Author manuscript, published in IEEE Trans. on Image Processing, vol. 17, no. 4, pp. 608-621, 2008

Bayesian Foreground and Shadow Detection in Uncertain Frame Rate Surveillance Videos

Csaba Benedek, *Student member, IEEE* and Tamás Szirányi, *Senior member, IEEE*

***Abstract*— In in this paper we propose a new model regarding foreground and shadow detection in video sequences. The model works without detailed a-priori object-shape information, and it is also appropriate for low and unstable frame rate video sources. Contribution is presented in three key issues: (1) we propose a novel adaptive shadow model, and show the improvements versus previous approaches in scenes with difficult lighting and coloring effects. (2) We give a novel description for the foreground based on spatial statistics of the neighboring pixel values, which enhances the detection of background or shadow-colored object parts. (3) We show how microstructure analysis can be used in the proposed framework as additional feature components improving the results. Finally, a Markov Random Field model is used to enhance the accuracy of the separation. We validate our method on outdoor and indoor sequences including real surveillance videos and well-known benchmark test sets.**

***Index Terms*— Foreground, Shadow, Texture, MRF.**

I. Introduction

FOREGROUND detection is an important early vision task in visual surveillance systems. Shape, size, number and position parameters of the foreground objects can be derived from an accurate silhouette mask and used by many applications, like people or vehicle detection, tracking and event classification.

The presence of moving cast shadows on the background makes it difficult to estimate shape [1] or behavior [2] of moving objects. Since under some illumination conditions $40-50\%$ of the non-background points may belong to shadows, methods without shadow filtering [3][4][5] can be less efficient in scene analysis.

In the paper we deal with an image segmentation problem with three classes: *foreground* objects, *background* and *shadow* of the foreground objects being cast on the background. We exploit information from local pixel-levels, microstructural features and neighborhood connection. We assume having a stable, or stabilized [6] static camera, since it is available for several applications. Note that there are papers [3][7][8] focusing on the presence of dynamic background and camera ego-motion instead of the various shadow effects.

Another important issue is related to the properties of the video flow. For several video surveillance applications high-resolution images are crucial. Due to the high bandwidth requirement, the sequences are often captured at a low [9] or unsteady frame rate depending on the transmission conditions. These problems appear, especially, if the system is connected to the video sources through narrow band radio channels or over saturated networks. For another example, quick off-line evaluation of the surveillance videos is necessary after a criminal incident. Since all the video streams corresponding to a given zone should be continuously recorded, these videos may have a frame rate lower than 1 fps to save up storage resources.

For these reasons, a large variety of temporal information, like pixel state transition probabilities [10][11][12], periodicity calculus [2][13], temporal foreground description [3], or tracking [14][15], are often hard to derive, since they usually need a permanently high frame rate. Thus, we focus on using frame rate independent features to ensure graceful degradation if the frame rate is low or unbalanced. On the other hand, our model also exploits temporal information for background and shadow modeling.

A technique used widely for background subtraction is the adaptive Gaussian mixtures method of [4], which can be used together with shadow filters of e.g. [16][17][18]. These methods classify each pixel independently, and morphology is used later to create homogenous regions in the segmented image. That way, the shape of the silhouettes may be strongly corrupted as it is shown in [12][19].

An alternative segmentation schema is a Bayesian approach [12]. The background, shadow and foreground classes are considered to be stochastic processes which generate the observed pixel values according to locally specified distributions. The spatial interaction constraint of the neighboring pixels can be modelled by Markov Random Fields (MRF) [20].

Some previous Bayesian methods [21][22] detect foreground objects by building adaptive models regarding the background and shadow, and the foreground pixels are purely recognized as non-matching points to these models. That way, background or shadow colored object-parts cannot be recognized. Spatial object description has been used both for interactive [23] and unsupervised image segmentation [24]. However, in the latter case, only large objects with typical color or texture are detected, since the model [24] penalizes the small segmentation classes. The authors in [3] have characterized the foreground by assuming temporal persistence of the color and smooth changes in the place of the objects. Nevertheless, in case of low frame rate, fast motion and overlaying objects, appropriate temporal information is often not available.

Manuscript received July 18, 2006; revised May 11, 2007. This work was partially supported by the EU project MUSCLE (FP6-567752). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Anil Kokaram.

The authors are with the Distributed Events Analysis Research Group, Computer and Automation Research Institute, Hungarian Academy of Sciences, H-1111 Budapest, Kende u. 13-17, and with the Faculty of Information Technology, Pázmány Péter Catholic University, H-1083 Budapest, Práter utca 50/A, Hungary (e-mail: bcsaba@sztaki.hu, sziranyi@sztaki.hu)

Digital Object Identifier: 10.1109/TIP.2008.916989

TABLE I

Comparison of different corresponding methods and the proposed model. Notes: * temporal foreground description, ** pixel state transitions

