Segmentation of Moving Objects with Information Feedback Between Description Levels

M. Rincón, E.J. Carmona, M. Bachiller, and E. Folgado

Dpto. de Inteligencia Artificial. ETSI Informatica. UNED.
Juan del Rosal 16, 28040 Madrid, Spain
{mrincon, ecarmona, marga, efolgado}@dia.uned.es

Abstract. In real sequences, one of the factors that most negatively affects the segmentation process result is the existence of scene noise. This impairs object segmentation which has to be corrected if we wish to have some minimum guarantees of success in the following tracking or classification stages. In this work we propose a generic knowledge-based model to improve the segmentation process. Specifically, the model uses a decomposition strategy in description levels to enable the feedback of information between adjacent levels. Finally, two case studies are proposed that instantiate the model proposed for detecting humans.

1 Introduction

In recent years, many researchers have focused their attention on detecting and tracking moving objects in video sequences given that it is the first significant step for many machine vision applications, such as semantic video annotation, pattern recognition, video surveillance, traffic control, detection and tracking of people, perceptual interfaces, etc. Obviously, depending on the application, it will be necessary to describe the scene with a different degree of detail, which implies applying different techniques to extract image information and a different degree of precision in segmenting the objects of interest. Thus, in some video surveillance applications it is necessary to distinguish the movement of the different parts of the body to recognise the specific action that the human is doing, for example, to determine whether he is carrying a briefcase or some dangerous object in his hands or not. By contrast, in other applications, the human can be treated as a rigid body, since only the system needs to detect his presence in a room, his passing through a specific area or, generally, the analysis of his path. Therefore, in the first instance a more precise segmentation is required than in the second one. This work focuses on obtaining a robust segmentation with enough degree of precision for the application.

The segmentation algorithm outputs, especially if we work with real scenes, generally contain noise. Noises are primarily due to the intrinsic noise of the video camera, to unwanted reflections, to objects that have a colour that matches the background totally or partially and the existence of sudden shadows and artificial

J. Mira and J.R. Álvarez (Eds.): IWINAC 2007, Part II, LNCS 4528, pp. 171–181, 2007.

or natural changes in the lighting. The total effect of these factors is twofold: first, it may mean that areas that do not belong to the moving objects are incorporated into the foreground (foreground noise), and second, that certain areas, which belong to the objects, do not appear in the foreground (background noise).

There are a number of methods for segmenting moving objects present in a video sequence. These are based, for example, on statistical methods [1][2], the subtraction of consecutive frames [3], optic flow [4], genetic algorithms [5] or on hybrid methods [6][7][8][9] that combine some of these techniques. However, due to the speed and ease of implementation, one of the most frequently used methods, with a fixed camera, is the one based on background subtraction and its many variants [10][11][12][13]. In all these works, a segmentation is generated whose goodness depends on adjusting the method parameter configuration for a specific type of scene, but there is no resegmentation of the scene in the event of error.

In real scenes it is difficult to obtain a precise segmentation in a first approach. Although, previously, knowledge on the type of objects was used to refeed the segmentation process [14][15][16], it was only used on static images and using basic generic characteristics of the objects of interest (continuity or smoothness properties of the contours, local uniformity of movement, etc).

Generally, the main problem in interpreting images is the huge semantic gap that exists between the physical signal level and the knowledge level. To facilitate this gap it is necessary to insert new levels and inject the knowledge available. Following the proposal by Nagel et al. [17], in this work we distinguish different description levels with an increasing degree of semantics: Pixel level, Blob level, Object level, and Activity-behaviour level. Specifically, the final aim is to show how the segmentation (blob level) results can be improved when there is an exchange of information between the object level and blob level assuming that models related to the type of objects of interest exist.

This paper is organized as follows. Section 2 describes in the first instance the generic segmentation model proposed, and after the specific model for human segmentation. Section 3 analyses two case studies that highlight the instantiation of the model in two different situations where the segmentation is improved by applying different operators. Finally, section 4 analyses the results obtained and the future work proposed for improving the proposal.

