Statistic and Knowledge-based Moving Object Detection in Traffic Scenes

R.Cucchiara, C.Grana*, M.Piccardi^, A.Prati*,*
*D.S.I. University of Modena, via Campi 213/b
41100 Modena, Italy {rita, prati, grana}@dsi.unimo.it
^D.I University of Ferrara, via Saragat 1 44100 Ferrara, Italy mpiccardi@ing.unife.it

Abstract

Vision-based systems for traffic surveillance have an impressive spread both for their practical application and interest as research issue. The most common approach used for vision-based traffic surveillance consists of a fast segmentation of Moving Visual Objects (MVOs) in the scene together with an intelligent reasoning module capable of identifying, tracking and classifying the MVOs in dependency of the system goal. In this paper we describe our approach for MVOs segmentation in an unstructured traffic environment. We consider complex situations with moving people, vehicles, infrastructures that have different aspect model and motion model. In this case we define a specific approach based on background subtraction with a statistic and knowledge-based background update. We show many results of real-time tracking of traffic MVOs in outdoor traffic scene such as roads, parking area, intersections, entrance with barriers.

1. Introduction

The availability of techniques and computation resources for processing images at frame rate opens big opportunities in application fields interested in surveillance and scene monitoring. Among them, one of the most promising is the context of Intelligent Transportation Systems, where fast and dynamic decision systems must be supported by intelligent and powerful information extraction units such as vision-based surveillance systems.

Many complete architecture of vision systems have been proposed aimed at extracting Moving Visual Object (MOVs) and reasoning about their flow in the observed scene. With the term *Visual Object* (VO) we define foreground elements of the scene, as image patterns that can be perceived by an observer as different from background. In dynamic scenes as traffic ones are, we should extend the observation to *Moving Visual Objects* (MVOs) that are VOs characterized by a non null (real or apparent) motion during the time of observation. We use this general definition to cover all objects of interest in traffic scenes, such as moving vehicles, people, stopped vehicles at intersections or in a parking area or moving infrastructures like barriers at toll-gates.

The most common approach of vision systems consists of two parts, more or less independent: a first, possibly fast, segmentation of moving visual objects (MVOs) in the scene or perceptual level, and a further reasoning level capable of identifying, tracking and classifying the MVOs in dependency of the system goal [1,2,3]. The basic but critical step is MVO segmentation. Many approaches have been proposed not only for traffic scenes The common aspect is that MVOs must be easy perceivable as different from background. Often the saliency of MOVs is based on visual features only (colour, texture, luminance gradient, area in pixel, motion field); in other cases the model of target MOVs is used for improving object segmentation and identification. In this work we do not consider model-based vision in the strict sense (i.e. object models do not guide the search) but we aim to exploit acquired knowledge of MVOs (e.g. people vs. vehicles) for improving segmentation.

In the paper we present our approach for MVOs segmentation in an unstructured traffic environment. Unlike many other works in the field [4,5], we do not address the relatively simpler road surveillance case, where one class of moving objects (i.e. vehicles) can be detected along some fixed observation directions . We consider instead more complex situations with moving people, vehicles, infrastructures that have a different shape, speed, trajectory. Possible applications are surveillance of public zone entrances such as monitored barriers (as Fig.1.a), monitoring of large urban parking or surveillance at highway toll-gates. (see Fig. 4).

In this context we have defined a specific approach called *S&KB Background update* based on background subtraction with a statistically adaptive and knowledge-based selective background updating. The statistical adaptation copes with the changing appearance of the scene during time, while the knowledge-based selection gives a feedback to the perceptual level in order to improve separation of foreground objects from background. In this way, we are able during time to decide when a stopped MVO should become part of the background (e.g. a parked vehicles) or not (for instance, a car stopped in proximity of a train barrier or at a traffic light), or when a moving pattern is only a false positive (apparent motion detection).

