of this paper is to propose and develop an innovative cognitive video surveillance system, which is able to detect anomalies by learning behavioural models from observations of crowd evolution and consequent human operator (re)actions. The system acquires the crowding states, by video analysis techniques, and it receives from the user his countermeasures, in order to maintain stability and to avoid abnormal situations. This knowledge (i.e. models of normal interactions) is transferred from human operator to the system, providing the system with crowding dynamic models augmented by user actions. A simulated crowd monitoring environment have been used for training and testing. The issue of modelling and simulating crowds will be discussed and motivated, as it represents a central matter in applying the theory which will be presented. The use of a simulator was necessary in order to gather enough data for training and testing, as video sequences of the desired kind are not available for training. A simple cognitive node application on a real video sequence is presented in order to show the capabilities of the system, which is however trained by simulated data.

The remaining of this work is organized as follows: section II gives a quick insight on the some issues rising in analysing and simulating crowds. Section III gives an overview of the cognitive system proposed in [6]. The applications of such models to crowd monitoring are presented in section IV. Section V describes the proposed approach for anomaly detections, while the results are given in section VI. Conclusions are drawn in section VII.

II. Crowd

As it will become clearer throughout the paper, the issue of simulating a crowd is a necessary step. This is basically due to two facts. On the one hand, we face the pressing need of having some kind of interaction between our system and the monitored people. Such an interaction cannot occur between the system and a recorded video sequence, where a crowd would have some clearly still behaviour and could not change its flow based on external instructions. We have been trying however to give more consistency to our work by showing some results in a situation of crowd flowing in a public dataset (section VI-A). On the other hand, arranging a real experimental setup of the desired type would imply hiring a huge amount of cooperative individuals for a considerable time period for each trial, which turns out to be hardly achievable in practice, considering the high number of trials to be run. Therefore, being quite essential to make use of a simulator, its realism become a central issue. Although not being strictly in the aim of this work, such an issue deserves a quick insight, as experimental results will rely on simulated data coming from simulated sensors.

Crowds need to be given an underlying dynamical model in order to be simulated. Actually, such a model is inherently in charge of depicting the evolution of some crowd features only. This raises the issue of how to describe crowds. Such a description includes a selection not only of the features one is interested in simulating, but also of the scale at which the model has to lie, in order to effectively describe the formers. Namely, a microscopic model can be given the task of simulating features at a more global level, while the opposite way is hardly practicable as it can be easily guessed. As a feature, we are interested in simulating the number of people in a selected environment of the given scenario. A realistic visual output is also needed in order for a plausible interaction with a human being to take place.

The dynamical model we have been employing for simulation is the well known Social Force Model developed by D. Helbing [13]. This method, and some of its following improvements [14], showed its validity in the context of video analysis (e.g [15]) and were also employed in simulation (e.g. [16]). A more detailed description of the simulator developed for testing will be given at the beginning of section VI.

III. THE PROPOSED FRAMEWORK

The proposed system relies on an general abstract structure, called Cognitive Node, which was described in [6]. We shortly review its main features in the following. The CN is implemented according to Damasio’s theories, and describes cognitive entities as complex systems capable of incremental learning based on the experience of relationships between themselves and the external world. A fundamental logical block of Damasio’s model consists in the Autobiographical
Memory (AM). During a learning phase any cognitive entity builds its own AM, which can be formalized as a sequence of cause-effect relationships, i.e., interactions, between two specific brain devices called proto-self and core-self. Such devices are specifically devoted to monitor and manage respectively the internal status of an entity (proto-self) and the relationships with the other entity, such as external world (core-self). In the crowd monitoring domain the two entities can be defined as: the operator (proto-entity) and whole crowd (core-entity). During the learning off-line stage, the CN receives data from sensors. A Data Fusion module provides to the system temporally and spatially aligned multidimensional proto and core state vectors, respectively $x^P(t)$ and $x^C(t)$; while an Event Detection block extracts information related to relevant changes in the signals acquired by sensors (the meaning of relevant will become shortly clear). It is possible to define proto $P$ and core $C$ events in order to develop a specific probabilistic model (the AM), that is able to describe different kinds of relationships between proto and core entities. According to Damasio’s theory, as proposed in [17], the sequences of proto and core events can be organized into two kinds of interaction patterns called triplets in order to account for interactions between entities: $\{e^P, e^C, e^P\}$ (passive interaction) and $\{e^C, e^P, e^C\}$, (active interaction), to represent the causal relationships, in terms of initial situation (first event), cause (second event) and consequent effect on the examined entity (third event) [18]. The interaction patterns are composed by a temporal sequences of interdependent events and then they can be seen as a stochastic processes described by two probability distributions: $p(e^P | e^C, e^P)$ and $p(e^C | e^P, e^C)$. Such structures are modelled by a Coupled Event based Dynamic Bayesian Networks (CE-DBNs) in order to have a representation able to statistically encode the relevant data variability. The topology of the above-mentioned AM resembles a CE-DBN structure [19].