2 Description of the Segmentation Model

Figure 1 shows the segmentation model proposed where the feedback cycle between levels is evident. The segmentation process begins by taking a video sequence frame as input. The result of this operation will be an initial proposal of the set of blobs associated with moving scene objects. This set of blobs, from the blob level, plays the role of findings at object level and is the input to the diagnosis task. The approach used at object level to refeed the blob level is based on the well-known strategy of diagnosis and planning of therapies used in medicine. Here, the quality of the segmentation is diagnosed based on normality

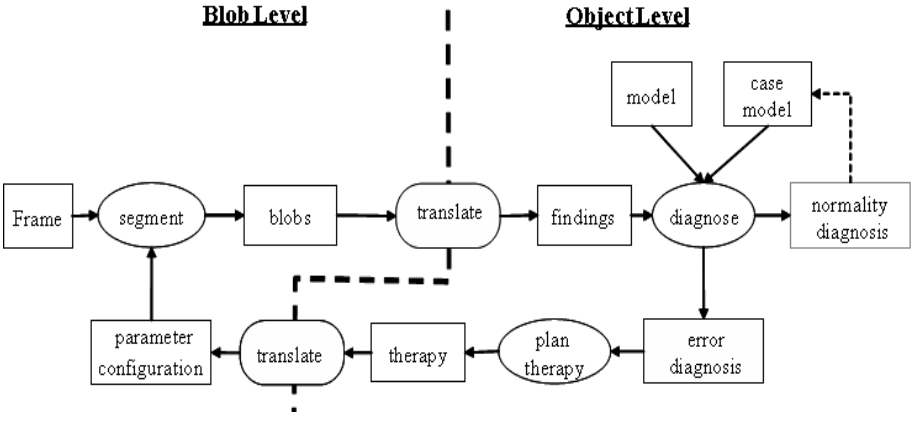

Fig. 1. Feedback structure proposed

models of the objects being recognized, distinguishing between normal situations (normality diagnosis) and abnormal situations (error diagnosis).

If a normality diagnosis is obtained, the resulting information updates the description of the object to that moment (case model). Conversely, if an abnormal situation has been detected, the feedback process is used to solve it by applying the appropriate therapy. This therapy is translated, at blob level, into a parameter configuration that affects the segmentation process. The feedback cycle is thus completed.

The separation into description levels means that it is necessary to introduce translation operations to adapt the entities that play the roles between adjacent levels. Thus, whereas at blob level, we speak of blobs associated with an object, these very blobs play the role of findings at object level. Similarly, the therapy planned to solve the segmentation problem becomes a new parameter configuration that will affect the segmentation process.

2.1 The Initial Segmentation

The process begins with an initial segmentation of the frame i of the video sequence using a method independent of the domain and tuned for the type of scene captured by the camera (context dependent). Specifically, in this work, as an initial segmentation method, the approach explained in [13] characterised by its ease of implementation and low computational cost was used. This approach is inspired by the subtraction background method and, therefore, it will be assumed that at every moment we will work with scenes taken with a fixed camera.

2.2 The Human Model

In this work we applied the segmentation model to detect humans. Consequently, a human model is required as a reference model for the diagnosis. The human

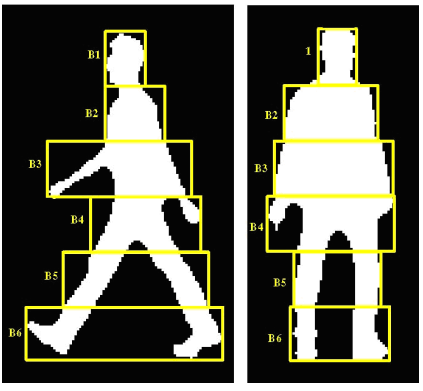

Fig. 2. A human blob in frontal and lateral position divided into blocks

model used here [18] consists on a block model. Basically it consists of dividing the blob corresponding to the human vertically into six regions the same height (figure 2). Each of these regions is defined by the rectangle that circumscribes it and that we will call block. Conceptually, the blocks of this division correspond to zones related to the physical position of specific parts of the body when usual actions are done with normal human movement (head, hands, feet, trunk). The main advantage of this division is that it enables us to study the human in parts. Thus, for example, if we are analysing the hands we know that in a normal situation these are between blocks B3 and B4, otherwise we would detect an abnormal situation. Another advantage of this model is that it is also possible to handle frontal and lateral human views homogeneously.

2.3 Therapy Diagnosis and Planning

In the diagnosis stage, the aim is to evaluate whether the result obtained in the segmentation is coherent with the human model described in the previous section. In our case, with this model it must be possible to detect, for example, the absence of some significant part of the body, the division of the body in several unconnected blobs, an unjustified change in the segmentation from one frame to the next, the presence of parts not related to the human (due to foreground noise), etc.

