Exemplar-based Background Model Initialization

Andrea Colombari
Dipartimento di Informatica
Università di Verona
Strada le Grazie 15
Verona, Italy
colombar@sci.univr.it

Marco Cristani
Dipartimento di Informatica
Università di Verona
Strada le Grazie 15
Verona, Italy
cristanm@sci.univr.it

Vittorio Murino
Dipartimento di Informatica
Università di Verona
Strada le Grazie 15
Verona, Italy
vittorio.murino@univr.it

Andrea Fusiello
Dipartimento di Informatica
Università di Verona
Strada le Grazie 15
Verona, Italy
andrea.fusiello@univr.it

ABSTRACT

Most of the automated video-surveillance applications are based on background (BG) subtraction techniques, that aim at distinguishing moving objects in a static scene. These strategies strongly depend on the BG model, that has to be initialized and updated. A good initialization is crucial for the successive processing. In this paper, we propose a novel method for BG initialization and recovery, that merges interesting ideas coming from the video inpainting and the generative modelling subfields. The method takes as input a video sequence, in which several objects move in front of a stationary BG. Then, a statistical representation of the BG is iteratively built, discarding automatically the moving objects. The method is based on the following hypotheses: (i) a portion of the BG, called *sure* BG, can be identified with high certainty by using only per-pixel reasoning and (ii) the remaining scene BG can be *generated* utilizing exemplars of the sure BG. The proposed algorithm is able to exploit these hypotheses in a principled and effective way.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Time-varying imagery*; I.5.1 [**Pattern Recognition**]: Models—*Statistical*; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*

General Terms

Algorithms, Security

VSSN'05, November 11, 2005, Singapore.

Keywords

Video surveillance, background initialization, background modelling, video inpainting, video analysis

1. INTRODUCTION

Analysis and understanding of video sequences is an active research field, whose importance is rapidly increased in the last years, due to the availability of more and more powerful hardware, to the development of effective real-time techniques, and to the potential vastity of the involved applications [1, 2]. Video surveillance is undoubtedly one of the most interesting applications of sequence analysis: human action recognition [3], semantic indexing of video [4], and, more generally, on-line discovering of unusual activities [5] are all tasks under investigations to partially or fully automate the surveillance. Typically, a video-surveillance system contemplates the monitoring of a site for long periods, using a static camera whose goal is to distinguish (and possibly classify) unusual behaviors from typical ones. To this end, the basic operation needed is the separation of the moving objects, the so-called *foreground* (FG), from the static information [6], the *background* (BG). This process is usually called background modelling.

The issues characterizing a BG modelling process are usually three: model representation, model initialization, and model adaptation. The first describes the kind of model (e.g., mixture of Gaussians) used to represent the BG; the second one regards the initialization of this model, and the third one relies to the mechanism used for adapting the model to the BG changes (e.g., illumination changes).

Recently, several techniques have been proposed in order to address the representation and the adaption issues, whereas the model initialization has received poor attention. In the BG model initialization problem, also called *bootstrapping* [7], the input is a short uncontrolled video sequence in which a number of moving objects may be present. The purpose is then to produce a BG model describing the observed scene. Actually, most of the BG models are built on a set of initial parameters that comes out from a short sequence, in which no FG objects are present [8]. This is

a too strong assumption, because in some situations it is difficult or impossible to control the area being monitored (e.g., public zones), which are characterized by a continuous presence of moving objects, or other disturbing effects.

In the literature, the initialization problem is typically disregarded, and only few methods are present. To the best of our knowledge, the main methods are devoted to algorithms initialization using a pixel-level analysis, disregarding higher-level information. Indeed, BG analysis could be carried out at different data-abstraction levels: pixel, region, and frame levels [7]. The pixel-level analysis processes independently each pixel, classifying it as FG or BG, and managing adaptation to changing BG [9]. In this modality, the analysis is performed at a very low level, and many problems of the BG subtraction remain unsolved, such as local or global sudden illumination changes [10]. The region-level analysis considers a higher level representation, modelling also inter-pixel relationships, so allowing a refinement of the modelling obtained at the pixel level. Finally, the frame-level analysis looks for changes in large parts of the image, and eventually swaps in more expressive BG models [11, 12].

In this paper, we propose a novel approach of BG model initialization that detects the visual appearance of the BG, inferring also a statistic description of each pixel of it by using a pixel-region scheme that draws on another active pattern recognition research field, the *video completion or video inpainting* subfield [15]. The techniques belonging to this research area try to fill-in user defined spatio-temporal holes in a video sequence using patches extracted in the existent spatio-temporal volume, according to consistency criteria evaluated both in time and space. The main application of these techniques is the one of removing unwanted objects in the scene, substituting them with visual patches in accord with the rest of the sequence. In the following, we consider a BG initialization process as an instance of video inpainting, in which we want to eliminate from the sequence all the FG objects, using the remaining visual information to estimate a statistics of the entire BG scene. Therefore, we want to define as hole every FG object present in the sequence, exploiting then spatio-temporal consistency to delete them coherently. The result is a representation of the static BG in which each pixel is modelled with a Gaussian density function that takes into account only for all the variations of the BG scene. These estimations can be used to correctly initialize BG subtraction methods based on probabilistic modelling of the pixel signals as the one proposed in [9]. The problem is that the FG and BG visual entities can not be so well discriminated due to several problems well known in the video surveillance literature, as the sleeping FG as first, the ghosting, the shadowing and so on (see [7]).

