3D Neural Model-Based Stopped Object Detection

Lucia Maddalena[1] and Alfredo Petrosino[2]

[1] ICAR - National Research Council
Via P. Castellino 111, 80131 Naples, Italy
lucia.maddalena@na.icar.cnr.it
[2] DSA - University of Naples Parthenope
Centro Direzionale, Isola C/4, 80143 Naples, Italy
alfredo.petrosino@uniparthenope.it

Abstract. In this paper we propose a system that is able to distinguish moving and stopped objects in digital image sequences taken from stationary cameras. Our approach is based on self organization through artificial neural networks to construct a model of the scene background and a model of the scene foreground that can handle scenes containing moving backgrounds or gradual illumination variations, helping in distinguishing between moving and stopped foreground regions, leading to an initial segmentation of scene objects. Experimental results are presented for video sequences that represent typical situations critical for detecting vehicles stopped in no parking areas and compared with those obtained by other existing approaches.

Keywords: moving object detection, background subtraction, background modeling, foreground modeling, stopped object, self organization, neural network.

1 Introduction

Stopped object detection in an image sequence consists in detecting temporally static image regions indicating objects that do not constitute the original background but were brought into the scene at a subsequent time, such as abandoned and removed items, or illegally parked vehicles.

Great interest in the stopped object detection problem has been given by the PETS workshops held in 2006 [8] and in 2007 [9], where one of the main aims has been the detection of *left luggage*, that is luggage that has been abandoned by its owner, in movies taken from multiple cameras. Another example of strong interest in the considered problem is given by the *i-LIDS bag and vehicle detection challenge* proposed in the AVSS 2007 Conference [20], where the attention has been driven on abandoned bags and parked vehicles events, properly defined.

A broad classification of existing approaches to the detection of stopped objects can be given as *tracking-based* and *non tracking-based* approaches. In *tracking-based* approaches the stopped object detection is obtained on the basis

P. Foggia, C. Sansone, and M. Vento (Eds.): ICIAP 2009, LNCS 5716, pp. 585–593, 2009.

of the analysis of object trajectories through an application dependent event detection phase. Such approaches include most of the papers in [8,9,20]. *Non tracking-based* approaches include pixel- and region-based approaches aiming at classifying pixels/objects without the aid of tracking modules, and include [4,12,13,17,19].

Our approach to the problem is non tracking-based. The problem is tackled as *stopped foreground subtraction*, that, in analogy with the background subtraction approach, consists in maintaining an up-to-date model of the stopped foreground and in discriminating moving objects as those that deviate from such model. Both background subtraction and stopped foreground subtraction have the common issue of constructing and maintaining an image model that adapts to scene changes and can capture the most persisting features of the image sequence, i.e. the background and the stationary foreground, respectively. For such modeling problem we adopt visual attention mechanisms that help in detecting features that keep the user attention, based on a self-organizing neural network.

We propose to construct a system for motion detection based on the background and the foreground model automatically generated by a self-organizing method without prior knowledge of the pattern classes. The approach, that is a variation of the one proposed for background modeling [15,16], consists in using biologically inspired problem-solving methods to solve motion detection tasks, typically based on visual attention mechanisms [2]. The aim is to obtain the objects that keep the users attention by referring to a set of predefined features.

The paper is organized as follows. In Section 2 we describe a model-based pixelwise procedure allowing to discriminate foreground pixels into stopped and moving pixels, that is completely independent on the background and foreground models adopted. In Section 3 we describe the model for both background and foreground modeling that we adopted in our experiments. Section 4 presents results obtained with the implementation of the proposed approach and compares them with those obtained by other existing approaches, while Section 5 includes concluding remarks.

2 Stopped Foreground Detection

In this section we propose a model-based approach to the classification of foreground pixels into stopped and moving pixels. A foreground pixel is classified as *stopped* if it holds the same color features for several consecutive frames; otherwise it is classified as *moving*.