Method	High frame rate requirement	Shadow detection	Shadow parameter update	Foreground estimation from current frame	indoor / outdoor	texture	Dynamic background
Mikic 2000 [21]	No	global, constant ratio	No	No	outdoor	No	No
Paragious 2001 [28]	No	illumination invariant	No	No	indoor	No	No
Salvador 2004 [29]	No	illumination invariant	No	No	both	No	No
Martel-Brisson 2005 [31]	No	local process	Yes	No	indoor	No	No
Sheikh 2005 [3]	Yes: tfd *	No	-	No	both	No	Yes
Wang 2006 [12]	Yes: pst **	global, constant ratio	No	No	indoor	first ordered edges	No
Proposed method	No	global, probabilistic	Yes	Yes	both	different microstructures	No

Our method (partly introduced in [25]) is a Bayesian technique which uses spatial color information instead of temporal statistics to describe the foreground. It assumes that foreground objects consist of spatially connected parts and these parts can be characterized by typical color distributions. Since these distributions can be multi-modal, the object-parts should not be homogenous in color or texture, while we exploit the spatial information without segmenting the foreground components.

In the literature, different approaches are available regarding shadow detection. Although there are some methods [26][27] which attempt to find and remove shadows in the single frames independently, their performance may be degraded [26] in video surveillance, where we must expect images with poor quality and low resolution, while the computational complexity is too high for practical use [27].

For the above reasons, we focus on video-based shadow modeling techniques in the following. Here the 'shadow invariant' methods convert the images into an illumination invariant feature space: they remove shadows instead of detecting them. This task is often performed by color space transformation. Widely used illumination-invariant color spaces are e.g. the normalized rgb [16][28] and $c_1c_2c_3$ spaces [29]. [30] exploits hue constancy under illumination changes to train a weak classifier as a key step of a more sophisticated shadow detector. We find an overview of the illumination invariant approaches in [29] indicating that several assumptions are needed regarding the reflecting surfaces and the light sources. These assumptions are usually not fulfilled in a real-world environment. Outdoors, for example, the illumination is the composition of the direct sunlight, the diffused light corresponding to the blue sky, and various additional light components reflected from the field objects with significantly different spectral distributions. Moreover, the camera sensors may be saturated, especially in the case of dark shadows, therefore the measured colors cannot be calculated by simplified physical models. Since some of these color spaces ignore the luminance components of the color, the resulting models become sensitive to noise.

In a 'local' shadow model [31] independent shadow processes are proposed for each pixel. The local shadow parameters are trained using a second mixture model similarly to the background in [4]. In this way, the differences in the light absorption-reflection properties of the scene points can be notably considered. However, a single pixel should be shadowed several times till its estimated parameters converge, whilst the illumination conditions should stay unchanged. This hypothesis is often not satisfied in outdoor surveillance environments, therefore, this local process based approach is less effective in our case.

We follow another approach: shadow is characterized with 'global' parameters in an image (or in each subregion, in case of videos having separated scene areas with different lightings), and the model describes how the background values of the different sites change, when shadow is projected on them. We consider the transformation between the shadowed and background values of the pixels as a random transformation, hence, we take several illumination artifacts into consideration. On the other hand, we derive the shadow parameters from global image statistics, therefore, the model performance is reasonable also on the pixel positions where motion is rare.

Color space choice is a key issue in several corresponding methods. We have chosen the CIE L*u*v* space for two well known properties: we can measure the perceptual distance between colors with the Euclidean distance [32], and the color components are approximately uncorrelated with respect to camera noise and changes in illumination [33]. Since we derive the model parameters in a statistical way, there is no need for accurate color calibration and we use the common CIE D65 standard. It is not critical to consider the exact physical meaning of the color components, which is usually environment-dependent [29]; we use only an approximate interpretation of the L, u, v components and show the validity of the model via experiments.

Besides the color values, we exploit microstructure information to enhance the accuracy of the segmentation. In some previous works [7][8] texture was used as the only feature for background subtraction. That choice can be justified in case of strongly dynamic background (like a surging lake), but it gives lower performance than pixel value comparison in a stable environment. We find a solution for integrating intensity and texture differences for frame differencing in [34]. However, that is a slightly different task than foreground detection, since we should compare the image regions to background/shadow models. Respect to the background class, our color-texture fusion process is similar to the joint segmentation approach of [12], which integrates gray level and local gradient features. We extend it by using different and adaptively chosen microstructural kernels, which suit better the local scene properties. Moreover, we show how this probabilistic approach can be used to improve our shadow model.

For validation we use real surveillance video shots and also test sequences from a well-known benchmark set [35]. Table I summarizes the different goals and tools regarding some of the above mentioned state-of-the-art methods and the proposed model. For detailed comparison see also Section VII.

In summary, the main *contributions* of this paper can be divided into three groups. We introduce a *statistical shadow model* which is robust regarding the forthcoming artifacts in real-world surveillance scenes (Section III-B.), and a corresponding automatic parameter update procedure, which is usually missing from previous similar methods (Section V-B). We introduce a non-object based, spatial description of the *foreground* which enhances the segmentation results also in low frame rate videos (Section IV). Meanwhile, we show how *microstructure analysis* can improve the segmentation in this framework (Section III-C).