Our method tries to face to this uncertainty, using the following BG hypotheses: (i) there is a portion of the static BG represented with high probability by i.i.d. pixel processes that follow a Gaussian distribution: we call this zone as *sure BG zone*; (ii) the sure BG is *sufficiently explaining* the visual appearance of the entire BG. The pixel locations outside the sure BG zone form the *confusion zone* or region; by considering this region along time we obtain the *related confusion volume*, in which FG and BG pixel values are mixed together. The term *sufficiently explaining* has to be taken in a generative modelling framework, in the sense that the BG regions belonging to the confusion zone can be thought as produced by sampling square patches or exemplars of varying size from the sure BG region. In order to exploit this hypothesis, we process the confusion volume frame by frame, trying to discover local spatial patches maximally similar with valid exemplars of the BG. This approach is similar only in the spirit with the video epitomes proposed in [13]; in facts, our approach is based on a incremental process that define automatically the concept of FG and BG, while in [13] this distinction is not present.

This method represents a novelty in both the fields of the classical video inpainting literature and the video surveillance field. In the video inpainting literature because the method is able to automatically detect a zone to inpaint in order to find the scene of the video sequence, without the help of the user. In the video-surveillance field our method represents an improvement in the state of the art, because it solves a situation in which the BG is not evincible using only per pixel statistic, being visible partially and for few frames.

Lastly, our method can be applied in those applications in which the appearance of the BG is the main object of interest and the FG could be visual clutter to eliminate. For example, in Fig. 1 is depicted a poster behind moving foliage. In this case, the BG is intuitively the white table, which spatial particulars (the letters, with different dimensions) have to be extracted and composed together in order to perceive the whole image, not corrupted. The result is shown in (Fig. 2).

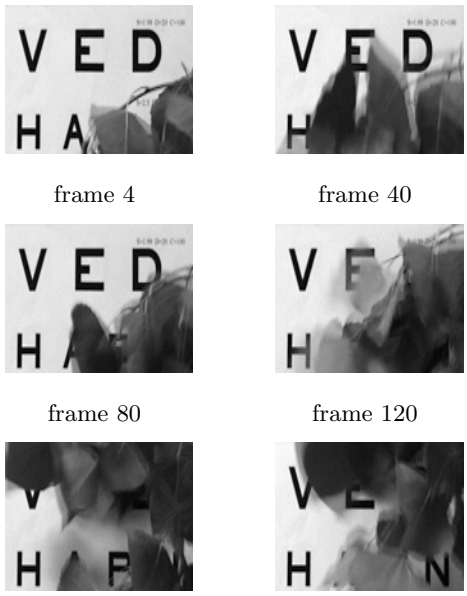


Figure 1: Few input frames from the "Sneller" sequence.

The rest of the paper is organized as follows. In Sec. 2, the state of the art of the BG initialization and the video inpainting subfields is examined. Sec. 3 is the core of this paper, in which the proposed approach is detailed. Experimental results are presented in Sec. 4 and, finally, Sec. 5 contains conclusions and future perspectives.

Figure 2: Output of the proposed approach: the static table with the letters as estimated from the proposed approach.

2. STATE OF THE ART

2.1 Background initialization

Actually, most of the BG models are built on a set of initial parameters that comes out from a short sequence, in which no FG objects are present [8]. This is a too strong assumption, because in some situations it is difficult or impossible to control the area being monitored (e.g., public zones), which are characterized by a continuous presence of moving objects, or other disturbing effects.

Most of the recent BG initialization methods discard the solution of computing a simple mean over all the frames, because it produces an image that exhibits blending pixel values in areas of FG presence. A general analysis regarding the blending rate and how it may be computed is present in [17]. In [18], the BG initial values are estimated by calculating the median value of all the pixels in the training sequence, assuming that the BG value in every pixel location is visible more than 50% of the time during the training sequence. Similarly, in [19] authors use median to model the BG of panning sequences where frames are aligned to form a mosaic of the BG. Even if those methods avoid the blending effects of the mean, the output of the median will contains large error when this assumption is false. Another proposed work [20], called adaptive smoothness method, avoids the problem of blending finding intervals of stable intensity in the sequence. Then, using some heuristics, the longest stable value for each pixel is selected and used as the value that most likely represents the BG. This method is similar to the recent Local Image Flow algorithm [16], which generates BG values hypotheses by locating intervals of relatively constant intensity, and weighting these hypotheses by using local motion information. Unlike most of the approaches, this method does not treat each pixel value sequence as an i.i.d. (independent identically distributed) process, but it considers also information generated by the neighboring locations. The problem of such approach is that, using a per pixel motion analysis, the problem of the sleeping FG is not faced, and whatever object visible for the most of the sequence is erroneously detected as BG, even if it has moved in few frames.