IV. AUTOBIOGRAPHICAL MEMORY DOMAIN APPLICATIONS: SURVEILLANCE AND CROWD MANAGEMENT SCENARIOS

In the previous section a probabilistic model based on CE-DBNs was briefly recalled in order to describe multiple entity interactions. The knowledge thus represented inside the proposed Cognitive Node can be employed in many different domains: surveillance scenarios and crowd analysis-management are just two limited examples.

In this section two aspects will be discussed, namely the probabilistic model learning phase and the detection phase for surveillance and crowd scenarios. During the (off-line) learning phase the CN observes both entities, i.e. the human operator and the crowd, storing their interactions within the Autobiographical Memory. As for the (on-line) detection phase, it will be shown how different definitions of the probabilistic model are needed.

The system is designed to support a human operator in crowd management during the on-line phase. This task is accomplished by recognizing specific operator-crowd abnormal interactions. Typically, in people flow redirection problems, an abnormal interaction can be detected whenever the user puts in action wrong countermeasures to avoid the panic or overcrowding situations. In this case the CN ought to drive the operator to choose correct actions by either simply signalling the anomaly or by suggesting actions to be performed based on its learned experience.

A. Learning phase: interaction representations

During an off-line phase, the Cognitive Node is able to learn and store into the AM a set of triplets (i.e. interactions) for different situations: $\{e^P, e^C, e^P\}$ and $\{e^C, e^P, e^C\}$ (i.e. the crowd configurations are captured by core sensors, while the operator actions is mapped into proto sensors). However, each generic triplet of events can be associated to an influence model, i.e. a specific AM can model the dynamic evolutions of interactions for a specific context. It is possible to define a switching variable $\theta$ as influence parameter [20].

Each triplet is associated to a probability, derived from an estimate of two conditional probability densities: $p(e^P | e^C, e^P, \theta)$ and $p(e^C | e^P, e^C, \theta)$, which are proportional to the number of votes that the particular triplet received, i.e. the number of occurrences observed during the AM training phase that represents a specific interaction (i.e. an influence model). Figure 1 shows an example of conditional relationship for a passive triplet: $e^P$ given the two previous events $e^C, e^P$ and the interaction model $\theta$.

![Fig. 1. Example of CE-DBNs for passive triplet, e.g. $\{e^P, e^C, e^P\}$, with a parameter $\theta$ tied across proto-core-proto transitions.](image)

B. Detection phase: Crowd management scenarios

In human-to-human interactions, at each state change of one entity typically corresponds a state change of the other. In this case it is possible to affirm that the entities have the same (or at least a similar) dynamic. On the contrary, in crowd scenarios, the dynamics of the entities are extremely different, namely the crowd changes its status more frequently than the operator. Generally the number of people, in a room or in a zone, can change without any operator actions. In all the cases in which the dynamics between entities show significant differences, the
AM can be considered as a *sparse* collection of triplets. In order to design a robust classification algorithm for abnormal interaction recognitions, an approach to encode a statistical sparse model using the Self Organizing Map is needed. The following section is dedicated to this scope.