2. MVO segmentation techniques

MVO segmentation in images is based on the extraction of image points in accordance with some visual and/or motion features; then, segmentation consists of the aggregation of these image points into objects, in function of their spatial connectivity and the homogeneity of computed features. In complex outdoor scenes, especially in the daytime, segmentation based on static visual features only is not practicable, since images contain noise, many non interesting details and object occlusions. Therefore, neither standard approaches based on edge detection and filling, nor region-based segmentation on colour and texture can be exploited [3]. The most important feature remains motion of points.

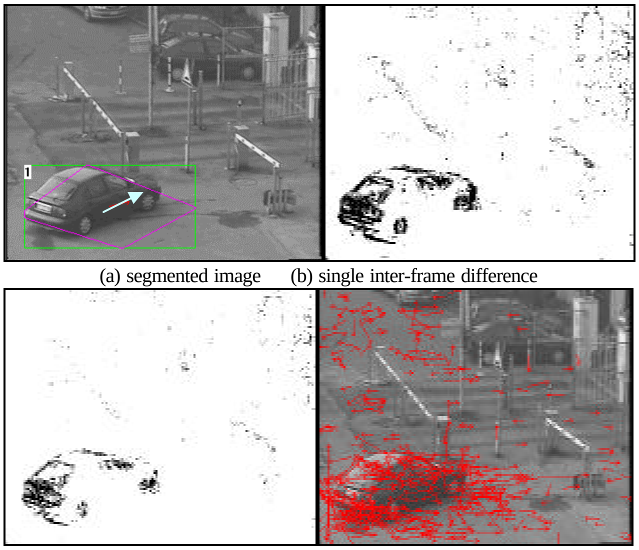
The best formalized image processing techniques for motion computation is based on *optical flow* (OF) [4,5,6]. Given **I**(x,y,t) the image frame at the time t, and using the Brightness Constancy Equation, $\frac{\partial I}{\partial x}\frac{dx}{dt}+\frac{\partial I}{\partial y}\frac{dy}{dt}+\frac{\partial I}{\partial t}=0$, many different solutions have been proposed for numerically solving the differential system formed by the equation and its boundary conditions and for adding different constraints. Fig. 1.d shows OF vectors, computed with the method proposed in [6]. OF is a powerful method used in wide application ranges, but seems not suitable enough in our complex and unstructured environment; if we try to use OF vectors for guiding clustering and for labelling points in object segmentation, we found some drawbacks :

a) OF is very time consuming and must be largely approximated for a real-time computation;
b) it is correct only if the basic condition of gradient constancy is satisfied; therefore grouping and labelling on the basis of OF feature is very hard for little and fragmentary objects such as people in wide-area images;
c) for large and uniform objects (e.g. monochromatic vehicles), many internal points can not be characterized by a significant optical flow vector.

Nevertheless, OF is effective when applied to small regions of interest, for characterizing the motion of already segmented objects (see the OF arrow superimposed on the vehicle in Fig.1a). Although OF is the most complete approach for point motion measurement, many other approaches based on *motion detection* only are generally exploited in traffic scenes. They are divided in two classes based on inter-frame distance and background subtraction.

In the first class, the temporal derivative of luminosity only is considered; a point is moving if its luminosity changes between two consecutive frames. These techniques have been proposed since many years with many further improvements, such as accumulative difference [10] or with double-difference between three frames [11,3]. Figg.1.b and 1.c show single difference and double difference. The second one is more precise in locating real objects. Inter-frame difference is interesting, since is very fast in frame-rate application. The drawback is that often these techniques do not produce close object contours and therefore morphological operations (e.g. dilation) must be added or an edge closure guided by the edge gradient [3]. However, these techniques can be useful when objects' motion is mainly along a known direction, as for instance when a road is monitored with coming or outgoing vehicles.

The most adopted technique is an extension of frame-difference, that is *background suppression*, consisting of subtracting a reference frame of background from the current frame. The main difference among the many proposals on background suppression is in the method of background update. The most trivial approach is the use of a static frame without MVOs as reference frame. In this case a first problem is that may not be easy to find a scene without moving objects and, in any case, is not automated; more important, during the day in an outdoor environment, luminance changes, atmospheric conditions and different day hours require a dynamic background upgrade. A first improvement consists of a dynamic background upgrade with an average of the last number of frames, which is a simple statistical assumption that the presence of moving objects will be removed by the averaging operation .