2.2 Video inpainting

Video-inpainting techniques can be evaluated as temporal development of the image inpainting techniques [14], that consist in substituting holes in an image choosing patches from the neighborhood and substituting them in the hole, in order to preserve spatial consistency, i.e. color smoothing and edge continuation.

Our approach belongs to those techniques that use small image patches or exemplar to account for high order statistic in image and video data. Similar to our approach is the video epitomes [13], in which the spatio-temporal volume is approximated using a smaller spatio-temporal volume, formed by local spatio-temporal cubes, that generated all the video data. The method is appealing, but the definition of BG and FG is not present. In [15] global consistency is captured by posing the problem of video inpainting as a global optimization problem with a well-defined objective function that assures spatio-temporal coherency, and solving it appropriately.

3. PROPOSED APPROACH

The proposed method takes as input a sequence of T frames, each one formed by N pixel values, ordered in raster scan. The output is a set of uni-modal Gaussian distributions, one for each pixel p_i, written as

$$\mathcal{N}(\mu_i, \sigma_i), \quad i = 1, \ldots, N \tag{1}$$

The algorithm consists in an iterative process that can be synthesized as follows:

1. Initialization
2. Patch-level BG expansion
3. Pixel-level BG expansion
4. Repeat steps 2 and 3 until the confusion zone is empty or it is not possible to proceed

In the following, we give an intuitive idea of the algorithm in order to fix the needed notation and to ease the reading of the next sections, where each single step is deeply described.

3.1 Overview of the method

3.1.1 Initialization

In this phase, we provide a partition of the scene in two regions: a region Φ, the sure BG zone, where pixel values satisfy the BG per-pixel hypothesis[1] and a region Ω, the confusion zone, where FG and the BG (pixel) values are mixed together in the sequence. Pixels in Φ are modelled by the set of Gaussian distributions $\{\mathcal{N}(\mu_i, \sigma_i)\}_{p_i \in \Phi}$. Moreover, for each $p_i \in \Phi$ we have the *membership index* G_i that indicates how many pixel values in the sequence are modelled by $\mathcal{N}(\mu_i, \sigma_i)$. In the following, we call the spatio-temporal volume spanned by all the pixels in Ω as the *volume relative to* Ω. These two regions are divided by a curve C belonging to Ω (see Fig. 3).

3.1.2 Patch-level BG expansion

In this step, we look in the confusion volume relative to Ω for a patch of BG. Starting from a pixel $p_i \in C$, we define a squared region $\Psi_{i,s}$ centered in p_i and with size $2s + 1$. Given a frame t, pixel values corresponding to the spatial locations in $\Psi_{i,s}$ form a patch $\Psi_{i,s}(t)$ (notice that the frame

[1]The first BG hypothesis mentioned in the Introduction and deeply explained in Sec. 3.2.

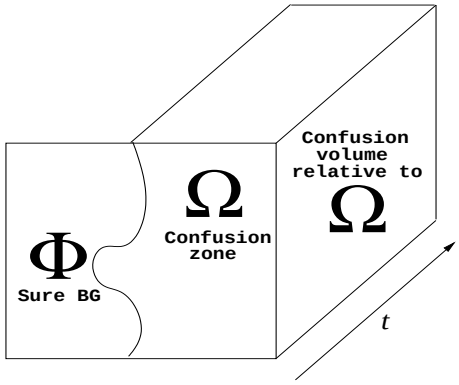


Figure 3: Definition of the sequence to analyze as spatio-temporal volume: note that only the confusion region determines a volume, in which the FG and BG pixels values are mixed together.

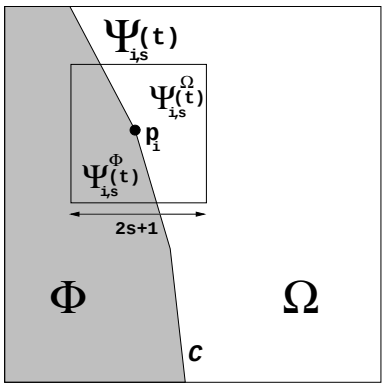


Figure 4: The patch-level BG expansion step: at iteration step m, we build a local patch $\Psi_{i,s}$, which value for each frame is indicated as $\Psi_{i,s}(t)$ and further divided in $\Psi^{\Omega}_{i,s}$ and $\Psi^{\Phi}_{i,s}$.

index t is used to specify a set of pixel values, otherwise we refer to an atemporal spatial location).