V. PROPOSED APPROACH FOR ABNORMAL INTERACTION DETECTION IN CROWD MONITORING DOMAIN

The proposed cognitive video surveillance system has two main purposes. The first and most important one is to detect the interaction anomalies between operator and crowd. The second is to substitute or to help the user during crowd management, recognizing anomalous interactions with the crowd. The presented cognitive system accomplishes both these goals by learning a specific behavioural model for operator-crowd interactions, in which the crowd is correctly controlled by an user. This model describes normal conditions of crowd management. The Cognitive Node can detect anomalous operator-crowd interactions as deviation from normal situations. In automatic operating mode, the system substitutes the operator and interacts directly with the crowd. When crowd reaction patterns are not conform to expected behaviour, an anomalous configuration (i.e. interaction) is detected. The method used for interaction modelling and anomaly detection is here presented. An interaction behaviour cannot be completely represented by a triplet alone: a set of triplets must be analysed in order to individuate a model. A common learning problem can be formalized as follows: the generic sequence of triplets $T_{Tr_j} = \{\vec{\epsilon}_{PC}^-, \vec{\epsilon}_C, \vec{\epsilon}_P^+, \vec{\epsilon}_{PC}^+\}$, $j = 1, \ldots, N_t$, where $N_t$ is the number of triplets in that specific sequence, can belong to different observed models $\theta_i$, $i = 1, \ldots, N_I$ ($N_I$ is the number of operator-crowd interaction models). Figure 2 shows triplet *encoding* by means of a mapping function $f(.)$.

For sparse collected data, i.e. sparse triplets, the *mapping function* defined as $f(\vec{\epsilon}_{PC}^-, \vec{\epsilon}_C, \vec{\epsilon}_P^+, \vec{\epsilon}_{PC}^+) = p(\vec{\epsilon}_{PC}^-|\vec{\epsilon}_C, \vec{\epsilon}_P^+, \vec{\epsilon}_{PC}^+, \theta_j)$ is not potentially useful in order to distinguish triplet associations with specific kinds of operator-crowd interaction models.

A different transform function, $\hat{f}(\vec{\epsilon}_{PC}^-, \vec{\epsilon}_C, \vec{\epsilon}_P^+, \vec{\epsilon}_{PC}^+)$, is defined for triplet mapping into 2-D space to decrease miss-classification errors. A specific dimensionality reduction method can be employed to encode the Autobiographical Memory. This way, it is possible to obtain a probabilistic model for *rare-interaction* detections, in order to describe high-complexity relationships between entities by means of simpler formulas [21].

A. Self Organizing Map: classification evaluation

A large number of methods have been addressed for dimensional reduction: they are typically classified in linear and non-linear methods. This section addresses a fundamental issue rising in reduction problems: interaction information contained in primary data must be preserved. A Self Organizing Map (SOM) [22] unsupervised classifier is employed in this work in order to detect generic proto and core events, $E_j$ where $J \in \{P, C\}$ defines proto or core entity, in term of relevant state changing. The clustering process, applied to $\vec{x}_P(t)$ and $\vec{x}_C(t)$, allows one to obtain a mapping of proto and core vectors in 2-D vectors, corresponding to the positions of the neurons in the SOM-map. These are called proto Super-states $I_P$ and core Super-states $I_C$ respectively.

The choice of SOMs to perform feature reduction and clustering processes is motivated by their capabilities to reproduce in a plausible mathematical way the global behaviour of the winner-takes-all and lateral inhibition mechanism shown by distributed bio-inspired decision mechanisms. Besides, a SOM allows for the establishment of a representation of reality based on analogies: similar (though not necessarily identical!) input vectors are likely to be mapped by the Kohonen map to the same neuron (in a non-injective way).

Taking into account a SOM layer formed by $n$ neurons, its dimensions are adapted in order to find the best matching couple $(l, w)$ such that $l \times w = n$. The number of core (or proto) Super-states is then $n$ and the total number of possible core (or proto) events is $n^2$, taking null-events as relevant as it will be explained. The parameter $n$ must be tuned: in fact, by decreasing the SOM-map size, many different input state vectors can fall into the same cluster: this fact generates a rougher classification but ensures that only relevant events are likely to be selected. On the other hand, by employing high-dimensional Kohonen’s layers, the classification is improved, whereas the total number of irrelevant events increases.

The dimension of the layer is a relevant parameter in our system, from which depends the resolution granularity of stored knowledge. A small layer allows the system to summarize its knowledge in a few concepts, which is positive, although classification of situations may result too rough in some cases. On the other hand, very large layers result in a very sparsely populated Superstate space, meaning that the system would need massive training in order to observe, and later recognize, any possible situation. This parameter is one of those studied in this work. A possible criterion to evaluate a classifier $C$, given a data set $D$, relies on *Average Mutual Information (AMI)* $I_M(D, C)$ defined by Equation 1:

$$I_M(D, C) = H(D) - H(D|C), \quad (1)$$

where $H(D)$ is the data set entropy and $H(D|C)$ is the conditional entropy. Table I lists entropies for different size of the SOM layer. Over $7 \times 7$ the quality of the classification have not significant improvements from AMI point of view.
VI. Results

The simulated monitored environment is shown in Fig. 3. The configuration of doors, walls and rooms is however customizable and for sure tests on different environments will be run in the future to give more consistency to the theory developed so far. The use of a graphical engine (freely available at http://www.horde3d.org/) has been introduced in order to make the simulation realistic in the Autobiographical Memory training phase. Here a human operator acts on doors configuration in order to prevent room overcrowding, based on the visual output of the simulator.