We also have a few assumptions in the paper. First, the camera stands in place and it has no significant ego-motion. Secondly, we expect static background objects (e.g. there is no waving river in the background). The third assumption is related to the illumination: we deal with one emissive light source in the scene, however, we consider the presence of additional diffused and reflected light components.

II. Formal model description

An image S is considered to be a two-dimensional grid of pixels (sites), with a neighborhood system on the lattice. The procedure assigns a label ω_s to each pixel $s \in S$ form the label-set: $\Phi = \{\text{fg,bg,sh}\}$ corresponding to three possible classes: foreground (fg), background (bg) and shadow (sh). Therefore, the segmentation is equivalent to a global labeling $\Omega = \{\omega_s \mid s \in S\}$. As it is typical, the label field Ω is modelled as a Markov Random Field based on [20].

The image data at pixel s is characterized by a 4 dimensional feature vector:

$$\overline{x}_s = [x_L(s), x_u(s), x_v(s), x_T(s)]^T \tag{1}$$

where the first three elements are the color components of the pixel in the CIE L*u*v* space, and $x_T(s)$ is a microstructural response which we introduce in Section III-C in detail. Set $X = \{\overline{x}_s | s \in S\}$ marks the global image data.

We use a Maximum A Posteriori (MAP) estimator for the label field, where the optimal labeling $\widehat{\Omega}$, corresponding to the optimal segmentation, maximizes the probability:

$$P(\widehat{\Omega}|X) \propto P(X|\widehat{\Omega}) \cdot P(\widehat{\Omega}) \tag{2}$$

We assume that the observed image data in the different pixel positions is conditionally independent given a labeling Ω [36]: $P(X|\Omega) = \prod_{s\in S} P(\overline{x}_s|\omega_s)$, while to present smooth connected regions in the segmented image, the a-priori probability of a labeling, $P(\Omega)$, is defined by the Potts model [37]. The key point in the model is to define the conditional density functions $p_k(s) = P(\overline{x}_s|\omega_s = k)$, for all $k \in \Phi$ and $s \in S$. For example, $p_{\text{bg}}(s)$ is the probability that the background process generates the observed feature value $\overline{x}_s$ at pixel s. Later on $\overline{x}_s$ in the background will also be featured as a random variable with the probability density function $p_{\text{bg}}(s)$.

We define the conditional density functions in Section III-V, and the segmentation procedure will be presented in Section VII in detail. Before continuing, note that in fact we minimize the minus-log of eq. (2). Therefore, in the following we use the $\epsilon_k(s) = -\log p_k(s)$ local energy terms, for easier notation.

III. Probabilistic model of the background and shadow processes

A. General model

We model the distribution of feature values in the background and in the shadow by Gaussian density functions, like e.g. [11][12][35].

Considering the low correlation between the color components [33], we approximate the joint distribution of the features by a 4 dimensional Gaussian density function with diagonal covariance matrix:

$$\overline{\overline{\Sigma}}_k(s) = \text{diag}\{\sigma^2_{k,L}(s), \sigma^2_{k,u}(s), \sigma^2_{k,v}(s), \sigma^2_{k,T}(s)\}$$

for $k \in \{\text{bg}, \text{sh}\}$.

Accordingly, the distribution parameters are $\overline{\mu}_k(s) = [\mu_{k,L}(s), \ldots, \mu_{k,T}(s)]^T$ mean, and $\overline{\sigma}_k(s) = [\sigma_{k,L}(s), \ldots, \sigma_{k,T}(s)]^T$ standard deviation vectors. With this 'diagonal' model we avoid matrix inversion and determinant recovery during the calculation of the probabilities, and the $\epsilon_k(s) = -\log p_k(s)$ terms can be directly derived from the one dimensional marginal probabilities:

$$\epsilon_k(s) = C + \sum_{i=\{L,u,v,T\}} \log \sigma_{k,i}(s) + \frac{1}{2}\left(\frac{x_i(s) - \mu_{k,i}(s)}{\sigma_{k,i}(s)}\right)^2 \tag{3}$$

with $C = 2\log 2\pi$. According to eq. (3), each feature contributes with its own additional term to the energy calculus. Therefore, the model is modular: the one dimensional model parameters, $[\mu_{k,i}(s), \sigma^2_{k,i}(s)]$, can be estimated separately.

