Jurnal Teknologi

Full paper

Indoor Overhead Video Camera for Efficient People Counting

Juan Serrano-Cuerda[a], José Carlos Castillo[a], Antonio Fernández-Caballero[a*]

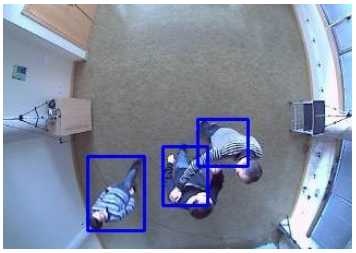
Universidad de Castilla-La Mancha, Departamento de Sistemas Informáticos, Instituto de Investigación en Informática de Albacete, 02071-Albacete, Spain

*Corresponding author: Antonio.Fdez@uclm.es

Article history

Received : 30 June 2013
Received in revised form :
21 February 2013
Accepted : 15 April 2013

Graphical abstract



Abstract

This article introduces a system for real-time people counting. People counting systems are challenging in the surveillance domain. The proposed system is built from INT3-Horus, a multi-agent based framework for intelligent monitoring and activity interpretation. The system uses an indoor overhead video camera that detects people moving freely in a hall or room. The people counting system is flexible in detecting individuals as well as groups. Counting is independent of the trajectories and possible occlusions of the humans present in the scene. The initial results offered by the system are very promising in terms of specificity, sensitivity and F-score.

Keywords: People counting system; overhead video camera setup; INT3-Horus framework; multi-agent system

■1.0 INTRODUCTION

Recent developments in monitoring and activity interpretation systems present significant improvements in the fast and efficient transmission of data, voice and video. An important feature of these systems is the real-time processing of the sensory information to predict the actions of the moving objects along the monitored environment.

This is why monitoring systems have been widely used in a range of domains, and it is nowadays possible to find proposals for means of transportation such as airports [1], ports [2], train and metro stations [3] or traffic control [4]. Also, there are proposals for public places surveillance [5], human activity monitoring [6] as well as industrial deployments [7].

Multisensory monitoring systems usually are composed by three main components: (1) sensors to acquire the information required for the detection, (2) processing algorithms combined with fusion techniques to identify the objects of interest in the scene, and, (3) software architectures to allow the easy integration of the algorithms as well as the scalability to process from several sensor sources.

The main goals of these systems can be summarized as follows:

- Environment monitoring by taking advantage of the different sensor technologies (fixed and moving cameras, laser, microphones, and so on), as well as the cooperation / collaboration among the different processing entities.
- Implementation and optimization of detection algorithms for a robust identification of the objects of interest along with their classification and tracking, dealing with real-time operation.
- Activity detection to identify the relations and behaviors among the objects in the scene, taking into account the low level processing.
- Smart actuation on the scenario through taking into account the special characteristics of the objects. This objective encompasses the development of human-machine interfaces to enhance the display of the information considering its nature.

Among the objectives related to the system proposed in this work, a key one is human detection. This process is also known as human segmentation and essentially consists on grouping image areas that contain significant information [8]. The algorithms are usually based on the decomposition of the images and the clustering of the selected pixels to generate more abstract entities.

Now, people counters have been widely addressed during the last few years, mainly for surveillance applications [9], [10]. This paper is focused on a system that calculates in real-time the number of people that are present in a given scenario, monitored by an indoor overhead video camera overlooking a scenario such as a hall or room [11], [12]. People move freely, as there is no clear entrance/exit at the monitored scene. Also, there is no initial limitation in the number of people to be detected; single humans appear in the scenario and also groups of people are allowed. A non-calibrated overhead camera is used to avoid the majority of occlusions present in lateral video camera installations.

The people counting system described in this paper has been developed from INT3-Horus, a framework for intelligent monitoring and activity interpretation. The paper introduces a description of the INT3-Horus framework, its particularization to develop the people counting system, and some promising results of the setup of the indoor overhead camera application.