After the problem has been detected, it is necessary to generate a plan to correct it. As the problems raised are related with the segmentation of an object, either because all the corresponding blobs have not been assigned to the object, or because more blobs have been assigned than necessary, operators must be applied to modify this segmentation. Obviously, if we tried to improve the segmentation globally on all the scene, we would encounter numerous problems, since segmentation operator behaviour is not usually homogeneous (whereas in some regions of the image the segmentation detects the objects precisely, in others it may introduce foreground noise, for example). Conversely, if we focus on

the problem region and reduce the analysis region, operator behaviour is probably more effective. In fact, the smaller the analysis region, the more homogeneous segmentation operator behaviour has to be.

2.4 Segmentation Operators

To improve the segmentation we exemplify here two operators that work at blob level: combination operator and restoration operator.

- The combination operator consists of modifying the set of blobs associated with an object, i.e., assigning or withdrawing some blob to/from the set under certain determining factors imposed by the domain and controlled from the object level via the therapy proposed.
- The restoration operator restores those human pixels associated with background noise. For example, let us assume that, following the previous example, the blobs corresponding to the head, arms and legs of the human have been located, but that one part of the body is so like the background that it has not been detected in the first instance by the segmentation algorithm. In this case, the restoration operator can be applied so that the missing part emerges focusing exclusively on the specific region at blob level of the human model and its associated blobs.

3 Case Studies

This section shows two segmentation examples of moving humans, using the model shown in figure 1 and each of the two segmentation operators described above.

Case 1

One example of combination operator use is that shown in figure 3. The human on the right has been correctly detected as one blob. This will happen as long as the human is notably different from the background. However, the human on the left has been divided into two blobs because his shirt has not been detected. In this instance, it is at object level where this situation has to be detected, i.e., the blob on the right will be recognised as human, but the same will not occur with the other two blobs when they are analysed separately and recognised as not being consistent with a complete human (error diagnosis). In this last instance, the operator will seek to match one of these blobs with different parts of the body, for example, the blob situated in the upper left part (figure 3b) has a high probability of corresponding to a head. Therefore, it is logical to think that the rest of the body must be within a region that has some dimensions matching the scale of a human in the type of scene considered and this region is in the centre of the lower part of the head. Thus, the combination operator will assign all the blobs found in that region to the set of blobs belonging to the human, thereby achieving an initial improvement of the object segmentation.

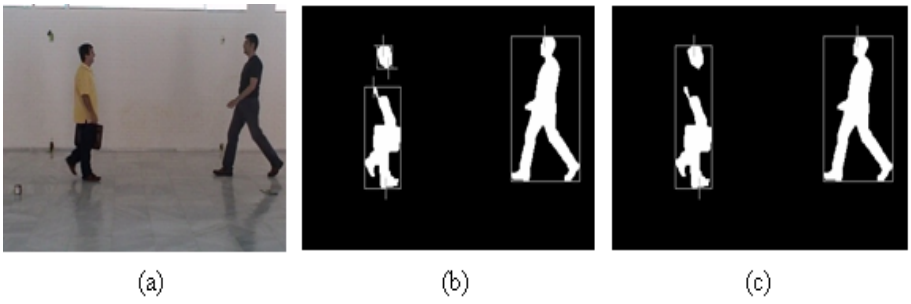

Fig. 3. Example of combination operator use: a) original frame b) the human on the left has been segmented into two very different blobs that have been initially assigned to two different objects c) using a combination operator both blobs have been grouped, and the bounding box has been obtained corresponding to the complete object

This implies passing, as information to the blob level, the reference blob (the one recognised as a head) and a region of interest or ROI where the blobs must be sought that have to be joined to this reference blob. The result of the process is shown in figure 3c.

Case 2

When the background subtraction method is used, the main causes of a background noise are usually associated with those pixels of the moving object whose RGB value is very like that of the background. Indeed, in this instance, the difference of the RGB values of this type of pixels with their background equivalents is not large enough to exceed the threshold and, thus, they are classified as not belonging to the foreground, when in fact they do belong there. Thus, the aim in this example is to use the feedback process to restore those parts of the human which, due to the presence of the background noise, were not detected in the first instance during the segmentation stage and were indeed detected as missing at object level from the human model. Observe that now, when making those parts of the human emerge which were missing, not only do we manage to restore the human silhouette, but also, as in the case analysed before, we are associating all those blobs belonging to the same individual explicitly.