Assuming we have a model BG_t of the image sequence background, we compute a function $E(x)$ of color feature occurrences for pixel $I_t(x)$ as follows

$$E(x) = \begin{cases} \min(\tau_s, E(x)+1) \text{ if } & I_t(x) \notin BG_t \text{ and } I_t(x) \in FG_t \\ \max(0, E(x)-1) \text{ if } & I_t(x) \notin BG_t \text{ and } I_t(x) \notin FG_t \\ \max(0, E(x)-k) \text{ if } & I_t(x) \in BG_t \end{cases} \quad (1)$$

where model FG_t of the sequence foreground is iteratively built and updated using image pixels $I_t(x)$ for which $E(x) > 0$.

Every time pixel $I_t(x)$ belongs to the foreground model ($I_t(x) \in FG_t$), $E(x)$ is incremented, while it is decremented if it does not belong to the foreground model. The maximum value τ_s for $E(x)$ corresponds to the *stationarity threshold*, i.e. the minimum number of consecutive frames after which a pixel assuming constant color features is classified as stopped. The value for τ_s is chosen depending on the desired responsiveness of the system.

On the contrary, if pixel $I_t(x)$ is detected as belonging to the background ($I_t(x) \in BG_t$), $E(x)$ is decreased by a factor k. The decay constant k determines how fast $E(x)$ should decrease, i.e. how fast the system should recognize that a stopped pixel has moved again. To set the alarm flag off immediately after the removal of the stopped object, the value of decay should be large, eventually equal to τ_s.

Pixels $I_t(x)$ for which $E(x)$ reaches the stationarity threshold value τ_s are classified as stopped, and therefore the set ST_t defined as

$$ST_t = \{FG_t(x) : E(x) = \tau_s\}$$

supplies a model for the stopped objects, while the remaining part of FG_t represents moving objects.

The described procedure is completely independent on the model adopted for the scene background and foreground. The model that we have adopted for the background and the foreground will be described in the following section.

3 Background and Foreground Update

For background and foreground modeling we adopt here a variation of the model presented in [15,16], according to which a self-organizing neural network, organized as a 3-D grid of neurons, is built up. Each neuron computes a function of the weighted linear combination of incoming inputs, with weights resembling the neural network learning, and can be therefore represented by a weight vector obtained collecting the weights related to incoming links. An incoming pattern is mapped to the neuron whose set of weight vectors is most similar to the pattern, and weight vectors in a neighborhood of such node are updated.

Specifically, for each pixel $p_t = I_t(x)$ we build a neuronal map consisting of L weight vectors $c^l(p_t), l = 1, \ldots, L,$. Each weight vector $c^l(p_t)$ is represented in the HSV colour space, that allows to specify colours in a way that is close to human experience of colours, and is initialized to the HSV components of the corresponding pixel of the first sequence frame $I_0(p_t)$. The complete set of weight vectors for all pixels of an image I with N rows and M columns is organized as a 3D neuronal map $\tilde{B}$ with N rows, M columns, and L layers. An example of such neuronal map is given in Fig. 1, which shows that for each pixel $p_t = I_t(x)$ we have a weight vector $\tilde{B}_t(x) = (c^1(p_t), c^2(p_t), \ldots, c^L(p_t))$.

By subtracting the current image from the background model $\tilde{B}$, each pixel p_t of the t-th sequence frame I_t is compared to the current pixel weight vectors to determine if there exists a weight vector that matches it. The best matching weight vector is used as the pixel's encoding approximation, and therefore p_t is

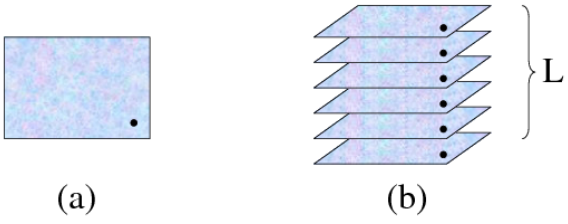

Fig. 1. A simple image (a) and the modeling neuronal map with L layers (b)

detected as foreground if no acceptable matching weight vector exists; otherwise it is classified as background.