B. Color features

The use of a Gaussian distribution to model the observed color of a single background pixel is well established in the literature, with the corresponding parameter estimation procedures such as in [4][38]. We train the color components of the background parameters $[\overline{\mu}_{\text{bg}}(s), \overline{\sigma}_{\text{bg}}(s)]$ in a similar manner to the conventional online K-means algorithm [4]. $[\mu_{\text{bg},L}(s), \mu_{\text{bg},u}(s), \mu_{\text{bg},v}(s)]^T$ vector estimates the mean background color of pixel s measured over the recent frames, while $\overline{\sigma}_{\text{bg}}(s)$ is an adaptive noise parameter. An efficient outlier filtering technique [4] excludes most of the non-background pixel values from the parameter estimation

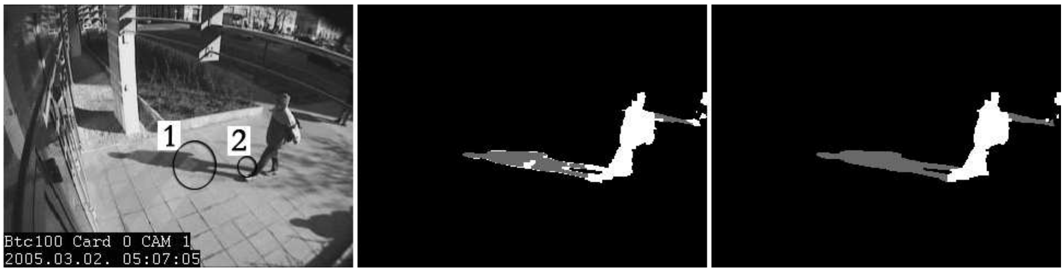


Fig. 1. Illustration of two illumination artifacts (the frame in the *left* image has been chosen from the ‘Entrance pm’ test sequence). 1: light band caused by a non-Lambertian reflecting surface (a glass door) 2: dark shadow part between the legs (more object parts change the reflected light). The constant ratio model (see image in the *middle*) causes errors, while the proposed model (*right* image) is more robust.

process, which works without user interaction.

As we have stated in the introduction, we characterize shadows by describing the background-shadow color value transformation in the images. The shadow calculus is based on the illumination-reflection model [39], which has been originally introduced for constant lighting, flat and Lambertian reflecting surfaces. Usually, our scene does not fulfill these requirements. The presented novelty is that we use a probabilistic approach to describe the deviation of the scene from the ideal surface assumptions, and get a more robust shadow detection.

1) Measurement of color in the Lambertian model: According to the illumination model [39] the response $g(s)$ of a given image sensor placed at pixel s can be written as

$$g(s) = \int e(\lambda, s)\rho(\lambda, s)\nu(\lambda)d\lambda \tag{4}$$

where $e(\lambda, s)$ is the illumination function, $\rho(s)$ depends on the surface albedo and geometry, $\nu(\lambda)$ is the sensor sensitivity. In the ‘background’, the illumination function is the composition of a direct and some diffused-reflected light components, while a shadowed surface point is illuminated by the diffused-reflected light only.

With further simplifications [39], eq. (4) implies the well-known ‘constant ratio’ rule. Namely, the ratio of the shadowed $g_{\mathrm{sh}}(s)$ and illuminated value $g_{\mathrm{bg}}(s)$ of a given surface point is considered to be constant over the image: $\frac{g_{\mathrm{sh}}(s)}{g_{\mathrm{bg}}(s)} = A$.

The ‘constant ratio’ rule has been used in several applications [11][12][21]. Here the shadow and background Gaussian terms corresponding to the same pixel are related via a globally constant linear density transform. In this way, the results may be reasonable when all the direct, diffused and reflected light can be considered constant over the scene. However, the reflected light may vary over the image in case of several static or moving objects, and the reflecting properties of the surfaces may differ significantly from the Lambertian model (See Fig. 1).

The efficiency of the constant ratio model is also restricted by several practical reasons, like quantification errors of the sensor values, saturation of the sensors, imprecise estimation of $g_{\mathrm{bg}}(s)$ and A, or video compression artifacts. Based on our experiments (Section VII), these inaccuracies cause poor detection rates in some outdoor scenes.

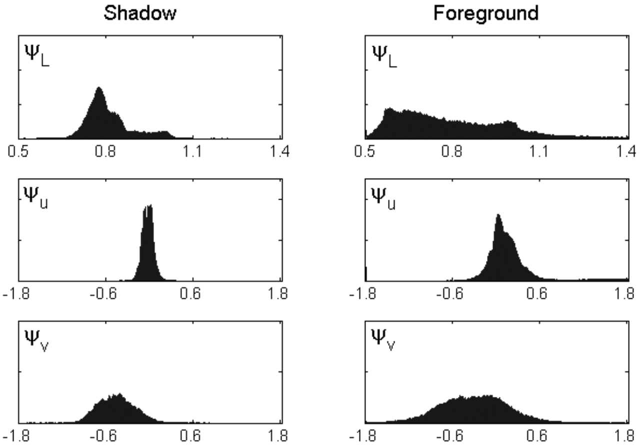


Fig. 2. Histograms of the ψ_L, ψ_u and ψ_v values for shadowed and foreground points collected over a 100-frame period of the video sequence ‘Entrance pm’ (frame rate: 1 fps). Each row corresponds to a color component.

2) Proposed model: The previous section suggests that the ratio of the shadowed and background luminance values of the pixels may be useful, but not powerful enough as a descriptor of the shadow process. Instead of constructing a more difficult illumination model, for example in 3D with two cameras, we overcome the problems with a statistical model. For each pixel s, we introduce the variable $\psi_L(s)$ by:

$$\psi_L(s) = \frac{x_L(s)}{\mu_{\mathrm{bg,L}}(s)} \tag{5}$$

where, as defined earlier, $x_L(s)$ is the observed luminance value at s, and $\mu_{\mathrm{bg},L}(s)$ is the mean value of the local Gaussian background term estimated over the previous frames [4].