(a) segmented image (b) single inter-frame difference

(c) double inter-frame difference (d) optical flow

Figure 1. Motion detection comparison.

Many approaches suggest an adaptive background update, based on a simple adaptive filter [12] or the Kalman filter [13,2]. Other proposals define the background update based on statistical assumptions, by modelling the background as a pre-defined statistical model [14,15,16]. In [16], Gaussian distributions model the background in order to cope with small and frequent luminance variations, typical for instance of tree leaf in wind conditions. This method allows maintaining a long-term background update, known the statistical distribution along the time. Finally, some authors propose selective update, known the motion of single points [15]. Our approach follows the way of combining both statistical model and selectivity according to a computed environment knowledge.

3. The proposed segmentation approach

Our approach aims to recognize different types of MVOs, namely vehicles, people, groups of people, moving infrastructure (e.g. barrier bars) in an unstructured and unknown urban traffic environment. The basic assumptions are

i) the background is unknown but is not in motion. Luminance can change, even considerably, due to atmospheric conditions, day hours, shadows. Limited, high-frequency camera motions (due for instance to the wind) must be corrected in a pre-processing phase);

ii) considered MVOs have different patterns and speed and are not a-priori known; this means that they may have different aspect: people and cycles have a non uniform textures in a small area and low motion; vehicles can have some uniform colour areas and some differing (e.g. the glasses), a variable size (cars, tracks…) and variable speed; moving barriers are thin and non uniform and with a pure rotation motion; vehicles and people can move with whatever trajectory;

iii) we look for a fast, real time process able to operate at a quasi-frame rate without special purpose devices. Last generation PCs and standard Windows OS- based programs written in C are used.

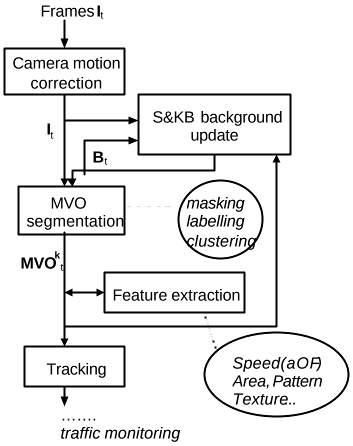


Figure 2. The architecture of the segmentation level.

The architecture of the segmentation level consists of several steps, that must be executed in real time (see Fig. 2). The initial camera motion correction process uses a calibrated fixed point in the frame and a frame-by-frame correlation for adjusting limited but unavoidable camera movements. The main process of MVO segmentation detects moving points by background subtraction; points are used as a mask in order to select pixels of interest in the frame. Labelling and a simple clustering are then performed, with the final goal to segment individual blobs corresponding to MVOs. Many visual features are computed such as oriented extent, area, inertia moments, textures, average speed, grey levels and colour histogram. The average MVO speed is computed as the average optical flow vector. These features are exploited in the further MVO tracking and classification. Actually, tracking could be provided based on area and speed only. Nevertheless, the other features are used in the higher level reasoning module for correcting ambiguities, and handling more general traffic control. The dynamically acquired knowledge of MVO is exploited in the S&KB background update process, described in the next section.

4. Statistic & Knowledge-based background update

The background image (i.e. the scene without MVOs) contains pixels which luminance value can be modelled as a statistical process. The computation of background image[1] $\boldsymbol{B}$ takes into account in an adaptive manner the information of pixels in previous frames.

Our approach, called S&KB background update, is based on two parts: a *statistic update* with *knowledge-based improvement*. Let us consider the first part.