■2.0 FROM INT3-HORUS FRAMEWORK TO PEOPLE COUNTING SYSTEM

This section starts with a general description of INT3-Horus. Then, the general structure of the people counting system is described. Finally, each level used for people counting from an indoor overhead camera is described extensively, with special focus in the algorithms developed to perform this specific task.

2.1 The INT3-Horus Framework

Most monitoring and activity interpretation systems define a set of processing levels to achieve specific goals such as people monitoring, trespassing detection or people counting. The problem arises from the lack of consensus among proposals, since different levels are usually proposed to carry out similar tasks. This problem complicates the integration of different algorithms to develop new monitoring systems without struggling with the hard task code integration

INT3-Horus is conceived as a framework to carry out monitoring and activity interpretation tasks (e.g. [13], [14], [15]). This is an ambitious goal given the huge variety of scenarios and activities that can be faced [16], [17], [18]. The framework establishes a set of operation levels where clearly defined input/output interfaces are defined. Inside each level, a developer places his/her code, encapsulated in a module in accordance with the operation performed. The purpose of this framework is easy and fast code integration and generation of real-world systems selecting the best combination to solve a problem from the available modules

Although a set of levels are proposed in INT3-Horus to cover all the steps of a generic multi-sensor and activity interpretation system [19], [20], [21], the philosophy underlying the framework allows a flexible set of levels to be adapted to a given final system [22], [23].

The framework infrastructure as well as the modules layout is based on the Model-View-Controller (MVC) paradigm [24], which allows isolating the user interface from the logical domain for an independent development, testing and management. The MVC paradigm divides an application into three main entities, defining their main roles as well as the connections among them. *Model* manages the application data, initializes objects and provides information about the application status. In event-driven systems, the *Model* informs the *View* and the *Controller* about information changes. *View* provides a representation of the *Model* information (performing just simple operations) to fit user requirements. Finally, *Controller* receives the inputs to the application and interacts with the *Model* to update its objects, and with the *View* to represent the new information.

Despite of the many benefits provided by the MVC paradigm, the union of the business logic and the data model presents a drawback when it comes to add new functionalities to the framework. To solve this issue the traditional MVC is extended for INT3-Horus, creating a new component to house each module's specific operation (see Figure 1 on the right). This way, each framework user receives a template with a series of components which are already integrated into INT3-Horus. The main task for the integration is to introduce the code into the component named "Algorithm" and tune up the rest of components if needed (e.g. adding controls to the module's *View* or data structures to the local *Model*). The module's controller provides the connections to the framework and the access to the global data model as well as the signals to control the execution.

In this sense, the framework allows easy code integration, providing users with module templates to put their code into them. These templates already have the necessary connections to access the rest of INT3-Horus components, not only the data model or the user interface, but also the controller to trigger each module's execution.

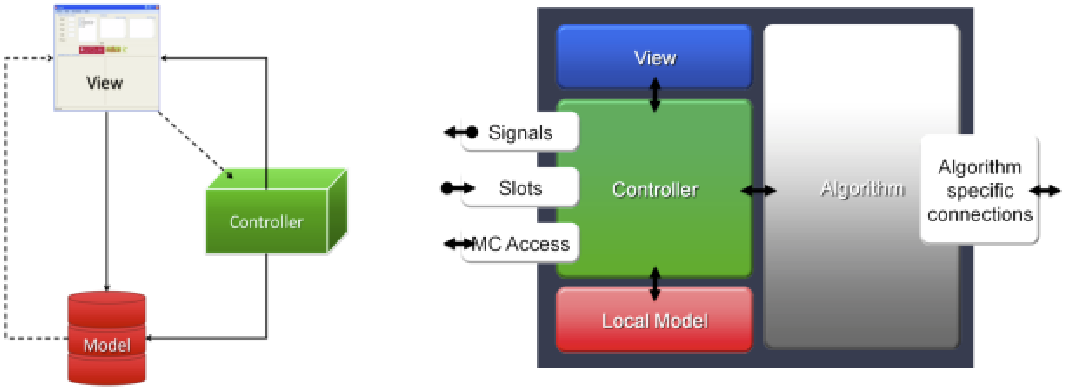


Figure 1 Model-View-Controller representation. The left image shows the traditional MVC layout. The right image shows the extended MVC