As depicted in Fig. 4 $\Psi_{i,s}(t)$ is partitioned in two parts: $\Psi^{\Omega}_{i,s}(t)$, located in the confusion zone, and $\Psi^{\Phi}_{i,s}(t)$, located in the sure BG zone. Pixels $p_i \in C$ are processed one by one following an order determined by a priority value. This ranking is based on the idea of processing first that pixel p_i for which the associated region $\Psi^{\Phi}_{i,s}$ has been built using a large amount of confident data. Then, we extract a patch $\Psi_{i,s}(t_{\text{Best}})$ (the "best hypothesis") at frame t_{Best}, formed by two fragments $\Psi^{\Phi}_{i,s}(t_{\text{Best}})$ and $\Psi^{\Omega}_{i,s}(t_{\text{Best}})$, such that $\Psi^{\Phi}_{i,s}(t_{\text{Best}})$ is similar to the existent BG model in that region, and the entire patch $\Psi_{i,s}(t_{\text{Best}})$ has to be maximally similar to a somewhat patch (the "source") in the sure BG model. The similarity criteria will be detailed in the following. In a generative sense, it is appropriate to consider the best hypothesis as *generated* from the source. If this hypothesis holds, is again licit that the best hypothesis inherits also the statistical properties of the source. Therefore, the best hypothesis assumes as *initial mean* the values $\Psi^{\Omega}_{i,s}(t_{\text{Best}})$, and as *initial standard deviation* the corresponding standard deviation values of the source. The obtained statistics is called *per-patch statistic* of $\Psi^{\Omega}_{i,s}(t_{\text{Best}})$.

3.1.3 Pixel-level BG expansion

In this step, we fit the initial per-patch statistics of $\Psi^{\Omega}_{i,s}(t_{\text{Best}})$ in the spatio-temporal volume relative to $\Psi^{\Omega}_{i,s}$ at pixel-level. In this way, we will obtain for each pixel location in $\Psi^{\Omega}_{i,s}$ a *per-pixel statistics*. This statistics models the likelihood of all the pixel values at a single pixel location $p_j \in \Psi^{\Omega}_{i,s}$ wrt to the per-patch statistics in the same location. The final *pixel-region statistics* of $\Psi^{\Omega}_{i,s}$ will be formed combining the two statistics, generating a set of Gaussian distributions $\mathcal{N}(\mu_j, \sigma_j)$, with $p_j \in \Psi^{\Omega}_{i,s}$. Finally, for each pixel p_j we obtain the number G_j of pixel values that are properly modelled by $\mathcal{N}(\mu_j, \sigma_j)$. After this step, the region Φ grows up and the spatial knowledge about the visual BG augments.

The last two steps are iterated until the confusion zone has been entirely processed, or when the similarity criteria has not being able to be satisfied for each pixel belonging to the contour $C \in \Phi$.

3.2 Step 1: Initialization

The initialization step is useful to build an initial partition of the scene in Φ and Ω. As stated in the Introduction, we define as BG pixels those pixels that shown a strictly uni-modal behavior. In order to find this characteristic, we tried two different approaches. First, we calculate for all the pixels signals $p_i(t), t = 1 \dots T$ of the sequence the median $\hat{\mu}_i$, and the MAD $\hat{\sigma}_i$ (a robust estimate of the standard deviation) in the sense that does not give undue weight to the tail behavior of the distribution. After this, we evaluate as BG pixel values all those median values which MAD estimates were lower than a threshold $\hat{\sigma}$, supposing this one as near as possible to the sensor noise standard deviation. This threshold was built evaluating a set of training sequences depicting only BG scenes. For each training sequence a particular standard deviation was found. At the end of the training session, the mean of all the σ's was evaluated, building the threshold $\hat{\sigma}^{Thresh}$. We built a set of thresholds for a set of different environment (indoor, outdoor with stable weather). Experimentally we found that this way was too strict, especially in the outdoor environment, being the BG subjected to other kind of noise, other than the sensor noise, due to weather visual fluctuations, shadows, etc.

Alternatively, we modelled each pixel signal as a mixture of Gaussians, which parameters are estimated using the Expectation Maximization (EM) algorithm [21], with number of component chosen automatically using the BIC criteria, derived from the most general MDL principle [22]. This parameter estimation process starts evaluating the likelihood of the observed pixel values given a starting maximally complex model (i.e. formed by a large number of components). Subsequently, the MDL principle is used in order to prune away the useless components, repeating the estimation process iteratively until convergence. At the end of the process, we obtain for each pixels sequence $p_i(t)$ a mixture of Gaussians (MoG) $\{\mathcal{N}\}_i$, composed by K_i Gaussian components $\mathcal{N}_{i,k}, k = 1, \dots, K_i$, with $\mu_{i,k}, \sigma_{i,k}, \pi_{i,k}$ indicating respectively the mean, the standard deviation and the mixing coefficient, that is proportional to the number of frames that the k-th component has been observed in the sequence.

Our task was to find for each pixel that Gaussian component that better depicts the BG, i.e. the component with

maximal mixing coefficient (that measures the persistency of a pixel value) *and* minimum standard deviation (the inverse of the precision). We combine these two components forming the *stability coefficient* $\rho_{i,k}$, for each pixel i and each component k, i.e.:

$$\rho_{i,k} = \pi_{i,k}\,(1 - \tilde{\sigma}_{i,k}) \tag{2}$$

where $\tilde{\sigma}_{i,k}$ is the standard deviation of the k-th Gaussian component, normalized by the sum of all the K standard deviations. This measure is similar to what used in [9], in order to choose those Gaussian components better modelling the BG of a battery of i.i.d pixel signals. Using this coefficient, we estimate the components that most probably model the BG appearance of the scene, by choosing $\mathcal{N}_{i,\hat{k}} = \underset{k}{\operatorname{argmax}}\ \rho_{i,k}$. Given this operation, the next step is to examine all the components selected by considering as BG those ones for which $\rho_{i,\hat{k}} > \tau$, where τ is a threshold fixed experimentally to 0.6. Lastly, we build for each p_i its *membership index* G_i, obtained by enumerating the pixel values for which hold

$$p_i(t) \in \mathcal{N}(\mu_j, \sigma_i) \quad t = 1, \ldots, T \tag{3}$$

that is satisfied when $p_i(t)$ falls in the range $[\mu_i - 3.5\sigma_i, \mu_i + 3.5\sigma_i]$. In the experiments, we show that this initialization method is able to model correctly portions of the scene depicting the static BG.