I) <u>*S-background*</u>

Since background is a–priori unknown, we compute it by defining a background pixel as *the statistically more probable during time*. The assumption is that, apart from MVOs, the intensity we can sample in the scene is due to the background points. An example is the point highlighted in Fig. 3.a which luminance variation during 1179 frames (sampled at a frame time Tf=1/10 s) is shown in Fig. 3.e. The ordinate is the grey level value; the two minima indicate the luminance of a vehicle (Fig. 3.b) and a group of two people (Fig 3.c) passing through that point. Apart from these periods, the luminance is about fixed with local variations due to sensors and small camera motions. Thus we define the background at time t as

$$\mathbf{B}_t = U\,(\mathbf{I}_t,\ \mathbf{I}_{t-\Delta t},\ldots,\mathbf{I}_{t-(n-1)\Delta t}) \qquad (1)$$

U is the update function, $\mathbf{I}_k$ is the image frame sampled at the time k. The function consider n frames, sub-sampled at time intervals D (that is a multiple of Tf). Thus nD is the window size of the observation time.

For improving the update we include a history of previous computed background with a term $w_b\mathbf{B}_{t-1,}$ being w_b an adequate weight:

$$\mathbf{B}_t = U\,(\mathbf{I}_t,\ \mathbf{I}_{t-\Delta t},\ldots,\mathbf{I}_{t-(n-1)\Delta t,}\ w_b\mathbf{B}_{t-1}) \qquad (2)$$

that we call *S-background* since is only due to statistical computation.

Let us discuss the choice of the U update function and the time parameters. In Fig. 3.e the mean, median and mode statistical functions in this long-term period are shown. The mean, although the most immediate value to compute, is not significant enough if the n number is small and the difference in luminance between background and moving objects is high. This can be seen in the particular enlarged in Fig. 3.f. The most correct value is given by the *mode* filter, that computes, by definition, the most probable luminance value as the maximum of the probability distribution. The a-posteriori distribution is shown in Fig. 3.d. Therefore as proposed in [11, 12], the U function should be the *mode* function.

Our approach aims to cope with strictly real-time requirements, so that many processes (from motion detection to recognition) should be performed efficiently; in our case, we should be capable to have a frame processed every 1/10 s. At the same time, computing the mode and the median in an incremental way (by eliminating the oldest pixel value when a new one is available) forces to keep in memory enormous data structures (a linked list with n nodes for each image pixel). Therefore, we are not allowed to use a high n value of samples.

[1] We use a bold notation for the vectors, (like images) and the correspondent normal notation for the vector element (i.e. the single pixel).