Together with the easy addition of new functionalities, a state of the art framework for monitoring and activity detection must take into account several information sources (sensors) and INT3-Horus is not an exception. These sources are mainly related to image sensors since they are the most widespread for monitoring tasks; but other sensor technologies, like commercial sensors and wireless sensor networks (WSNs), are also integrated to show the generic purpose of INT3-Horus.

Among the advantages provided by a formal architecture for monitoring and activity interpretation, three contributions are considered especially important:

- Modularity to incorporate new features depending on the system requirements. This implies the easy addition of new processing modules and the arrangement of the components into a set of operation levels abstracted from the trends found in the literature. These levels have also to include data fusion to maximize the quality of the information sources.
- Environment modeling is usually forgotten in most proposed architectures, but it is crucial to achieve the previous contribution. Several aspects can be modeled such as environment and sensors, as monitored areas usually are big and the cooperation among the sensors is essential to carry out tasks such as object tracking, regardless of the information source.
- Execution model must be selected according to the system demands. In the case of monitoring systems, where different sensor information is processed in real time and fused to provide a unified view of the scenario status, the hybrid or hierarchical execution model seems a reasonable option. This model proposes the use of remote nodes in charge of the collection of the information in the first levels of the processing stack whilst a central node unifies the information and performs higher level operations such as activity detection.

2.2 INT3-Horus Levels for People Counting

The main goal of an efficient people counting system is to obtain the number of humans inside the field of view of a camera. In this particular case, we describe a system based on an indoor overhead camera with the highest possible accuracy when counting people. For this purpose, four processing levels are selected from INT3-Horus framework.

The lower one is in charge of collecting data from an overhead camera *acquisition*. The next level, *segmentation*, uses an approach based on background subtraction to isolate the humans in the scene. Some filtering and heuristics are applied in the next level, *blob detection*, to enhance segmented humans, dealing with false positives and splitting groups of people into individuals.

2.2.1 The Data Acquisition Level

At *acquisition* level, a specific module is in charge of capturing images from Axis cameras. This module uses VAPIX (http://www.axis.com/techsup/cam_servers/dev/index.htm), an HTTP-based application programming interface that provides functionality for requesting images, controlling network camera functions (pan-tilt-zoom, relays, etc.) and setting/retrieving internal parameter values. The proposed people counting system only needs images obtained from a networked camera.

2.2.2 The Segmentation Level

Image segmentation can be defined as the process by which an image is divided into parts or objects that constitute it [8]. The objective is the location of significant areas of the image, such as imperfections in a tool, urban areas in case working on a map, humans, etc. This process involves two major tasks. On the one hand it is necessary to perform the decomposition of the images for further analysis and, second, the pixels of the images should be organized in higher-level units that acquire meaning for further analysis.