3.3 Step 2: Patch-level BG expansion

This step can be performed at each iteration m until no confusion zone is present in the scene. The estimated BG region and the confusion region at iteration m should correctly be called $\Phi^{(m)}$ and $\Omega(m)$ respectively, but we omit the (m) notation for clarity, if it is not explicitly needed. Together with the two regions, we obtain a corresponding contour that separates the zones $C \in \Omega$. The next step is the choice of the pixel $p_i \in C$, around that the squared region $\Psi_{i,s}$ is built (with starting *maximal* size $s = s_{\text{Init}}$), with which we explore the relative confusion volume.

The order with which the pixel p_i is chosen is determined by a priority queue. The basic idea of this ranking is that we want to process first that pixel p_i which associated region ${}^{\text{BG}}\Psi_{i,s}$ has been built using a large amount of data. In order to do this, we want to privilege the pixel p_i, which associated region $\Psi^{\Phi}_{i,s}$ is supported by the largest amount of pixel values with small variance. Therefore, we rank in descending order all the $p_i \in C$ by a *per-patch stability coefficient* $\hat{\delta}_{i,k}$:

$$\hat{\delta}_{i,k} = \sum_{j \in \Psi^{\Phi}_{i,s}} \delta_{j,k} \tag{4}$$

with

$$\delta_{j,k} = \frac{G_j}{T}\,(1 - \tilde{\sigma}_{j,k}) \tag{5}$$

where G_j in the number of pixel value used previously to the estimation of p_j. Once we selected the starting pixel p_i, we select the best frame t_{Best} for which

$$\Psi^{\Phi(m)}_{i,s}(t_{\text{Best}}) \in \Phi(m) \quad \wedge \tag{6}$$

$$t_{\text{Best}} = \underset{t,i'}{\operatorname{argmin}} \quad d(\Psi_{i,s}(t), \Psi_{i',s}) \tag{7}$$

where $\Psi_{i',s}$ is a squared region centered in i' and such that $(\Psi_{i',s} \cup \Omega(m) = \emptyset)$ and d is a whatever distance function. The condition of Eq. 6 is satisfied when each one of the pixel values $p_j(t) \in \Psi_{i,s}(t)$ falls in the range $[\mu_j - 3.5\sigma_j, \mu_j + 3.5\sigma_j]$, where μ_j and σ_j model the region $\Psi^{\Phi}_{i,s}$.

In other words, in this step we are looking for a frame t_{Best} in which the BG portion of the patch is conform with the corresponding sure BG model (Eq. 6), and the entire patch is maximally similar with a patch of the sure BG (Eq. 7); if both of the equations above hold, we found a spatial patch intersecting the confusion area where the component pixel values discarded as BG during the initialization, here are re-evaluated. This re-evaluation is due to their joint similarity with a patch of sure BG. At this point, the patch fragment $\Psi^{\Omega}_{i,s}(t_{\text{Best}})$ inherits from the BG region $\Psi_{i',s}$ the standard deviation value ${}^{\text{PP}}\sigma_j$ for each pixel $p_j \in \Psi^{\Omega}_{i,s}$, as depicted in Fig, maintaining as mean values ${}^{\text{PP}}\mu_j$ the pixel values $\Psi^{\Omega}_{i,s}(t_{\text{Best}})$ ("near" the values of the means that model $\Psi_{i',s}$). Moreover, for each pixel p_j we extract ${}^{\text{PP}}G_j$ as the number of pixels values with which the estimation of the corresponding value of the source p'_j has been built. This statistics is called *initial per patch statistics* relative to $\Psi^{\Omega}_{i,s}(t_{\text{Best}})$. Anyhow, this statistics will be updated in the next step, in order to improve the modelling of all the pixel values in the volume relative to $\Psi^{\Omega}_{i,s}(t_{\text{Best}})$. If one of the equations (6) and (7) does not hold for all the pixels $p_i \in C$, the process is repeated decreasing the size of the squared patch i.e. using $s' < s_{Init}$, until a minimum size s_{min} is reached. If no one match occurs even with the minimal patch size, this means that the BG present in the confusion zone is too much different wrt the sure BG, therefore no similarity criteria holds, and the process stops.