Thus, if the $\psi_L(s)$ value is close to the estimated shadow darkening factor, s is more likely to be a shadowed point. More precisely, in a given video sequence, we can estimate the distribution of the shadowed ψ_L values globally in the video parts. Based on experiments with manually generated shadow masks, a Gaussian approximation seems to be reasonable regarding the distribution of shadowed ψ_L values (Fig. 2 shows the global ψ statistics regarding a 100-frame period of outdoor test sequence ‘Entrance pm’). For comparison, we have also plotted the statistics for the foreground points, which follows a significantly different, more uniform distribution.

Due to the spectral differences between the direct and ambient illumination, cast shadows may also change the u and v color components [40]. We have found an offset between the shadowed and background u values of the pixels, which can be efficiently modelled by a global Gaussian term in a given scene (similarly as for the v component). Hence, we define $\psi_u(s)$ (and $\psi_v(s)$) by

$$\psi_u(s) = x_u(s) - \mu_{\mathrm{bg,u}}(s) \tag{6}$$

As Fig. 2 shows, the shadowed $\psi_u(s)$ and $\psi_v(s)$ values follow approximately normal distributions.

Consequently, the shadow color process is characterized by a three dimensional Gaussian random variable:

$$\forall s \in S : \overline{\psi}(s) = [\psi_L(s), \psi_u(s), \psi_v(s)]^T \leftarrow N[\overline{\mu}_\psi, \overline{\sigma}_\psi]$$

According to eq. 5 and 6, the color values in the shadow at each pixel position are also generated by Gaussian distributions,

$$[x_L(s), x_u(s), x_v(s)]^T \;\leftarrow\; N[\overline{\mu}_{\mathrm{sh}}(s), \overline{\sigma}_{\mathrm{sh}}(s)]$$

with the following parameters:

$$\mu_{\mathrm{sh},L}(s) = \mu_{\psi,L} \cdot \mu_{\mathrm{bg},L}(s) \tag{7}$$

$$\sigma^2_{\mathrm{sh},L}(s) = \sigma^2_{\psi,L} \cdot \mu^2_{\mathrm{bg},L}(s) \tag{8}$$

Regarding the u (and similarly to the v) component:

$$\mu_{\mathrm{sh},u}(s) = \mu_{\psi,u} + \mu_{\mathrm{bg},u}(s), \quad \sigma^2_{\mathrm{sh},u}(s) = \sigma^2_{\psi,u} \tag{9}$$

The estimation and the time dependence of parameters $[\overline{\mu}_\psi, \overline{\sigma}_\psi]$ are discussed in Section V-B.

C. Microstructural features

In this section, we define the 4^{th} dimension of the pixels' feature vectors (eq. (1)), which contains local microstructural responses.

1) Definition of the used microstructural features: Pixels covered by a foreground object often have different local textural features from the background at the same location, moreover, texture features may identify foreground points with background or shadow like color. In our model, texture features are used together with color components and they enhance the segmentation results as an additional component in the feature vector. Therefore, we make restrictions regarding the texture features: we search for components that we can get by low additional computing time from the existing model elements, in exchange for some accuracy.

According to our model, the textural feature is retrieved from a color feature-channel by using microstructural kernels. For practical reasons, and following the fact that the human visual system mainly percepts textures as changes in intensity, we use texture features only for the 'L' color component. A novelty of the proposed model is (as being explained in Section III-C.3) that we may use different kernels at different pixel locations. More specifically, there is a set of kernel coefficients for each site s: $K_s = \{a_s(r) | r \in N_s\}$, where N_s is the set of pixels around s covered by the kernel. Feature $x_T(s)$ is defined by:

$$x_T(s) = \sum_{r \in N_s} a_s(r) \cdot x_L(r) \tag{10}$$

2) Analytical estimation of the distribution parameters: Here, we show that with some further reasonable assumptions, the features defined by eq. (10) have also Gaussian distribution, and the distribution parameters $[\mu_{k,T}(s), \sigma_{k,T}(s)]$, $k \in \{\mathrm{bg}, \mathrm{sh}\}$ can be determined analytically.

As a simplification we exploit that the neighboring pixels have usually the same labels, and calculate the probabilities by:

$$p_k(s) = P(x_s | \omega_s = k) \approx P(x_s | \omega_r = k, r \in N_s)$$

This assumption is inaccurate near the border of the objects, but it is a reasonable approximation if the kernel size (and the size of set N_s) is small enough. To ensure this condition, we use 3×3 kernels in the following.