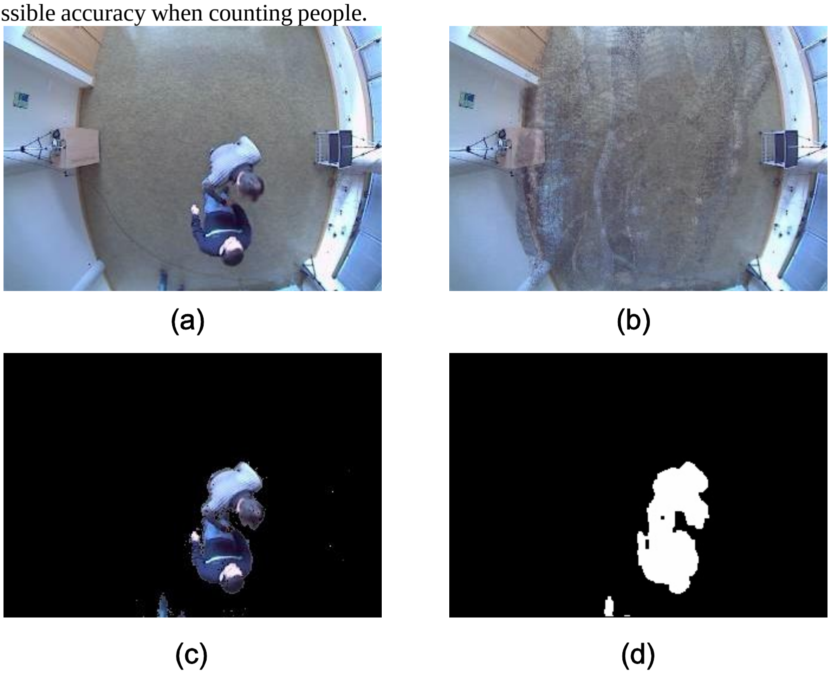


Figure 2 Images generated by the segmentation level. (a) Current frame. (b) Background model. (c) Foreground. (d) Segmented image

Statistical and probabilistic techniques use temporal consistency of the sequences [25], [26]. Region growing techniques partition images into different regions with common features [27]. Graph-based techniques interpret the images as a set of vertices and nodes [28]. Finally, learning-based techniques select a set of pixels as prior knowledge (supervised learning) or estimate the optimal number of regions in the images (unsupervised learning) [29], [30].

The main objective of the *segmentation* level is to perform the initial detection of the humans present in the scene. An adaptive Gaussian background subtraction is performed on input image I_Z obtained from the indoor overhead camera, as shown in Figure 2a. The subtraction is based on the *OpenCV* implementation of a well-known algorithm [9]. The algorithm builds an adaptive model of the scene background based on the probabilities of a pixel to have a given color level. An example of this model is shown in Figure 2b. A shadow detection algorithm, based on the computational color space used in the background model, is also used. After the background segmentation is performed, an initial background segmentation image I_B is obtained as shown in Figure 2c.

However, the resulting image contains some noise which must be eliminated. For this, an initial threshold θ_0 (experimentally fixed as a 16th of the number of possible gray levels in the image) is applied, as shown in equation (1):

$$I_{th}(x,y) = \begin{cases} min, if\ I_B \leq \theta_0 \\ max, otherwise \end{cases} \quad (1)$$

and

$$\theta_0 = \frac{max}{16}$$

where min is fixed to 0 (since we are obtaining binary images) and max is the maximum gray level value that a pixel can have in I_B (e.g. 255 for an 8-bit image).

After this operation, two morphological operations, namely opening and closing, are performed to eliminate the remaining noise of the image, obtaining I_S as shown in Figure 2d. After the first noise reduction, the number of white pixels (corresponding to possible humans) is counted in the image. If this value is greater than a 50% of the area of the image during a predefined amount of time Δt (usually one second), it is estimated that a big lighting change has occurred in the scene (e.g. a light switch turned on/off or a door was opened/closed). In this case, the algorithm is reinitialized to build a new background based on the new lighting conditions of the scene.

2.2.3 The Blob Detection Level

Now, human candidates must be extracted from binary image I_S, paying special attention to the existence of groups of people. For this purpose, the concept of region of interest (ROI) is explained. A ROI is defined as the minimum rectangle containing a human. It can be characterized by a pair of coordinates $\{(x_{min},y_{min}),(x_{max},y_{max})\}$, corresponding to the upper-left and lower-right limits of the ROI, respectively. All detected ROIs are used to annotate the humans detected in the scene in a list L_B. A summary of the stages of the blob detection level can be seen in Figure 3.

In first place, human candidates are extracted from the scene. With this objective, connected components (blobs) are extracted from I_S. Next, blobs with a ROI area lower than A_{min} (with a value experimentally fixed according to scene features such as the scene area or the height where the camera is placed) are discarded. A new area threshold A_G is also established based on similar factors.