Matching for the incoming pixel $p_t = I_t(x)$ is performed by looking for a weight vector $c^b(p_t)$ in the set $\tilde{B}_t(x) = (c^1(p_t), \ldots, c^L(p_t))$ of the current pixel weight vectors satisfying:

$$d(c^b(p_t), p_t) = \min_{i=1,\ldots,L} d(c^i(p_t), p_t) \leq \varepsilon \tag{2}$$

where the metric $d(\cdot)$ and the threshold ε are suitably chosen as in [15].

The best matching weight vector $c^l(p_t) = \tilde{B}_t(x)$ belonging to layer l and all other weight vectors in a $n \times n$ neighborhood N_{p_t} of $c^l(p_t)$ in the l-th layer of the background model $\tilde{B}$ are updated according to selective weighted running average:

$$\tilde{B}^l_t(x) = (1 - \alpha_t(x))\tilde{B}^l_{t-1}(x) + \alpha_t(x) I_t(x), \quad \forall x \in N_{p_t} \tag{3}$$

where $\alpha_t(x)$ is a learning factor, later specified, belonging to [0,1] and depends on scene variability. If the best match $c^b(p_t)$ satisfying (2) is not found, the background model $\tilde{B}$ remains unchanged. Such selectivity allows to adapt the background model to scene modifications without introducing the contribution of pixels not belonging to the background scene.

Spatial coherence is also introduced in order to enhance robustness against false detections. Let $p = I(x)$ the generic pixel of image I, and let N_p a spatial square neighborhood of pixel $p \in I$. We consider the set Ω_p of pixels belonging to N_p that have a best match in their background model according to (2), i.e.

$$\Omega_p = \{q \in N_p : d(c^b(q), q) \leq \varepsilon\} \ .$$

In analogy with [5], the *Neighborhood Coherence Factor* is defined as:

$$NCF(p) = \frac{|\Omega_p|}{|N_p|}$$

where $|\cdot|$ refers to the set cardinality. Such factor gives a relative measure of the number of pixels belonging to the spatial neighborhood N_p of a given pixel p that are well represented by the background model $\tilde{B}$. If $NCF(p) > 0.5$, most of the pixels in such spatial neighborhood are well represented by the

background model, and this should imply that also pixel p is well represented by the background model. Values for $\alpha_t(x)$ in (3) are therefore expressed as

$$\alpha_t(x) = M(p_t)\ \alpha(t)\ w(x), \quad \forall x \in N_{p_t}, \tag{4}$$

where $w(x)$ are Gaussian weights in the neighborhood N_{p_t}, $\alpha(t)$ represents the learning factor, that is the same for each pixel of the t-th sequence frame, and $M(p_t)$ is the crisp hard-limited function

$$M(p_t) = \begin{cases} 1 & \text{if} \quad NCF(p_t) > 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

that gives the background/foreground segmentation for pixel p_t, also taking into account spatial coherence.

The described model $\tilde{B}_t$ has been adopted for both the background model BG_t and the foreground model FG_t described in Section 2 for the classification of stopped and moving pixels.

4 Experimental Results

Experimental results for the detection of stopped objects using the proposed approach have been produced for several image sequences. Here we present results on parked vehicle sequences *PV-easy*, *PV-medium*, and *PV-hard* belonging to the publicly available *i-LIDS 2007* dataset[1]. Such scenes represent typical situations critical for detecting vehicles in no parking areas, where the street under control is more or less crowded with cars, depending on the hour of the day the scene refers to. For all the scenes the main difficulty is represented by the strong illumination variations, due to clouds frequently covering and uncovering the sun. For the purpose of the AVSS 2007 contest [20], the no parking area is defined as the main street borders, and the stationarity threshold is defined as $\tau_S = 1500$. This means that an object is considered irregularly parked if it stops in the no parking area for more than 60 seconds (scenes are captured at 25 fps).