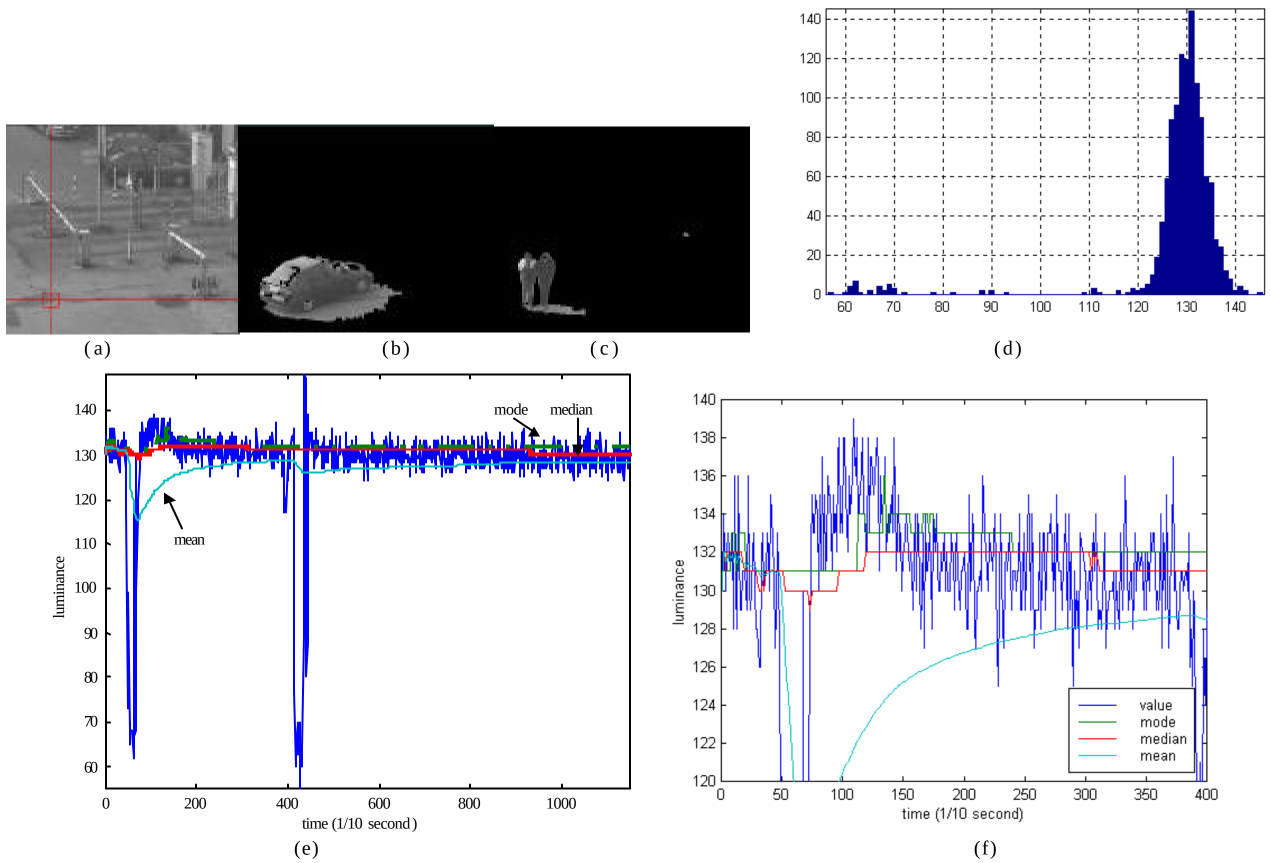


Figure 3.The S-background

We could decide either to decrease *n*, or to increase Δt (starting from its minimum value *Tf*). Decreasing *n* means to become more reactive to luminance change and to have a short history. This causes the erroneous inclusion in S-background of parts of moving objects and may create many false positives (blobs recognized as MVOs since they differ from background). This is particularly true for slow MVOs, such as people chatting or cars incoming to barriers. On the contrary, increasing Δt means sub-sampling the observation window. This means that the distribution probability could have a less evident peak value (a more spread maximum); in particular, with a small *n* value, the value of background may not be the maximum of the distribution. In this case, the *median* empirically approximates better the background. For this reason, even if with a statistically significant n and for lim Δt→*0* we should use mode filter, we adopt the *median* filter.

Thus for each point I∈ **I**, we compute each S-background point B ∈ **B** as

$$B_t = \mathrm{Median}\,(I_t,\ I_{t-\Delta t},\ldots,\ I_{t-(n-1)\Delta t},\ 2B_{t-1}) \qquad (3)$$

In our experiments we found an optimal value of n=9 only and Δt varying between 10*Tf* and 50 *Tf*.

We fixed n=9 for computational constraints so that each background update consists of ordering nine values (plus old background) for each pixel. The weight of old background has been empirically fixed to 2. As example, in Table 1 we report the approximation provided by the median computed with various *n* and Δt parameters w.r.t. the correct mode value computed on all the previous frames.

Parameter	Maximum deviation	Average deviation
n=100 Δt=1*Tf*	2	0,5292
n =50 Δt=1 *Tf*	13	0,7433
n=20 Δt=1 *Tf*	66	2,5688
n =50 Δt=2 *Tf*	2	0.6454
n =20 Δt=5 *Tf*	3	0.6797
n=10 Δt=10 *Tf*	4	0.6673

Table 1

With a very small *n* (10-20), some problems on background arise. In particular if Δt is too short the background could easily include parts of moving objects.