Crowd behaviour within the simulator is modelled based on Social Forces [14]. This model assimilates each character on the scene to a particle subject to 2D forces, and treats it consequently from a strictly physical point of view. Its motion equations are derived from Newton’s law \( \mathbf{F} = \mathbf{ma} \). The forces a character is driven by are substantially of three kinds. An attractive motivational force pulls characters toward some scheduled destination, while repulsive physical forces and interaction forces prevent from collision into physical objects and other characters. The model shows a very (qualitatively) realistic visual output. Characters are also animated to simulate walk motion.

The simulator also includes (simulated) sensors. These try to reproduce (processed) sensor data coming from different cameras looking at different subsets (rooms) of the monitored scene. A virtual people estimation algorithm outputs the number of people by simply adding some noise to the mere number of people framed by the virtual camera.

The state vector of the system (which corresponds to the external object (eso)) is

\[
X_{Cr}(t) = \{x_{Cr_1}(t), x_{Cr_2}(t), \ldots, x_{Cr_n}(t)\},
\]

with \( n = 6 \) in our case (six cameras, one for each room). \( x_{Cr_n}(t) \) is the number of people in room \( n \).

The people flow starts in a hall room, that corresponds to \( x_{Cr_1} \). A \( 7 \times 7 \) 2D SOM is then trained in order to clusterize the state vector space. The SOM Super States (better say, their variations) define events. The internal (endo) state of the system (namely, the rooms’ configuration) is simply modelled by a binary vector storing the state of each door (true if open, false if closed). Variations of such a vector define proto events.

An Autobiographical Memory (IV) is then trained by a human operator who opens virtual gates in order to let the crowd stream outside in case high occupancy states are reached and, at the same time, to minimize the time gates remain open.

In our case, four kinds of simulated scenarios for different crowd behaviours (see Table II), have been taken into consideration, in order to memorize the interactions between a human operator (proto-self) and the crowd (core-self). For instance, the first crowd behaviour, identified by 1d, has \( \mu = \sigma^2 = 1 \) for the Poisson probability mass function regulating the “birth” of new characters, weak interaction force, and a relatively short interaction range.

After mapping the AM into a 2-D space, by means of a Self Organizing Map, the operator’s reactions to different crowd fluctuations, stored and encoded by \( \hat{f} \), can be used on-line to choose an optimal strategy, i.e. to emulate the actions of a human operator, by predicting not only his behaviour but also crowd’s reaction to it.

A reference model \( \theta_i \) for operator-crowd interactions is then designed (refer to figure 2). We define a sequence of passive triplets \( \{Tr_k\} \) (related to \( i = 1d \) crowd behaviour, Table II), where \( Tr = \{e_p, e_c, e_p^\prime\} \). The mapping function \( \hat{f}(Tr_k) = S_k \) def.ines a corresponding sequence of Super states into the SOM-map as follows: \( \{S_k\} = (S_1, S_2, \ldots, S_k) \). In the simulation, the maximum time between two subsequent Super states (\( \ldots, S_k, S_{k+1}, \ldots \)) is taken as \( 8[s] \). After such a time lapse, a new interaction (Super state) is considered. The \( k^{th} \) Super state probability is a vector \( P \) whose elements are defined as: \( P_k = P(S_k) \); it corresponds to the number of occurrences of \( S_k \) over \( \{S_k\} \) with \( k = 1, \ldots, K \). The elements of the transition probability matrix \( M \), are defined as: \( M(S_k, S_{k+1}) = P(S_{k+1}|S_k) \).