Results obtained for sequence *PV-easy* are reported in Fig. 2. Since an empty initial background is not available for this scene, we artificially inserted 30 empty scene frames at the beginning of the sequence (starting from frame 251) in order to not be puzzled with bootstrapping problems for background modeling. As soon as the white van stops (starting from frame 2712) the function $E(x)$ described in Section 2 starts incrementing for pixels belonging to the van; such pixels are inserted into the foreground model FG_t and used for the model update. After approximately τ_S=1500 frames, $E(x)$ reaches the stationarity threshold τ_S, thus signaling the first stopped object (frame 4119). From this moment till to the end of the stopped car event the stopped object model allows to distinguish moving objects from the stopped object. When the van leaves again (from frame 4875), the part of the scene uncovered by the van is again recognized as belonging to the background model, and previously stopped pixels are deleted from the stopped object model.

[1] ftp://motinas.elec.qmul.ac.uk/pub/iLids/

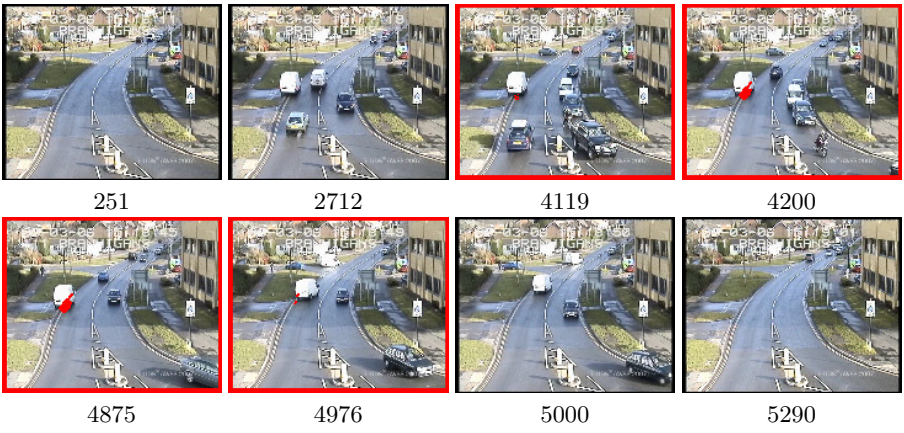

Fig. 2. Detection of stopped objects in sequence *PV-easy*. The van first stops in frame 2712. The first stationary object is detected in frame 4119; further stationary pixels are later detected, as shown in frames 4200 and 4875. The van is detected as a stationary object till to frame 4976, and no more stopped objects are detected till to frame 5290 (end of the sequence).

It should be stressed that illumination conditions have changed quite a bit between the stopping and leaving of the van. This results in an uncovered background very different from the background that was stored before the van stop. Our background model, however, could recognize it again as background since it includes a mechanism for distinguishing shadows and incorporating them into the background model (here not described for space constraints).

Moreover, it should be clarified that we do not identify the whole white van, but only its part belonging to the no parking area, since we restrict our attention only to the street including the no parking area (masking out the remaining part of the scene).

Results obtained for sequence *PV-medium* are reported in Fig. 3. In this case the empty scene available at the beginning of the sequence (starting from frame 469) allows to train a quite faithful background model. The role of the FG_t model is quite clear here, since many cars pass in front of the stopped car (e.g. in frame 2720, where the white car covers the stopped car) and, without a comparison with such model, could be taken erroneously as part of the stopped car.
Analogous considerations can be drawn by looking at results obtained for sequence *PV-hard*, whose images have not been reported here for space constraints.

We compared results obtained with our approach with those obtained with other approaches for the same sequences. Specifically we considered results obtained by four tracking-based approaches to the detection of stopped objects presented by Boragno et al. [1], who employ a DSP-based system for automatic visual surveillance where block matching motion detection is coupled with MOG-based foreground extraction; Guler et al. [11], who extend a tracking system, inspired by the human visual cognition system, introducing a stationary object