Blobs with a ROI area lower than A_G are considered to contain a single human and the ROI containing it is enlisted in L_B. Otherwise, blobs B_G with a ROI area greater than A_G are analyzed separately, since they are considered to possibly contain a group of humans. Now, each human belonging to these groups is extracted individually. To do so, a new sub-image I_G is created containing the ROI delimiting B_G, as shown in Figure 4a. Then, a new series of morphological openings are performed, since occlusions are less frequent in an overhead view than in a lateral view, obtaining a new image I_G. An example of the result of these operations is offered in Figure 4b. Next, blobs are searched in this new image. Now, blobs with an area greater than A_{min} are annotated in a list of group blobs L_{BG}, whilst the others are discarded. If, at the end of the search, L_{BG} is empty, the original ROI with the blob B_G is enlisted as a single human; but, it will be marked as a possible group that could not be separated. Finally, the blobs from L_{BG} are enlisted in L_B, where the number of humans in the scene (people counting) is the number of blobs contained in L_B. The detected humans in the scene are shown in Figure 4c for this running example.

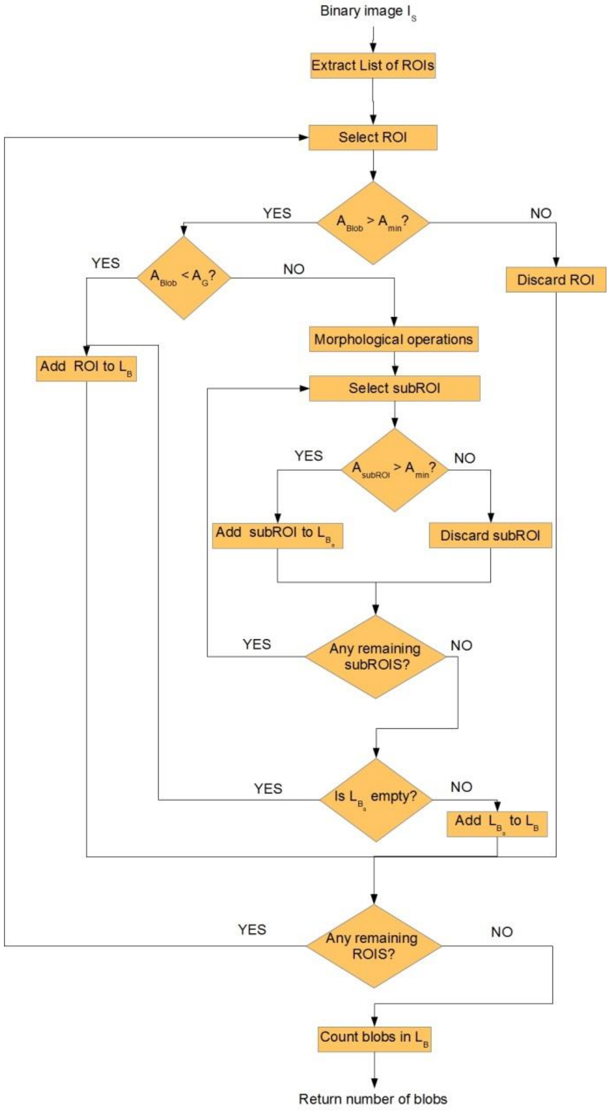


Figure 3 Stages of the blob detection level

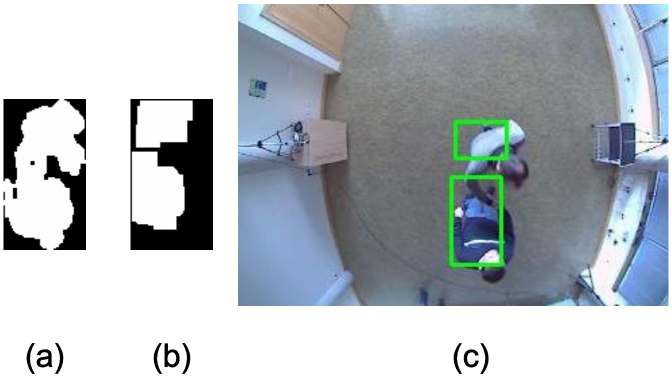


Figure 4 Results of the blob detection level. (a) Original ROI. (b) Separated ROIs. (c) Final Result