A test sequence of passive triplets \( \{Tr^{(1D)}_i\} \) (one for each crowd behaviour listed in Table II) is simulated and processed by \( S^{(ID)}_k = \hat{f}(Tr^{(1D)}_i) \) in order to generate \( \{S^{(ID)}_k\} \) with \( k = 1, \ldots, K \). A weighted average of transition probabilities between subsequent Super states \( \{\ldots, S^{(ID)}_k, S^{(ID)}_{k+1}, \ldots\} \) is computed as follows:

\[
p^{(ID)}_i(t) = \frac{1}{W} \sum_{k=1}^{W} p^{(ID)}_k p^{(ID)}_{k+1|k} \tag{3}
\]

where \( p^{(ID)}_k = P(S^{(ID)}_k | \theta_i) \) and \( p^{(ID)}_{k+1|k} = M(S^{(ID)}_k, S^{(ID)}_{k+1} | \theta_i) \), while
$W$, called moving evaluation windows, defines the number of test sequence triplets considered at each step $t$. We define the probability to reach $k+1^{th}$ super state starting from the $k^{th}$, as follows: $P_{k^{th} \rightarrow k+1^{th}}^D \rightarrow D$. The recognition of the interaction model is performed by taking the maximum probability: $(i^*, t) = \arg \max_i P_{i^*}^D(t)$ with $i = 1, \ldots, N_f$ and $t = 1, \ldots, T$. The couple $(i^*, t)$ defines the kind of recognized operator-crowd interaction $\theta$, and also the maximum time $W \cdot 8 + t \cdot 8[s]$ in which the detection is performed.

The four interaction behaviours (red curve) are compared with the reference model (blue curve). Different averages of transition probabilities curves, with $W = 2, 5, 10, 15$ and $T = 10$ steps, are taken for training; figure 4 shows an example with $W = 10$. Using only a few triplets (i.e. lower $W$ values, e.g. $W = 2$ and $W = 5$) for each time step, some behaviour models result confused. The separation distance between the curves increases when the evaluation window values increase, e.g. with $W = 10$ and $W = 15$.

The Mean Square Error (MSE) is computed, in order to evaluate the distances between the observed interaction behaviour curves and the reference behaviour model. The minimum MSE provides a similarity measure between interaction behaviours. At each time step $t = 1, \ldots, T$ as follows: 

$$
MSE(t) = \frac{1}{W} \sum_{k=1}^{W} (P_{k^*}^D - P_{k}^D)^2
$$

where $P_{k^*}^D$ and $P_{k}^D$ correspond to probability values over $S_k$ and $S_k^*$, i.e. reference and observed sequences. The anomalous interactions between an operator and the monitored crowd could emerge after a normal behaviour, e.g., a careless user does not open some doors. In this situation the CN, working in its on-line modality, is able to recognize anomalous crowd management. Figure 5 shows the normal behaviour, in the specific case of $ID = 1d$ (blue curve) and compares it with the observed operator-crowd interactions (red curve). Using an evaluation window equal to $W = 10$, two processes start to drift away at $t = 6.4$. In a corresponding manner MSE grows. The rule of detection is $\nabla MSE(t) > 0$ for $t \in [t_{min}, t_{max}]$. The system detects operator-crowd anomalous interactions when the curve gradient is positive for an evaluation period $\tau_{ep} = t_{max} - t_{min}$.

A. Application on a real video sequence

In order to give consistence to the proposed cognitive video surveillance system, an experiment has been conducted on a available video sequence from PETS dataset for single camera (S3 High level, Time 14 – 16, View_0001, sequence length is 223 frames, frame rate is $\sim 7$ [fps]). The main target of this experiment is to demonstrate how the system is able to recognize interactions between an operator and the crowd behaviour in a video sequence. For this purpose the real environment has been partitioned in three zones, which are supposed to be monitored by cameras, as shown in figure 6(a). In the simulated environment, the zones correspond to three rooms, figure 6(b). In the sequence, two crowd behaviours corresponding to different people flows have been individuated. The fist flow direction when the people go across the scene from zone 1 to zone 3 (i.e. from frame_0000 to frame_0107), while the second flow when the people move from zone 3 to zone 1 (i.e. from frame_0108 to frame_0222).

By using the simulator these two different people behaviours have been reproduced: the people source in zone 1 and leave in zone 3 (i.e. first flow), the people source in zone 3 and leave in zone 1 (second flow). In the simulator an human operator can manage the crowd flow, from a room to another, by the doors, $d1$ $d2$ $d3$ $d4$. The user opens the door when the people are near to it. In order to reproduce the same interaction using the real video sequences, it has been supposed to have the same configuration of the doors. A people counting algorithm [23] provides an estimate of the total number of people in the zones present in video sequences, figure 6(a). In virtual environment a people counting module is simulated in order to count the people into a sub-area of the room, dashed circular areas figure 6(b).