3.4 Pixel-level BG expansion

At this point, we have a patch $\Psi_{i,s}(t_{\text{Best}})$ most similar with respect to $\Psi^{\Phi}_{i,s}$ for what concerns the BG fragment *and* most similar with respect to a certain $\Psi_{i',s}$ (the "source"). But $\Psi_{i,s}(t_{\text{Best}})$ is only one spatial configuration of pixel values that, considered alone, can be present in more than one frame. Therefore, the aim of this step is to build a *per-pixel statistics* that takes into account for this consideration. In order to do this, we first find all the pixel values $\{p_j(t)\}$ in the volume relative to $\Psi^{\Omega}_{i,s}$ such that $p_j(t) \in \mathcal{N}({}^{\text{PP}}\mu_j, {}^{\text{PP}}\sigma_j)$ (as defined in Eq. 3) and their related membership number ${}^{\text{Pp}}G_j$. Once this membership function has evaluated, for each pixel we estimate the final per-pixel mean and standard deviation ${}^{\text{Pp}}\mu_j, {}^{\text{Pp}}\sigma_j$. This estimation represents a strictly local estimation of the single pixel values in accord with the per-patch statistic. As last step, we want to fuse the per-pixel statistic with the per-patch statistic. Therefore, we learn the final Gaussian parameters that, for each pixel p_j, are a weighted sum that takes into account for ${}^{\text{PP}}G_j$, i.e. the number of pixels with which p_j has been derived in the per-patch statistics, and the number of pixels ${}^{\text{Pp}}G_j$. In formulae:

$$\mu_j = \frac{{}^{\text{PP}}G_j\,{}^{\text{PP}}\mu_i + {}^{\text{Pp}}G_j\,{}^{\text{Pp}}\mu_j}{{}^{\text{PP}}G_j + {}^{\text{Pp}}G_j} \tag{8}$$

$$\sigma_j = \frac{{}^{\text{PP}}G_j\,{}^{\text{PP}}\sigma_j + {}^{\text{Pp}}G_j\,{}^{\text{Pp}}\sigma_j}{{}^{\text{PP}}G_j + {}^{\text{Pp}}G_j} \tag{9}$$

Lastly we calculate the number of pixels G_j with which the pixel-region statistic has been built as

$$G_j = \frac{{}^{\text{PP}}G_j + {}^{\text{Pp}}G_j}{2} \tag{10}$$

With these incremental learning we drift the Gaussian components of the new pixel of BG taking into account for the patch with which $\Psi_{i,s}$ is generated and the local statistic in which the patch $\Psi_{i,s}$ is immersed.

4. EXPERIMENTS

We show here four experiments performed on different kind of sequences in order to give an idea of both the potentialities and the limits of the proposed method. The algorithm is written in MATLAB code and has been ran on a Pentium IV 3GHz CPU with 1Gb of RAM. The mean elapsed time for a computation is around 1 hour and half.

The first sequence, the "Sneller" sequence, depicts a white poster with letters of different dimensions. In front of this poster there is moving foliage that occludes a wide region (Fig. 1). As explained in Sec. 3.2, we first apply the initialization step that finds the sure BG region. In order to validate this step, we apply also the median based initialization. In Fig. 5 (a) is depicted the median of the sequence, over which we have to choose the sure BG region. In Fig. 5 (b), is depicted the mean of the component with highest stability coefficient, found by the Gaussian clustering. This comparison is enough to state that the second method outgoes the first one, by (correctly) depicting on the left side white pixels values, while in the median the corresponding values are gray because influenced by the leaves gray level. The selection of the sure BG performed by our method is depicted in Fig. 5 (c). Once the initialization is performed, the inpainting process can be ran. In Fig. 6 a sketch of the process is shown. In the first column, the reader can see the inpainting process evolution, in which the sure BG is enlarged at each step using the squared patch highlighted by a red solid line. This patch results maximally similar to the one highlighted by a dashed blue line. It is interesting to note that the patch size decreases during the process. In fact, as reader can notice, from iteration $m = 152$ the patch size is smaller (25×25) than the previous size (31×31). This is because with the initial size no similarity is discovered along all the contour that divides the sure BG zone and the confusion zone. In the central column standard deviation values are shown. The column on the right shows the number of pixels that are involved in the statistics that models the sure BG identified up to now: the brighter the color is, the more the pixel values involved are. The final results, obtained after 319 iterations, are depicted in Fig. 2 (mean pixel values) and in Fig. 7 (standard deviation values).

(a) (b) (c)

Figure 5: Initialization method: (a) median based; (b) mixture based; (c) sure BG zone detected from the initialization step, opposed to the black-masked confusion zone.

The second sequence, the "Classroom" sequence, depicts two persons that are walking and staying in front of a written board (Fig. 8). In this case, the details on the board imply the use of smaller patches (13×13 in average). The satisfying results are displayed in Fig. 9.