The exchange of information between the different levels can also be described following figure 1. At the object level, using the human model described in secton 2.2 as a reference and the set of blobs resulting from the segmentation process as input (findings), all those blobs that may belong to the human are recognised. Using the human model, for example, the aim is to find whether some block exists with a significant absence of pixels (error diagnosis). If this is the case, this region of the image will be proposed as the region of interest where the missing blobs should be sought (therapy). Another error diagnosis could be the detection of unconnected blobs, which suggests the need to connect them by obtaining new regions belonging to the object (therapy). In either of these two cases, the position of the box and its dimensions are used as information feedback

(parameter configuration). Already at blob level, the restoration operator focuses its analysis on the region of interest (ROI) of the image to locate new blobs not detected before. The aim is to make the largest amount of pixels belonging to the human and only to the human emerge which were not detected in the initial segmentation process. Then all the blobs associated with the object are grouped to generate a new set of blobs. These once more pass to the object level, where the diagnosis process will again check whether the degree of restoration of the human silhouette is sufficiently acceptable for the application needs or, on the contrary, whether it is necessary to repeat the cycle. Observe that, although the process description was done based on only one block of the human model with the need for restoration, really there is no limit to the number of boxes that can be refed at lower levels.

Restoration Operator

An important element in the description of all the previous process exposed in case 2 is the restoration operator of the ROIs where it is hypothesised that part of the object must be detected. The characteristics of this operator are based on the truncated cones method [13] for background substraction. Basically, the idea is the following, as can be seen in Figure 4a, for each point p of the background model, a revolution cone can be built using the straight line containing its associated RGB vector, B_t^p, as the axis and another straight line as a generator which, passing through the origin, forms an angle, ω, with the previous straight line. If we now trace three perpendicular planes to the vector B_t^p, one containing the point p (reference plane), and the other two, situated above and below this, at a distance h_1 and h_2, respectively, these planes will delimit, together with the cone surface, two adjacent regions of interest: a truncated cone situated in the upper part of the reference plane and another in the lower part. Since the pixels associated with parts of the object not detected in the first segmentation present RGB values very close to the background RGB values, it is obvious that if sufficiently small h_1, h_2 and ω values are chosen, these parameter values will delimit a region where this type of pixels are contained. However, the disadvantage is that in this region all the pixels belonging to the fluctuation noise are also included. Indeed, the RGB level of the image pixels that do not belong to moving objects is not exactly the same as the RGB level of the background model pixels, but rather the RGB level of the image pixels presents small fluctuations around the RGB level of the background model pixels.

Therefore, the problem is how to appropriately choose the value of the parameters that define the truncated cone region to separate, if it is possible, the two types of pixels mentioned. For this, we unfold each of the three parameters defined before into two to divide the original truncated cone volume into new subregions. Thus, as is indicated in figure 4b, $h_{21}$and h_{22} (together with ω) will make it possible to delimit a truncated cone volume in the upper part of the original volume. Similarly, h_{11} and h_{12}(together with ω) will make it possible to

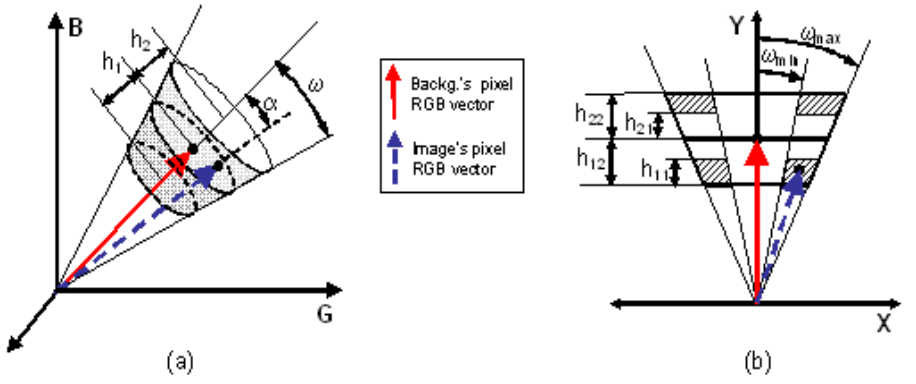

Fig. 4. Background truncated cones associated with a background point (a) in the RGB space and (b) as a projection on the XY plane (the Y axis is made to coincide with the vector RGB of the background point)

delimit another truncated cone volume included in the lower part of the original volume. Analogically, ω_{min} and ω_{max} (together with h_1 and h_2, in figure 4a) will make it possible to delimit a truncated cone crown also included in the original volume. Finally, if we calculate the intersection of these three volumes, we obtain two truncated cone crowns, whose section (striped area) can be seen in figure 4b. The idea is that if the value of these 6 parameters is appropriately chosen, a pixel belonging to a fluctuation noise, by the extreme nearness of its RGB value to the background, will not have much probability of being confined in these two crowns and will have much more probability of being in the remaining original volume. On the other hand, if we admit as a working hypothesis that the number of object pixels that are extremely similar to the background is very low, then we can affirm that if a pixel belongs to the object, there is a high probability that this pixel belongs to one of the two truncated cone crowns defined above.