The test is composed by two parts: learning and detection (on-line). During the learning phase, the cognitive system has learned two probabilistic models by using the simulator, i.e. two Autobiographical Memories, in order to describe two crowd behaviours. The rules used to memorize such two models are specified as follows: if the operator sees the people moving from zone 1 to zone 3 must open only $d1$ $d2$ in...
Fig. 6. Example of real ambient (a) and simulated scenario (b) used for the test, the virtual rooms correspond to the zones. The red dashed line correspond to people flow direction from zone 1 to zone 3; the blue dashed line describes people movement from zone 3 to zone 1. Dashed circular areas qualitatively correspond to the parts of the rooms monitored by cameras equipped with people counter module. \(d_1\) \(d_2\) \(d_3\) \(d_4\) are to the doors.

Fig. 7. Sample frames in four different crowd-environment interactions. Different people flows are presented: two opposite directions of movement (a) (b), people splitting (c), people merging (d). (a) and (b) represent normal behaviours, while (c) and (d) represent two abnormal behaviours.

Fig. 8. The qualitative results of the normal and anomalous operator-crowd interaction detection, during the operator support phase. The ground truth bar represents the operator actions in corresponding with video frames. Prediction and action bar represents the cognitive system actions.

According with the people flow; the user has to act on \(d_3\) \(d_4\) if people flow is in opposite direction. During the second part the system works on real video sequences. The Cognitive Node recognizes the observed situations in according with the memorized knowledges. During autonomous phase, the CN, to the end to interact directly with the crowd, must discriminate different crowd-environment configurations. Figure 7 presents four sample frames about different crowd behaviours: in case (a) the people flow moves leading red arrow (i.e. from zone 1 to zone 3), in case (b) the opposed people movement direction is presented (i.e. from zone 3 to zone 1). In cases (c) and (d) the groups of the people have different movement directions: people splitting and merging. In these last two cases, the system does not find any correspondence between observed scene and stored interaction models. In particular, the CN can consider the scene (c) as anomalous crowd-environment interaction compared with (a) situation. The same consideration can be done for (d) and (b) situations. When anomalous crowd-environment interactions are detected, the cognitive system sends an alarm message in order to inform the human operator. After this phase, the CN is able to predict most likely future actions and when it will be performed. During the operator support phase, the cognitive system individuates anomalies in terms of differences between predicted actions and user actions. In figure 8 the cognitive system predictions and detections of normal and abnormal interactions between an operator and the crowd are shown. Considering the case (a), the system predicts the first zone crossing (i.e. from zone 1 to zone 2) as to open \(d_1\) (specified by the blue triangle). In this case, the operator action results to be corresponding with predicted action (i.e. \(d_1\)). During the second zone crossing (i.e. from zone 2 to zone 3) the system detects an anomalous operator-crowd interaction behaviour: the user opens the wrong door (i.e. \(d_3\) indicated by blue circle). Case (b) presents the same analysis for a different people flow direction.
VII. CONCLUSIONS

An automatic systems called Cognitive Surveillance Node (CSN), which is part of a complex called JDL-based and bio-inspired architecture was presented in this paper. Also, a bio-inspired structure was proposed, for encoding and synthesizing signals for modelling cause-effect relationships between entities. In particular, the case where one of such entities is a human operator was studied. Interaction models are stored within the Autobiographical Memory during a learning phase. Knowledge is thus transferred from a human operator towards the CSN. Learned representations can be used, at different levels, either to support human decisions by detecting anomalies in interaction models and thus compensating for human shortcomings, or, in an automatic decision scenario, to identify anomalous patterns and choose the best strategy to preserve stability of the entire system. Results are shown in a simulated visual 3D environment in the context of crowd monitoring. The simulated crowd is modelled according to the Social Forces Model. The results show two possible applications of the CSN for crowd surveillance applications: first, the system can support the operator for crowd management and people flow redirection by detecting drift from some learned interaction models; secondly, to work in automatic mode and thus autonomously detecting anomalies in crowd behaviour. Furthermore, it has been shown how user-crowd interaction knowledge, learned from the simulator and modelled as proposed, is useful in order to detect anomalies on real video sequence. Future developments of this work will include a detailed study on an other classification algorithm, namely the Instantaneous Topological Map, which in our opinion could enhance feature of constructing a representation of reality based on minimum amount of information extracted from the environment, e.g. limiting the number of the controlled areas, needed for crowd monitoring.

REFERENCES