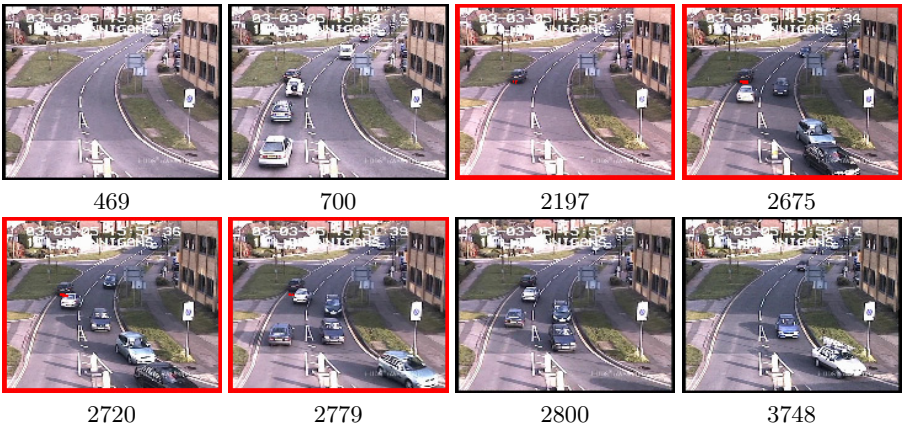

Fig. 3. Detection of stopped objects in sequence *PV-medium*. The car first stops in frame 700. The first stationary object is detected in frame 2197; further stationary pixels are later detected, even if the stopped object is occluded by foreground pixels (e.g. in frame 2720, where the white car covers the stopped car). The car is detected as stopped till to frame 2779, and no more stopped objects are detected till to frame 3748 (end of the sequence).

model where each region represents hypotheses stationary objects whose associated probability measures the endurance of the region; Lee et al. [14], who present a detection and tracking system operating on a 1D projection of images; and by Venetianer et al. [23], who employ an object-based video analysis system featuring detection, tracking and classification of objects.

In Table 1 we report stopped object event start and end times provided with the ground truth and those computed with all considered approaches. Corre-

Table 1. Comparison of ground truth (GT) stopped object event start and end times (in minutes) with those computed with our approach and with different approaches reported in [1,11,14,23], for considered sequences. Related absolute errors (ε) are expressed in seconds; total error is computed as the sum of absolute errors over the three sequences.

Sequence	**Event**	**GT**	**Our**	ε	**[1]**	ε	**[11]**	ε	**[14]**	ε	**[23]**	ε
PV-easy	Start	02:48	02:45	3	02:48	0	02:46	2	02:52	4	02:52	4
”	End	03:15	03:19	4	03:19	4	03:18	3	03:19	4	03:16	1
PV-medium	Start	01:28	01:28	0	01:28	0	01:28	0	01:41	13	01:43	15
”	End	01:47	01:51	4	01:55	8	01:54	7	01:55	8	01:47	0
PV-hard	Start	02:12	02:12	0	02:12	0	02:13	1	02:08	4	02:19	7
”	End	02:33	02:34	1	02:36	3	02:36	3	02:37	4	02:34	1
Total error				12		15		16		37		28

sponding absolute errors show that generally our approach compares favorably to the other approaches, and this is still more evident if we consider the total error over the three considered sequences. It should be emphasized that, since our approach to stopped object detection is pixel-based and no region-based post-processing is performed in order to identify objects, in our case a stopped object event starts as soon as a single pixel is detected as stopped and ends as soon as no more stopped pixels are detected.

5 Conclusions

The paper reports our approach to the problem of *stopped foreground subtraction*, consisting in maintaining an up-to-date model of the stopped foreground and in discriminating moving objects as those that deviate from such model. For such modeling problem we adopt visual attention mechanisms that help in detecting features that keep the user attention, based on a 3D self-organizing neural network, without prior knowledge of the pattern classes. The approach consists in using biologically inspired problem-solving methods to solve motion detection tasks, typically based on visual attention mechanisms [2]. The aim is to obtain the objects that keep the user attention in accordance with a set of predefined features, by learning the trajectories and features of moving and stopped objects in a self-organizing manner. Such models allow to construct a system able to detect motion and segment foreground objects into moving or stopped objects, even when they appear superimposed.