Accordingly, with respect to eq. (10), $x_T(s)$ in the background (and similarly in the shadow) can be considered as a linear combination of Gaussian random variables from the following set Λ_s:

$$\Lambda_s = \{x_L(r) |\; r \in N_s\} \tag{11}$$

where $x_L(r) \leftarrow N[\mu_{\mathrm{bg},L}(r), \sigma_{\mathrm{bg},L}(r)]$. We assume that the $x_L(r)$ variables have joint normal distribution, therefore, $x_T(s)$ is also Gaussian with parameters $[\mu_{\mathrm{bg},T}(s), \sigma_{\mathrm{bg},T}(s)]$. The mean value $\mu_{\mathrm{bg},T}(s)$ can be determined directly [41] by

$$\mu_{\mathrm{bg},T}(s) = \sum_{r \in N_s} a_s(r) \cdot \mu_{\mathrm{bg},L}(r) \tag{12}$$

On the other hand, to estimate the $\sigma_{\mathrm{bg},T}(s)$ parameter, we should model the correlation between the elements of Λ_s.

In effect, the $x_L(r)$ variables in Λ_s are non-independent, since fine alterations in global illumination or camera white balance cause correlated changes of the neighboring pixel values. However, very high correlation is not usual, since strongly textured details or simply the camera noise result in some independence of the adjacent pixel levels. While previous methods have ignored this phenomenon e.g. by considering the features to be uncorrelated [12], our goal is to give a more appropriate statistical model by estimating the order of correlation for a given scene.

We model the correlation factor between the 'adjacent' pixel values by a constant over the whole image. Let q and r be two sites in the neighborhood of s ($q, r \in N_s$), and denote the correlation coefficient between q and r by $c_{q,r}$. Accordingly,

$$c_{q,r} = \begin{cases} 1 & \text{if } q = r \\ c & \text{if } q \neq r \end{cases}$$

where c is a global constant. To estimate c, we randomly choose some pairs of neighboring sites. For each selected site pair (q, r), we make a set $I_{q,r}$ from time stamps corresponding to common background occurrences of pixels q and r. Thereafter, we calculate the normalized cross correlation $\hat{c}_{q,r}$ between time series $\{x_L^{[t]}(q) | t \in I_{q,r}\}$ and $\{x_L^{[t]}(r) | t \in I_{q,r}\}$, where t indices are time stamps of the x_L measurements. Finally, we approximate c by the average of the collected correlation coefficients $\hat{c}_{q,r}$ over all selected site pairs.

Thereafter, we can calculate $\sigma^2_{\mathrm{bg},T}(s)$ according to the variance theorem for sum of random variables [41]:

$$\sigma^2_{\mathrm{bg},T}(s) = \sum_{q,r \in N_s} a_s(q) \cdot a_s(r) \cdot \sigma_{\mathrm{bg},L}(q) \cdot \sigma_{\mathrm{bg},L}(r) \cdot c_{q,r} \tag{13}$$

Similarly, the Gaussian shadow parameters regarding the microstructural components by using eq. (7), (8), (12):

$$\mu_{\mathrm{sh},T}(s) = \sum_{r\in N_s} a_s(r)\cdot\mu_{\psi,L}\cdot\mu_{\mathrm{bg},L}(r) = \mu_{\psi,L}\cdot\mu_{\mathrm{bg},T}(s) \tag{14}$$

$$\sigma^2_{\mathrm{sh},T}(s) = \sigma^2_{\psi,L}\sum_{q,r\in N_s} b_{q,r}(s) \tag{15}$$

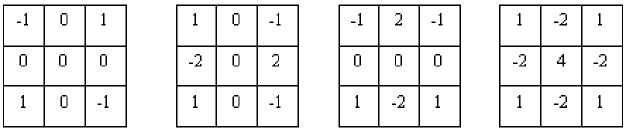
where

$$b_{q,r}(s) = a_s(q)\cdot a_s(r)\cdot\mu_{\mathrm{bg},L}(q)\cdot\mu_{\mathrm{bg},L}(r)\cdot c_{q,r}$$

3) Strategies for choosing kernels: In the following we deal with zero-mean kernels ($\forall s: \ \sum_{r\in N_s} a_s(r) = 0$) as a generalization of simple first-order edge features by [12]. Here we face an important problem from an experimental point of view. Each kernel has an adequate pattern, for which it generates a significant nonzero response, while most of the pixel-neighborhoods in an image are 'untextured' with respect to it. Therefore, one single kernel is unable to discriminate an 'untextured' object point on an 'untextured' background.

An evident enhancement uses several kernels which can recognize several patterns. However, increasing the number of the microstructural channels would intensify the noise, because at a given pixel position all the 'inadequate' kernels give irrelevant responses, which are accumulated in the energy term eq. (3).

To overcome the above problem, we use one microstructural channel only (see eq. (1)), and we use the most appropriate kernel at each pixel. Our hypothesis is: if the kernel response at s is significant in the background, the kernel gives more information for the segmentation there. Therefore, after we have defined a kernel set for the scene, at each pixel position s the kernel having the highest absolute response in the background centered at s is used. According to our experiments, different kernel-sets, e.g. corresponding to the Laws-filters [42], or the Chebyshev polynomials [43][42], produce similar results. In the following sections we use the kernels shown in Fig. 3, which we have found reasonable for the scenes. Regarding the 'Entrance pm' sequence, each kernel of the set corresponds to a significant number of background points according to our choice strategy (distributed as 44-19-22-15%), showing that each kernel is valuable.