Fig. 4 shows frames from a toll-gate of the Italian freeways. In frame 35 a car is passing (Fig. 4.a); in frame 290 another car is approaching and a people is passing through (Fig.4.d). Wrong S-backgrounds at frame 35 computed with a median filter with a too small Δt are shown in Figg. 4.b (*n*=9 and Δt =1 *Tf*) and Fig.4.c (*n*=9 and Δt= 5*Tf*): Δt must be increased to 20 *Tf* in order to have in this frame a correct S-background (not shown). Fig.4.e shows a wrong S-background for frame 290 where parts of both the people and the car are included in the background.

What are the consequences of wrong background computation? When background subtraction is performed with a wrong S-background the detected MVO could be composed by two parts: one due to the true moving object different from background, and one part of a "virtual moving object" due to the fact that the background is not correct. The consequence is that the area of MVO could be wrong (very large) and also all the other computed visual features are not correct. Moreover, other difficulties arise since separate moving objects are easily confused into a single one.

For overcoming these limitations Δt must be high (e.g. 20 or 30*Tf* for vehicles approaching barrier). In this case, if the probability of MVO is not too high, it is easier to compute a correct background but the responsiveness is low. The system is not reactive enough and background is updated after a longer period.

Finally, the most critical aspect to adopt S-background only is that the optimal value of Δt is not estimable a priori and may vary for different frame of the same sequence, and also for different MVOs in the same frame. Therefore an only statistical method, is not suitable in complex environments where people, vehicles, barriers, cycles and so on can concurrently happen.

II) <u>*S&KB background*</u>

A possible way to improve this mechanism is to provide a *selective* update that computes a new background value only if a point is not marked as motion point, as in [12]. The advantage is that is it possible to use a short Δt and a small *n* without the risk of including moving objects in the background. The selectivity calls for a feedback after the segmentation process and therefore obviously takes into account all possible errors and difficulties of the higher level of computation.

A specific problem arises whenever moving objects are stopped for a long time and become part of the background. Then, when these objects start again, a false positive is detected in the area where they were stopped. This will persist for all the following frames, preventing the area to be updated in the background image forever.

In order to cope with this critical problem, in [12] a long-term non selective background update is combined with a short term selective update. In this selective approach a point detected as a moving point is not included in the background computation anymore; therefore, whenever a false positive is detected it remains in the short-term background but its corrected by the long-term background.

Our approach differs since we do not reason on single moving points but on detected and recognized moving objects. We exclude *from the S&KB background update y those points that t belong to detected MVOs and exhibit a non null value of the motion feature*. Therefore, after a MVO is detected, its average optical flow is computed. This value is used for deciding if the MVO is in true motion or not. Thus we define S&KB background as

$$\begin{aligned} B_t &= B_{t-1} \text{ if } I_t \in MVO^k_t \text{ and } OF^k > TH \text{ for any } k; \\ &= \text{Median}\,(I_t, I_{t-\Delta t}, \ldots, I_{t-(n-1)\Delta t}, 2B_{t-1}) \text{ otherwise} \qquad (4) \end{aligned}$$

In all points corresponding to true moving objects (for any kth MVO with an optical flow value greater than a threshold TH) S&KB background is not updated, while in the other points the S-background is computed.

For instance a parking car, a barrier bar, a stopping people or a virtual object, as seen above, are characterized by a null motion field (i.e., below a given threshold). A later classification could discriminate between real objects and virtual objects and adopt different policies; for instance a timeout for deciding when a MVO should be regarded as part of background can be adopted.

The advantage of our approach is evident in Fig. 5. Fig.5.a represents the background with three closed barriers. In Fig.5.b the bar is arisen and a car is passing. In this frame four MVOs are detected, the car, the people, the rising bar and a virtual bar in the horizontal position (in the circled area). This one is a false positive and can be discriminated since its OF is null. Therefore its points will be included in the background update. Instead the points of the other three MVOs will be masked and conversely used in the further tracking phase (see Fig. 5.c).

The selectivity itself allows short observation window (nΔt) but presents many false positives; the knowledge-based update improvement also abates the false positives. In Fig. 4.e the correct background computed with the S&KB approach with n=9 Δt =1 is shown.