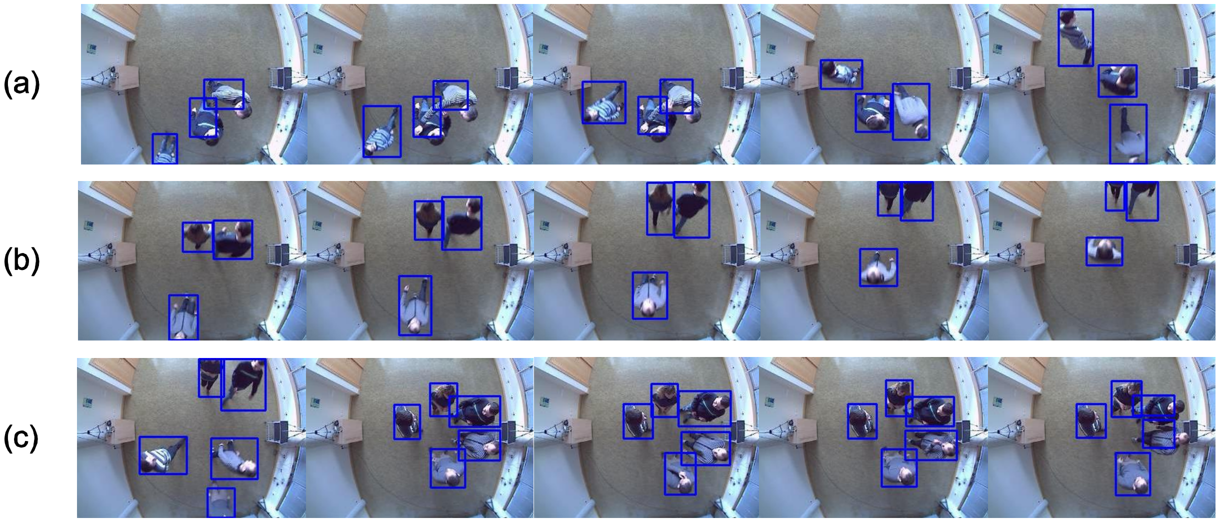


Figure 5 Qualitative results for the three recorded sequences. (a) Video 1. (b) Video 2. (c) Video 3

■3.0 DATA AND RESULTS

Three different video sequences were recorded from an Axis camera to test our proposal. The first sequence shows different people walking along a hallway (individually or in groups of two or three individuals). Generally, the people do not stop and do not cross their paths, except for the final frames of the video, where two people meet and talk for a while in the center of the scene and another person approaches them (see Figure 5a). The second sequence is similar to the first one, although more occlusions appear as different people intersect their paths. Another meeting takes place in this second video; this time, there are three people remaining still in the scene for a minute without being added to the background model. An example of the group separation in this sequence is shown in Figure 5b.

The final sequence is the most complex one. In this video, up to five people appear in the scene, crossing and intersecting their paths, which results in a greater amount of occlusions than in the previous sequences. They also meet for two minutes without being added to the background and partially occluding themselves. Nevertheless, they are detected most of the time as shown in Figure 5c.

Table 1 shows some quantitative results extracted from the three sequences. In order to evaluate the performance, we have used measures of specificity, sensitivity and F-score. These are usual statistics in image processing, calculated as shown in equations (2), (3) and (4) respectively.

$$specificity = \frac{TP}{TP + FP} \quad (2)$$

$$sensitivity = \frac{TP}{TP + FN} \quad (3)$$

$$F - score = \frac{2 \cdot specificity \cdot sensitivity}{specificity + sensitivity} \quad (4)$$

TP (true positives) is the number of correct detections in the scene, *FP* (false positives) is the number of humans detected but actually not present, and *FN* (false negatives) is the number of humans present in the scene that have not been detected by our algorithm.