-1	0	1
0	0	0
1	0	-1

1	0	-1
-2	0	2
1	0	-1

-1	2	-1
0	0	0
1	-2	1

1	-2	1
-2	4	-2
1	-2	1

Fig. 3. Kernel-set used in the experiments: 4 of the impulse response arrays corresponding to the 3×3 Chebyshev basis set proposed by [43]

IV. Foreground probabilities

The description of background and shadow characterizes the scene and illumination properties, consequently it has been possible to collect statistical information about them in time. In our case, the color distribution regarding the foreground areas is unpredictable in the same way. If the frame rate is very low and unbalanced, we must consider consecutive images containing different scenarios with different objects. Previous works [21][22] used uniform distribution to describe the foreground process which agrees with the long-term color statistics of the foreground pixels (Fig. 2), but it presents a weak description of the class. Since the observed feature values generated by the foreground, shadow and background processes overlap strongly in numerous real world scenes, many foreground pixels are misclassified that way.

Instead of temporal statistics we use spatial color information to overcome this problem by using the following assumption: whenever s is a foreground pixel, we should find foreground pixels with similar color in the neighborhood. Consequently, if we can estimate the color statistics of the nearby foreground sites, we can decide if a pixel with a given color is likely part of the foreground or not. Unfortunately, when we want to assign a probability value to a given pixel describing its foreground membership, the positions of the nearby foreground pixels are also unknown. However, to estimate the local color distribution, we do not need to find all foreground pixels, just some samples in each neighborhood. The key point is that we identify some pixels which *certainly* correspond to the foreground: these are the pixels having significantly different levels from the locally estimated background and shadow values, thus they can be found by a simple thresholding:

$$\omega^0_s = \begin{cases} \text{fg} & \text{if } \ (\epsilon_{\text{bg}}(s) > \zeta) \text{ AND } (\epsilon_{\text{sh}}(s) > \zeta) \\ \text{bg} & \text{otherwise} \end{cases} \tag{16}$$

where ζ is a threshold (which is analogous with the uniform value in previous models [22] choosing $\epsilon_{\text{fg}}(s) = \zeta$), and ω^0_s is a 'preliminary' segmentation label of s.

Next, we estimate for each pixel s the local color distribution of the foreground, using the *certainly* foreground pixels in the neighborhood of s. The procedure is demonstrated in Fig. 4 (for easier visualization with 1D grayscale feature vectors). We use the following notations: F denotes the set of pixels marked as *certainly* foreground elements in the preliminary mask:

$$F = \{r \mid r\in S,\ w^0_r = \text{fg}\}$$

Note that F may be a coarse estimation of the foreground (Fig. 4b).

Let be V_s the set of the neighboring pixels around s, considering a rectangular neighborhood with window size $m\times m$ (Fig. 4a). Thereafter, F_s is defined with respect to s as the set of neighboring pixels determined as 'foreground' by the preprocessing step: $F_s = F\cap V_s$ (Fig. 4c).

The foreground color distribution around s can be characterized by a normalized histogram h_s over F_s (Fig. 4d). However, instead of using the noisy h_s directly, we approximate it by a 'smoothed' probability density function, $f_s(\overline{x})$, and determine the foreground probability term as $p_{\text{fg}}(s) = f_s(\overline{x}_s)$.[1]

To deal with multi-colored or textured foreground components, the estimated $f_s(.)$ function should be multi-modal (see a

[1] In the spatial foreground model, we must ignore the textural component of $\overline{x}$, since different kernels are used in different pixel locations, and the microstructural responses of the various pixels may be incomparable. Thus in this section, $\overline{x}$ is considered to be a three dimensional color vector, and h_s a three dimensional histogram.