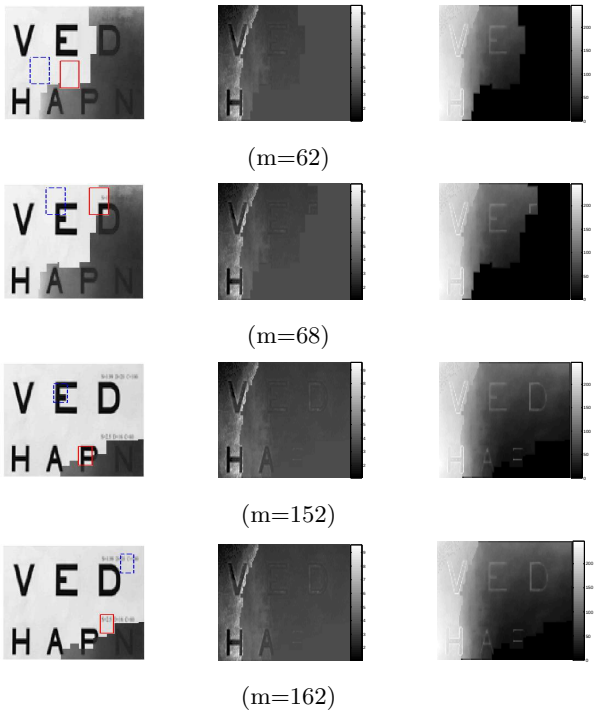


Figure 6: Sketch of the inpainting process.

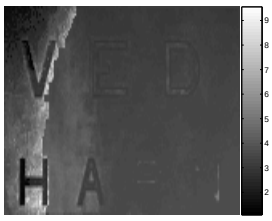


Figure 7: Final standard deviation values detected for the BG of the "Sneller" sequence.

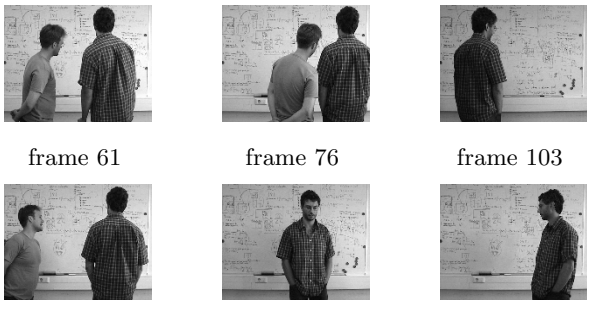
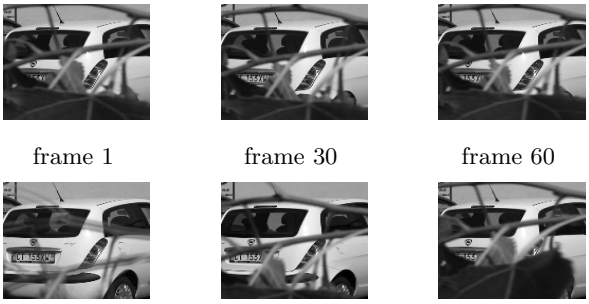


Figure 8: Few frames of the "Classroom" sequence.

(a) (b)

Figure 9: "Classroom" sequence results: BG (a) mean values and (b) standard deviation values.

The third sequence, the "Car" sequence (Fig. 10), is characterized by a highly irregular BG and moving foliage. In this case, the algorithm is very slow because of difficulties to find similarity between the sure BG and the BG in the confused zone. And in fact, it stops at iteration $m = 1627$, because the similarity criteria does not hold for any pixel of the contour C, and for any size of the patch. The problem here is that the sure BG is not able to explain the BG in the remaining confused zone and consequently the second BG hypothesis mentioned in the Introduction falls and the method stops. Anyhow, the result obtained (see Fig. 11, (a) and (b)) shows that the partial BG estimation models the car behind the foliage in a good way. Fig. 12 shows which part of the initial confused zone has been processed and which part has not.



frame 1 frame 30 frame 60

frame 90 frame 120 frame 130

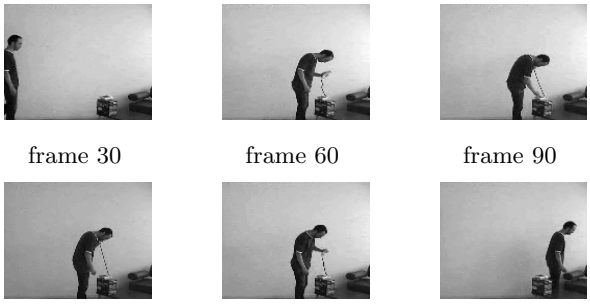
Figure 10: Few frames of the "Car" sequence.

(a) (b)

Figure 11: "Car" sequence results: BG (a) mean values and (b) standard deviation values.

Figure 12: Too irregular BG ("Car" sequence): the not processed confused zone is shown in white whilst the shadowed area represents the processed confused zone and the rest of the image represents the initial sure BG.

The last sequence, the "Phone" sequence, is an example of when the proposed method terminates, but the result is not very good. The sequence depicts an almost uniform BG with a moving person that stays almost still for a long time in the center of the scene for a phone call (See Fig. 13). As we see from the results (Fig. 14), the moving person has been removed, but the visual appearance of the BG is not smooth as expected since the person silhouette can be recognized in the phone-call zone. This is because the mean values computed in that zone differ from the rest of the BG (and the low variance witnesses this). In fact, the number of pixel values involved in the person substitution are inferior wrt the rest of the scene, in which the BG has been built using more evidence, taking into account for more BG fluctuations.



frame 30 frame 60 frame 90

frame 120 frame 150 frame 180

Figure 13: Few frames of the "Phone" sequence.