To tune the value of the six parameters that characterise the operator four probability distributions will be used. On the one hand, in order to estimate the background noise, we will consider how all those pixels are distributed which according to the initial segmentation method do not belong to the foreground of the whole frame under analysis. In the first place, we will calculate this distribution for different values of angle α (see figure 4a), i.e., the angle existing between the vector RGB of a point of the image and the vector RGB equivalent to the background model. Similarly, on the other hand, we will calculate the same distribution, but for the pixels of the ROI that we want to restore (see figure 5a). The two remaining distributions are obtained in the same way but depending on different h values normalised according to the background vector module (see figure 5b).

Comparing the curve slopes in figure 5a enable us to establish that approximately both curve slopes are conserved from 0^o to the angle value 0.7^o. This means that most of the ROI pixels that are within this range belong to the

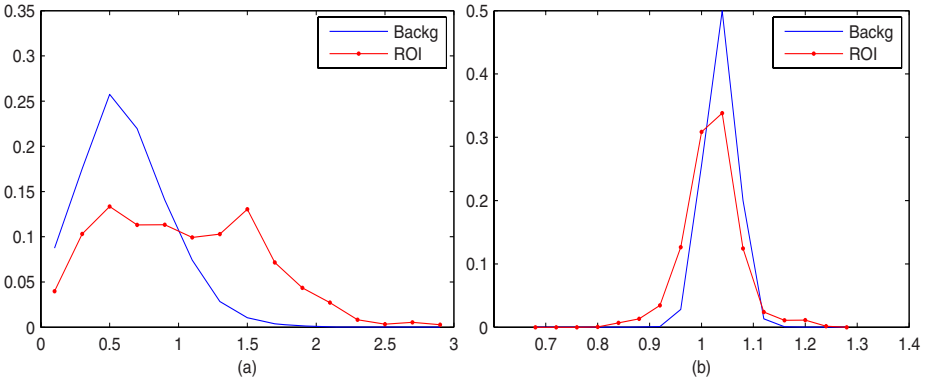

Fig. 5. Probability distributions of the number of points that according to the segmentation stage do not belong to the foreground, (a) depending on the value of angle α and (b) depending on the offset h

fluctuation pixel category. From 0.7^o the trend for both curves is different, thereby indicating the presence of points associated to the object in the ROI. This frontier value enables us to initialise the ω_{min} value. Observe that the ω_{max} value is not critical, a value will simply be chosen that is large enough to guarantee that no object pixel remains outside, for example, $\omega_{max} = 10^o$. The h parameter values are determined by inspecting figure 5b. In this figure it is observed that in the range $(0, 1.05)$, because of the disparity of slopes between the two curves, there is a greater probability of finding an object pixel than a fluctuation pixel in the ROI. Conversely, in the interval $(1.05, 1.09)$, this probability is inverted. Finally, in the interval $(1.09, 2)$ the probability of finding object pixels in the ROI increases once more. In the light of this analysis, for the frame under analysis, $(h_{11}, h_{12}, h_{21}, h_{22}) = (0, 1.05, 1.09, 2)$ will be taken. Observe that the critical h values are now h_{12} and h_{21} because they are the ones that mark the frontier between fluctuation and object pixels. On the other hand, the h_{11} and h_{22} parameters are no longer as critical because they do not mark the frontier limits between both types of pixels, suffice it be to assign them a value small and large enough so as not to leave any object pixel outside.

Finally, figure 6 shows the result of applying all the steps indicated to one of the frames of a scene with a human which presents parts of her body with RGB values very similar to the background and whose distribution curves were those represented in figure 5. Thus, taking the frame indicated in figure 6a as input, a segmentation is done that produces as a result a set of blobs indicated in figure 6b. An analysis at object level of the blobs obtained reveals that there are several boxes associated with the human model with no pixels. The application of the restoration operator to the ROIs associated with these boxes, together with the blobs already existing in the segmentation stage, produces the result shown in figure 6c.