Notice that the results are really outstanding. The first sequence shows worse results since several small lighting changes took place during the recording. Also small misdetections have greater impact in the final statistics as fewer humans appear in this sequence than in the later ones. It is also important to highlight that the results of the final sequence show the difficulty of the video, since a lot of humans appear and they are occluding themselves most of the time.

Table 1 Results of people counting in different video sequences

Sequence	Humans in the sequence	Humans detected	TP	FP	FN	Specificity	Sensitivity	F-score
1	1060	1024	1001	23	59	0.944	0.978	0.960
2	1698	1702	1656	46	42	0.975	0.973	0.974
3	3199	3274	3128	146	71	0.978	0.955	0.966
Total	**5957**	**6000**	**5785**	**215**	**172**	**0.971**	**0.964**	**0.967**

■4.0 CONCLUSIONS

This paper has introduced an efficient people counting system. The system based on an indoor overhead camera counts the number of people that are present in a given scenario in real-time. There is no restriction in the motion of the people. Even, there is no limitation in the number of people to be detected. The people counting system accepts individual as well as groups of people.

The people counting system described in this paper has been developed from INT3-Horus, a framework for intelligent monitoring and activity interpretation. The paper has demonstrated the usefulness of the framework and the accuracy of the developed system.

Acknowledgements

This work was partially supported by Spanish Ministerio de Economía y Competitividad / FEDER under TIN2010-20845-C03-01 grant.

References

[1] del Corte, A., Gutiérrez, O. and Gómez, J. 2012. New Location Techniques Based on Ray-tracing for Increasing Airport Safety in Apron and Stand Areas. *Frontiers in Computer Education*. 515–522.

[2] Huck, R.C. 2011. A Building Block Approach to Port Security. PhD Dissertation, School of Electrical and Computer Engineering. The University of Oklahoma.

[3] Svenonius, O. 2012. The Stockholm Security Project: Plural Policing, Security and Surveillance. Information Polity. 17: 35–43.

[4] Casas, P., Mazel, J. and Owezarski, P. 2012. Knowledge-independent Traffic Monitoring: Unsupervised Detection of Network Attacks. *IEEE Network*. 26(1): 13–21.

[5] Cucchiara, R., Grana, C., Prati, A., Tardini, G. and Vezzani, R. 2004. Using Computer Vision Techniques for Dangerous Situation Detection In Domotic Applications. Proceedings of the IEEE Workshop on Intelligent Distributed Surveillance Systems. 1–5.

[6] Kieran, D. and Yan, W. 2011. A Framework for an Event Driven Video Surveillance System. *Journal of Multimedia*. 6(1): 3–13.

[7] Geradts, Z. and Bijhold, J. 2000. Forensic Video Investigation. Multimedia Video Based Surveillance Systems, chapter 1. 3–12.

[8] Gonzalez, R. C. and Woods, R. E. 2007. *Digital Image Processing*. 3rd Edition. Prentice-Hall.

[9] Boltes, M. and Seyfried, A. 2013. *Collecting Pedestrian Trajectories. Neurocomputing*. 100: 127–133.

[10] Moreno-Garcia, J., Rodriguez-Benitez, L., Fernández-Caballero, A. and López, M.T. 2010. Video Sequence Motion Tracking by Fuzzification Techniques. *Applied Soft Computing*. 10(1): 318–331.

[11] Rao, R., Taylor, C. and Kumar, V. 2006. Experiments in Robot Control from Uncalibrated Overhead Imagery. Experimental Robotics IX. 491–500.