5. CONCLUSIONS

In this paper we proposed a novel method for BG initialization and recovery. The method takes as input a video sequence in which several moving objects can be present, possibly for long time, and builds a per-pixel statistical representation of the static BG. The iterative method here explained represents an improvement in the video surveillance literature, because is able to give a statistics of the static BG appearance even for that zones in which the scene is visible for few frames. This estimation is built by exploit-

(a) (b)

Figure 14: "Phone" sequence results: BG (a) mean values and (b) the standard deviation values.

ing some hypotheses: a spatial portion of the BG has to be visible with high per-pixel certainty (forming the sure BG), and this portion visually should explain in a generative way the visual appearance of the rest of the scene (the confusion zone). These two hypotheses are combined and used by a method similar to a video inpainting technique. The basic idea is that those visual patches in the confusion zone highly similar to some visual exemplars of the sure BG are themselves evaluated as sure BG. The approach is valid, in the sense that finds good estimations of the BG visual appearance, even if some heuristic is present. A future perspective is to find a more elegant and formal way to describe the novel proposed approach, for example finding a generative model similar to the one proposed in [13] in which is embedded the definition of FG and BG.

6. REFERENCES

[1] Y. Deng and B. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(8):800–810, 2001.

[2] H. Tao, H. Sawhney, and R. Kumar. Object tracking with bayesian estimation of dynamic layer representations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(1):75–89, 2002.

[3] Y. Yacoob and M. Black. Parametrized modeling and recognition of activities. *Computer Vision and Image Understanding: CVIU*, 73(2):232–247, 1999.

[4] M. Petkovic and W. Jonker. Content-based video retrieval by integrating spatio-temporal and stochastic recognition of events. In *Proc. of IEEE Workshop on Detection and Recognition of Events in Video*, pages 75–82, 2001.

[5] I. Haritaoglu, D. Harwood, and L. Davis. W^4: real-time surveillance of people and their activities. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):809–830, 2000.

[6] N. Friedman and S. Russell. Image segmentation in video sequences: A probabilistic approach. In *Uncertainty in Artificial Intelligence: Proceedings of the Thirteenth Conference (UAI-1997)*, pages 175–181, San Francisco, CA, 1997. Morgan Kaufmann Publishers.

[7] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *Int. Conf. Computer Vision*, pages 255–261, 1999.

[8] B. Gloyer, H. K. Aghajan, K. Y. Siu, and T. Kailath. Video-based freeway monitoring system using recursive vehicle tracking. In *IS&T-SPIE Symposium on Electronic Imaging: Image and Video Processing*, 1995.

[9] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Int. Conf. Computer Vision and Pattern Recognition*, volume 2, 1999.

[10] M. Cristani, M. Bicego, and V. Murino. Integrated region- and pixel-based approach to background modelling. In *Proc. of IEEE Workshop on Motion and Video Computing*, pages 3–8, 2002.

[11] B. Stenger, V. Ramesh, N. Paragios, F.Coetzee, and J. M. Buhmann. Topology free hidden Markov models: Application to background modeling. In *Int. Conf. Computer Vision*, volume 1, pages 294–301, 2001.

[12] N. Ohta. A statistical approach to background subtraction for surveillance systems. In *Int. Conf. Computer Vision*, volume 2, pages 481–486, 2001.

[13] V. Cheung, B.J. Frey and N. Jojic. Video Epitomes. In *Int. Conf. Computer Vision and Pattern Recognition*, volume 1, pages 42–49, 2005.

[14] A. Criminisi, P. Prez and K. Toyama. Region Filling and Object Removal by Exemplar-Based Inpainting. *IEEE Trans. on Image Processing*, 13(9):1200–1212, 2004.

[15] Y. Wexler, E. Shechtman and M. Irani. Space-Time Video Completion. In *Int. Conf. Computer Vision and Pattern Recognition*, volume 1, pages 120–127, 2004.

[16] D. Gutchess, M. Trajkovic, E. Cohen-Solal, D. Lyons and A.K. Jain. A Background Model Initialization Algorithm for Video Surveillance. In *Int. Conf. Computer Vision*, pages 733-740, 2001.

[17] X. Gao, T. Boult, F. Coetzee, and V. Ramesh. Error analysis of background adaption. In *Proc. of IEEE Conf. on Computer Vision Pattern Recognition*, volume I, pages 503–510, 2000.

[18] B. Gloyer, H. Aghajan, K.-Y. Siu, and T. Kailath. Video-based freeway monitoring system using recursive vehicle tracking. In C. M. Bishop and B. J. Frey, editors, *Proc. SPIE - The International Society for Optical Engineering*, volume 2421, pages 747–757, 1995.

[19] A. Colombari, A. Fusiello, and V. Murino. Segmentation and tracking of multiple objects in video sequences. In *Proc. of the 5th International Workshop on Image Analysis for Multimedia Interactive Services*, 2004.

[20] W. Long and Y. Yang. Stationary background generation: An alternative to the difference of two images. *Pattern Recognition*, 23:1351–1359, 1990.

[21] A.P. Dempster, N.M. Laird and D.B. Rubin. Maximum Likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39:1–38, 1977.

[22] M.A.T. Figueiredo and A.K. Jain. Unsupervised learning of finite mixture models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.