5. Conclusion

This paper presents a novel approach for moving object segmentation with an improved technique for background computation, called S&KB background update. It exploits both statistical properties of background points and the knowledge of the already segmented moving objects. This S&KB approach allows to

a) use a limited number of frames for background update, suitable for real-time computation;
b) the frame can be sampled with a short time interval, allowing good responsiveness in background adaptation;
c) the selectivity together with the knowledge of average motion of detected objects abates the false positive typical of highly responsive background update.

We thank the Italian freeway company, Autostrade S.p.A., for providing us with image sequences from toll-gates.

References

[1] Koller, D., Weber, J., Huang, T., Malik, J., Ogasawara, G., Rao, B., Russel, S., “Towards Robust Automatic Traffic Scene Analysis in Real-Time”, Proc. Int’l Conf. Pattern Recognition, pp. 126-131, 1994.

[2] R.Cucchiara, P.Mello, M. Piccardi, Image Analysis and Rule-Based Reasoning for a Traffic Monitoring, Proceedings 1999 IEEE/IEEJ/JSAI International Conference on Intelligent Transportation Systems (ITSC99)

[3] R.Cucchiara, M.Piccardi, Vehicle Detection Under Day and Night Illumination, Proc. of 3rd International ICSC Symposium on Intelligent Industrial Automation, 1999

[4] B.K.P. Horn, B.G. Schunck, Determining optical flow, Artificial Intelligence, 17 (1981) 185-203.

[5] A. Bainbridge-Smith, R.G. Lane, Determining optical flow using a differential method, Image and Vision Computing 15 (1997) 11-22.

[6] B.D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in Proc. DARPA Image Understanding Workshop, 1981, pp. 121-130.

[7] R.C. Jain, Difference and Accumulative Difference Pictures in Dynamic Scene Analysis, Image and Vision Computing (2), No. 2, May 1984, pp. 99-108.

[8] Y. Kameda, M. Minoh, A Human Motion Estimation Method using 3-successive video frames, Proceedings of International Conference on Virtual Systems and Multimedia'96, pp.135-140, 1996.

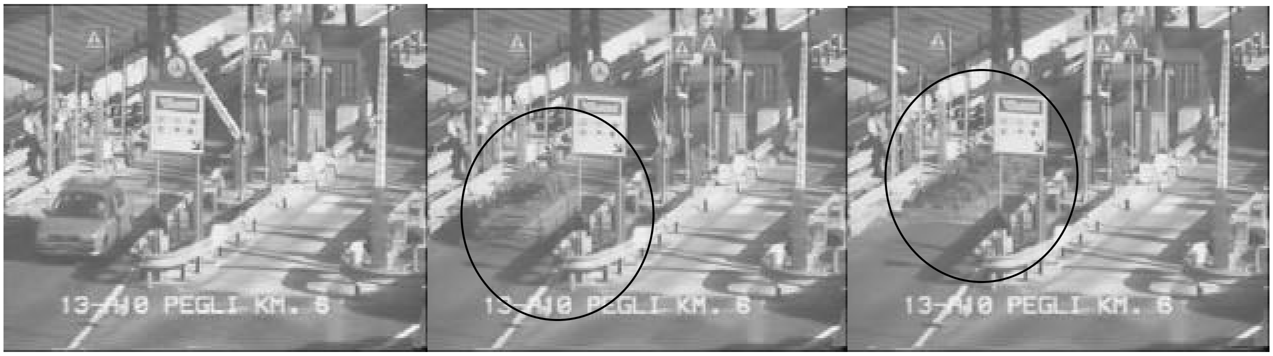
[9] C.R. Wren, A. Azarbayejani, T.J. Darrell, A.P. Pentland, Pfinder: Real-Time Tracking of the Human Body, IEEE Trans. Pattern Analysis and Machine Intelligence (19), No. 7, July 1997, pp. 780-785.