[12] Serrano-Cuerda, J., Castillo, J. C., Sokolova, M. V. and Fernández-Caballero, A. 2013. Efficient People Counting from Indoor Overhead Video Camera. *Trends in Practical Applications of Agents and Multiagent Systems*. 129–137.

[13] Gascueña, J. M. and Fernández-Caballero, A. 2011. On the Use of Agent Technology in Intelligent, Multisensory and Distributed Surveillance. *The Knowledge Engineering Review*. 26(2): 191–208.

[14] Gascueña, J. M., Fernández-Caballero, A., López, M.T. and Delgado, A.E. 2011. Knowledge Modeling Through Computational Agents: Application to Surveillance Systems. *Expert Systems*. 28(4): 306–323.

[15] Gascueña, J.M., Navarro, E. and Fernández-Caballero, A. 2012. Model-driven Engineering Techniques for the Development of Multi-Agent Systems. *Engineering Applications of Artificial Intelligence*. 25(1): 159–173.

[16] Kieran, D. and Yan, W. 2011. A Framework for an Event Driven Video Surveillance System. *Journal of Multimedia*. 6(1): 3–13.

[17] Castillo, J. C., Fernández-Caballero, A., Serrano-Cuerda, J. and Sokolova, M.V. 2012. Intelligent Monitoring and Activity Interpretation Framework - INT3-Horus Ontological Model. *Advances in Knowledge-Based and Intelligent Information and Engineering Systems*. 980–989.

[18] Sokolova, M.V., Castillo, J. C., Fernández-Caballero, A. and Serrano-Cuerda, J. 2012. Intelligent Monitoring and Activity Interpretation Framework-INT3-Horus General Description. *Advances in Knowledge-Based and Intelligent Information and Engineering Systems*. 970–979.

[19] Pavón, J., Gómez-Sanz, J., Fernández-Caballero, A. and Valencia-Jiménez, J.J. 2007. Development of Intelligent Multi-sensor Surveillance Systems with Agents. *Robotics and Autonomous Systems*. 55(12): 892–903.

[20] Rivas-Casado, A., Martinez-Tomás, R. and Fernández-Caballero, A. 2012. Multiagent System for Knowledge-based Event Recognition and Composition. *Expert Systems*. 28(5): 488–501.

[21] Fernández-Caballero, A., Castillo, J.C. and Rodríguez-Sánchez, J.M. 2012. Human Activity Monitoring by Local and Global Finite State Machines. *Expert Systems with Applications*. 39(8): 6982–6993.

[22] Carneiro, D., Castillo, J. C., Novais, P. and Fernández-Caballero, A. 2012. Multimodal Behavioural Analysis for Non-invasive Stress Detection. *Expert Systems with Applications*. 39(18): 13376–13389.

[23] Costa, A., Castillo, J. C., Novais, P., Fernández-Caballero, A. and Simoes, R. 2012. Sensor-driven Agenda for Intelligent Home Care of the Elderly. *Expert Systems with Applications*. 39(15): 12192–12204.

[24] Reenskaug, T. 1979. Thing-model-view-editor an Example from a Planning System. XEROX PARC Technical Note. May 1979.

[25] Kass, M., Witkin, A. and Terzopoulos, D. 1988. Snakes: active Contour Models. *International Journal of Computer Vision*. 1(4): 321–331.

[26] Mikolajczyk, K., Schmid, C. and Zisserman, A. 2004. Human Detection Based on a Probabilistic Assembly of Robust Part Detectors. Proceedings of the Conference on Computer Vision. 1: 69–82.

[27] Jung, C. R. and Scharcanski, J. 2005. Robust Watershed Segmentation Using Wavelets. *Image and Vision Computing*. 23(7): 661–669.

[28] Shi, J. and Malik, J. 1997. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 22: 888–905.

[29] Yu, Z., Wong, H.-S. and Wen, G. 2011. A Modified Support Vector Machine and its Application to Image Segmentation. *Image and Vision Computing*. 29(1): 29–40.

[30] Lloyd, S. P. 1982. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*. 28: 129–137.