Fig. 6. Result of the ROI restoration process: (a) current frame (b) set of blobs obtained in the segmentation stage (c) set of segmentation blobs after the restoration

4 Conclusions and Future Works

This work presents a knowledge-based segmentation model based on different description levels, which makes it possible to feedback information from the more abstract object level to the blob level to improve the segmentation process results. In order to test the viability of the model, it is instantiated in the two case studies, each of which uses a different resegmentation operator and whose results support the validity of this model.

In future works, the study will focus on developing tasks to diagnose and plan therapy belonging to the object level. Thus, for example, the task of diagnosing implies recognising humans from the human model and also parts of their body. Similarly, at blob level, it will be necessary to refine the already existing operators and develop new operators that make it possible to do the segmentation therapies proposed at object level at this level.

Acknowledgments

The authors would like to thank the CiCYT for financial support via project TIN-2004-07661-C0201 and the UNED project call 2006.

References

[1] T. Horprasert, D. Harwood, L. S. Davis. *A statistical approach for realtime robust background subtraction and shadow detection.* In Proc. of IEEE Frame Rate Workshop, pp 1-19, Kerkyra, Greece, 1999.

[2] C. Stauffer, W. Grimson. *Adaptive background mixture models for real-time tracking.* In Proc. of the IEEE Computer SocietyConference on Computer Vision and Pattern Recognition, pp.246-252, 1999.

[3] A. J. Lipton, H. Fujiyoshi, R. S. Patil. *Moving target classification and tracking from real-time video.* In Proc. of WorkshopApplications of Computer Vision, pages 129-136, 1998.

[4] L. Wang, W. Hu, T. Tan. *Recent developments in human motion analysis*. Pattern Recognition. vol. 36(3), pp. 585-601, March 2003.
[5] E. Y. Kim, S. H. Park. *Automatic video segmentation using genetic algorithms*. Pattern Recognition Letters 27 (11), pp. 1252-1265, 2006.
[6] E. J. Carmona, J. Martínez-Cantos and J. Mira. *Posprocesamiento morfológico adaptativo basado en algoritmos genéticos y orientado a la detección robusta de humanos*. Campus Multidisciplinary in Perception and Intelligence, CMPI-2006, pp. 249-261, Albacete (Spain), July 2006.
[7] J. Martínez-Cantos, E.J. Carmona, A. Fernández-Caballero, M.T. López. *Mejora paramétrica de la interacción lateral en computación acumulativa*. Campus Multidisciplinary in Perception and Intelligence, CMPI-2006, pp. 262-273, Albacete (Spain), July 2006.
[8] R. T. Collins, A. J. Lipton, A. J, T. Kanade. *Special Issue on Video Surveillance*. IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) 2000.
[9] Y. Dedeoglu. *Moving Object Detection, Tracking and Classification for Smart Video Surveillance*. Ph.D. Thesis, 2004.
[10] I. Haritaoglu, D. Harwood, L.S. Davis. *W4: Real-time Surveillance of People and Their Activities*. PAMI, 22(8), pp. 809-830, Aug. 2000.
[11] C. Stauffer, W. Grimson. *Learning Patterns of Activity Using Real-Time Tracking*. IEEE Trans. Pattern Analysis and MachineIntelligence, vol. 22, no. 8, pp. 747-757, 2000.
[12] R. Cucchiara, M. Piccardi, A. Prati. *Detecting Moving Objects, Ghost and Shadows in Video Streams*. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 25, no. 10, pp. 1337-1342, 2003.
[13] E. J. Carmona, J. Martínez-Cantos and J. Mira. *A new video segmentation method of mobile objects based on blob-level knowledge*. Pattern Recognition Letters. (under review), 2007.
[14] C. Dillon, T. Caelli. *Learning Image Annotation: the CITE system*. Videre , 1, 2, pp. 90-123, 1998.
[15] S. Sing and A. Sowmya. *RAIL: Road Recognition from Aerial Images Using Inductive Learning*. In International Archives of Photogrammetry and Remote Sensing, volume (32) 3/1, pp. 367-378, 1998.
[16] J. Tani. *Model-based learning for mobile robot navigation from the dynamical systems perspective*. IEEE Transactions on Systems, Man and Cybernetics, Part B, pp. 421-436, Vol : 26, Issue: 3, Jun 1996.
[17] H. H. Nagel. *Steps toward a cognitive vision system*. AI Mag. 25(2), pp.31-50,2004.
[18] E. Folgado, M. Rincón, E.J. Carmona, M. Bachiller. *A block-based model for monitoring of human activity*. IEEE trans. on Pattern Analysis and Machine Intelligence. (under review). 2007.