References

1. Boragno, S., Boghossian, B., Black, J., Makris, D., Velastin, S.: A DSP-based system for the detection of vehicles parked in prohibited areas. In: [20]
2. Cantoni, V., Marinaro, M., Petrosino, A. (eds.): Visual Attention Mechanisms. Kluwer Academic/Plenum Publishers, New York (2002)
3. Cheung, S.-C., Kamath, C.: Robust Techniques for Background Subtraction in Urban Traffic Video. In: Proceedings of EI-VCIP, pp. 881–892 (2004)
4. Collins, R.T., Lipton, A.J., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N., Hasegawa, O., Burt, P., Wixson, L.: A System for Video Surveillance and Monitoring. The Robotics Institute, Carnegie Mellon University, Tech. Rep. CMU-RI-TR-00-12 (2000)
5. Ding, J., Ma, R., Chen, S.: A Scale-Based Connected Coherence Tree Algorithm for Image Segmentation. IEEE Transactions on Image Processing 17(2), 204–216 (2008)
6. Elhabian, S.Y., El-Sayed, K.M., Ahmed, S.H.: Moving Object Detection in Spatial Domain using Background Removal Techniques - State-of-Art. Recent Patents on Computer Science 1, 32–54 (2008)
7. Elgammal, A., Duraiswami, R., Harwood, D., Davis, L.S.: Background and Foreground Modeling Using Nonparametric Kernel Density Estimation for Visual Surveillance. Proceedings of the IEEE 90(7), 1151–1163 (2002)
8. Ferryman, J.M. (ed.): Proceedings of the 9th IEEE International Workshop on PETS, New York, June 18 (2006)

9. Ferryman, J.M. (ed.): Proceedings of the 10th IEEE International Workshop on PETS, Rio de Janeiro, Brazil, October 14 (2007)
10. Fisher, R.B.: Change Detection in Color Images, http://homepages.inf.ed.ac.uk/rbf/PAPERS/iccv99.pdf
11. Guler, S., Silverstein, J.A., Pushee, I.H.: Stationary objects in multiple object tracking. In: [20]
12. Herrero-Jaraba, E., Orrite-Urunuela, C., Senar, J.: Detected Motion Classification with a Double-Background and a Neighborhood-based Difference. Pattern Recognition Letters 24, 2079–2092 (2003)
13. Kim, K., Chalidabhongse, T.H., Harwood, D., Davis, L.S.: Real-time Foreground-Background Segmentation using Codebook Model. Real-Time Imaging 11, 172–185 (2005)
14. Lee, J.T., Ryoo, M.S., Riley, M., Aggarwal, J.K.: Real-time detection of illegally parked vehicles using 1-D transformation. In: [20]
15. Maddalena, L., Petrosino, A.: A Self-Organizing Approach to Background Subtraction for Visual Surveillance Applications. IEEE Transactions on Image Processing 17(7), 1168–1177 (2008)
16. Maddalena, L., Petrosino, A., Ferone, A.: Object Motion Detection and Tracking by an Artificial Intelligence Approach. International Journal of Pattern Recognition and Artificial Intelligence 22(5), 915–928 (2008)
17. Patwardhan, K.A., Sapiro, G., Morellas, V.: Robust Foreground Detection in Video Using Pixel Layers. IEEE Transactions on PAMI 30(4), 746–751 (2008)
18. Piccardi, M.: Background Subtraction Techniques: A Review. In: Proceedings of IEEE Int. Conf. on Systems, Man and Cybernetics, pp. 3099–3104 (2004)
19. Porikli, F., Ivanov, Y., Haga, T.: Robust Abandoned Object Detection Using Dual Foregrounds. EURASIP Journal on Advances in Signal Processing (2008)
20. Proceedings of 2007 IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2007). IEEE Computer Society (2007)
21. Radke, R.J., Andra, S., Al-Kofahi, O., Roysam, B.: Image Change Detection Algorithms: A Systematic Survey. IEEE Transactions on Image Processing 14(3), 294–307 (2005)
22. Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower: Principles and Practice of Background Maintenance. In: Proceedings of the Seventh IEEE Conference on Computer Vision, vol. 1, pp. 255–261 (1999)
23. Venetianer, P.L., Zhang, Z., Yin, W., Lipton, A.J.: Stationary target detection using the objectvideo surveillance system. In: [20]
24. Wren, C., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: Real-Time Tracking of the Human Body. IEEE Transactions on PAMI 19(7), 780–785 (1997)