[10] K.P. Karmann, A. von Brandt, Moving Object Recognition Using an Adaptive Background Memory, in V Cappellini (ed.), Time-Varying Image Processing and Moving Object Recognition, 2, Elsevier, Amsterdam, The Netherlands, 1990.

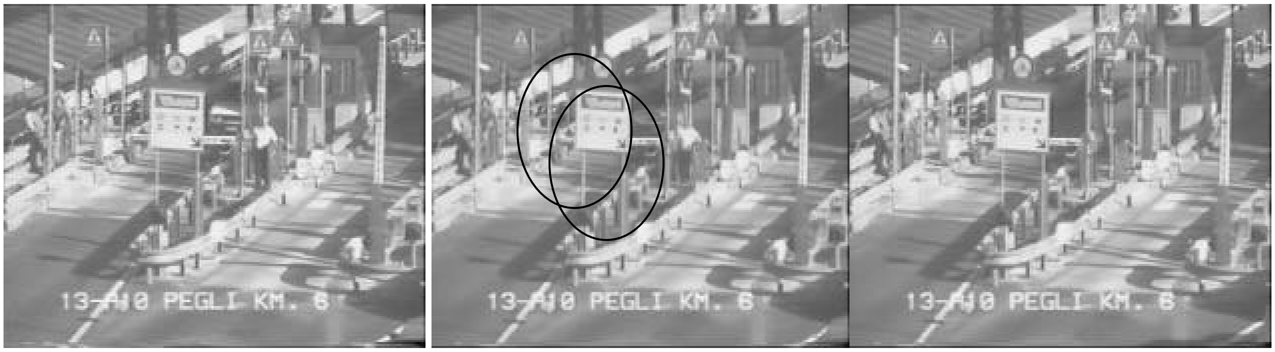
[11] A. Shio, J. Sklansky, Segmentation of People in Motion, Proceedings IEEE Workshop on Visual Motion, Princeton, NJ, October 7-10, 1991, 325-332.

[12] A. Elgammal, D. Harwood, L. Davis, Non-parametric Model for Background Subtraction, FRAME-RATE: Frame-rate Applications, Methods and Experiences with Regularly Available Technology and Equipment, Corfu, Greece, September 21, 1999.

[13] W.E.L. Grimson, C. Stauffer, Adaptive background mixture models for real-time tracking, Proceedings of in CVPR, 1999.

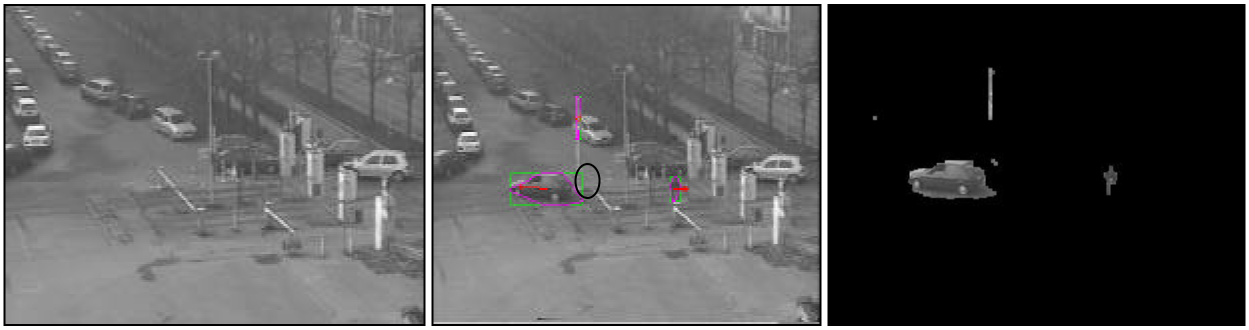


(a) Frame 35: a car passes through the barrier; wrong S-background with n=9 and DT=1 (b) and DT=5(c): DT must be increased



(d) Frame 290: a car and a people passing through, (e) wrong S-background with n=9 and DT=10 (f)correct S&KB background n=9 DT=1

Figure 4. Examples of background computation.



(a) (b) (c)

Figure 5. Detection of real moving vehicles by S&KB background technique.