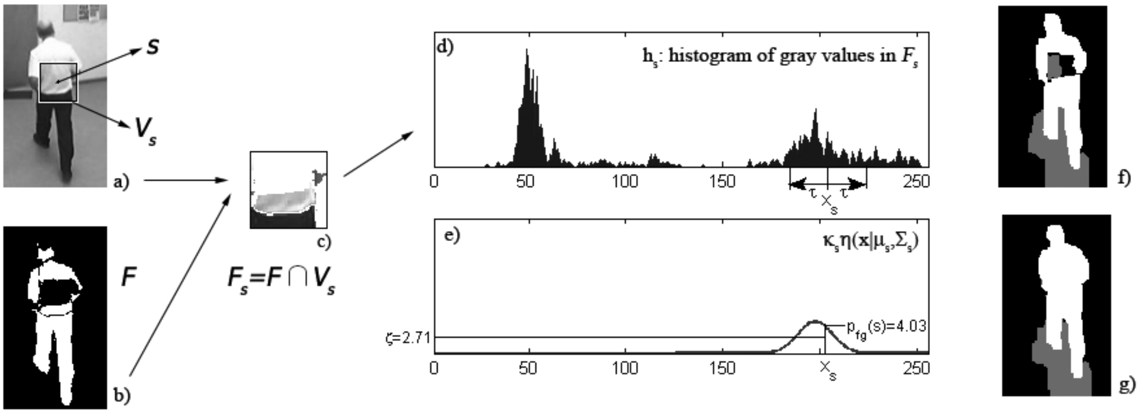


Fig. 4. Determination of the foreground conditional probability term for a given pixel s (demonstrated in grayscale). a) video image, with marking s and its neighborhood V_s (with window side $m = 45$). b) noisy preliminary foreground mask c) Set F_s: preliminary detected foreground pixels in V_s. (Pixels of $V_s \backslash F_s$ are marked with white.) d) Histogram of F_s, marking x_s, and its τ neighborhood e) Result of fitting a weighted Gaussian term for the $[x_s - \tau, x_s + \tau]$ part of the histogram. Here, $\zeta = 2.71$ is used (it would be the foreground probability value for each pixel according to the 'uniform' model), but the procedure increases the foreground probability to 4.03. f) Segmentation result of the model optimization with the uniform foreground calculus g) Segmentation result by the proposed model

bimodal case in Fig. 4d). Note that we use $f_s(.)$ only to calculate the foreground probability value of s as $f_s(\overline{x}_s)$. Thus, it is enough to estimate the parameters of the mode of $f_s(.)$, which covers $\overline{x}_s$ (see Fig. 4e). Therefore, we consider $f_s(.)$ as a mixture of a weighted Gaussian term $\eta(.)$ and a residual term $\vartheta_s(.)$, for which we only prescribe that $\vartheta_s(.)$ is a probability density function and $\vartheta_s(\overline{x}) = 0$ if $\|\overline{x}_s - \overline{x}\| < \tau$. ($\kappa_s$ is a weighting factor: $0 < \kappa_s < 1$.) Hence,

$$f_s(\overline{x}) = \Big[\kappa_s \cdot \eta(\overline{x}|\overline{\mu}_s, \overline{\overline{\Sigma}}_s) + (1 - \kappa_s) \cdot \vartheta_s(\overline{x})\Big]$$

Accordingly, the foreground probability value of site s is statistically characterized by the distribution of its neighborhood in the color domain:

$$\epsilon_{\text{fg}}(s) = -\log f_s(\overline{x}_s) = -\log \kappa_s - \log \eta(\overline{x}_s|\overline{\mu}_s, \overline{\overline{\Sigma}}_s)$$

The steps of the foreground energy calculation are detailed in Fig. 5. We can speed up the algorithm, if we calculate the Gaussian parameters by considering only some randomly selected pixels in F_s [19]. We describe the parameter settings in Section V-A and in Table II.

V. Parameter settings

Our method works with scene-dependent and condition-dependent parameters. *Scene-dependent* parameters can be considered constant in a specific field, and are influenced by, e.g. camera settings, a-priori knowledge about the appearing objects or reflection properties. We provide strategies on how to set these parameters if a surveillance environment is given. *Condition-dependent* parameters vary in time in a scene, therefore, we use adaptive algorithms to follow them.

We emphasize two properties of the presented model. Regarding the background and shadow processes, only the one dimensional marginal distribution parameters should be estimated (Section III-A). On the other hand, we should estimate here the color-distribution parameters only, since the mean-deviation values corresponding to the microstructural component are determined analytically (see Section III-C.2).

Algorithm 1: foreground probability calculation

1) The pixels of F_s whose pixel values are close enough to $\overline{x}_s$ are collected into a set:
$$F_s^D = \{r \mid r \in F_s,\ \|\overline{x}_s - \ \overline{x}_r\| < \tau\}$$
2) The empirical mean and deviation values are calculated regarding the color levels of set F_s^D: $\overline{\mu}_s^D$, $\overline{\sigma}_s^D$. These values estimate the mean and deviation parameters of the Gaussian component $\eta(.)$.
3) Denote by $\#H$ the number of the elements in a given set H. $\kappa_s^{(1)} = \frac{\#F_s^D}{\#F_s}$ is introduced as the ratio of the number of pixels with similar color to s and all pixels, among the neighboring foreground initialized sites.
4) An extra term is used to keep the probability low if there are none or only a few foreground pixels in the neighborhood. Denote by $\kappa_s^{(2)} = \frac{\#F_s}{m^2}$ the ratio of the number of pixels in F_s and the size of the neighborhood V_s. This term biases the weight through a sigmoid function:
$$\kappa_s = \kappa_s^{(1)} \cdot \frac{1}{1 + \exp\left[-(\kappa_s^{(2)} - \kappa_{\min}/2)\right]} \qquad (17)$$
5) Finally, the energy term is calculated as:
$$\epsilon_{\text{fg}}(s) = -\log \kappa_s - \log \eta(\overline{x}_s, \overline{\mu}_s^D, \overline{\sigma}_s^D) \qquad (18)$$

Fig. 5. Algorithm for the estimation of the foreground probability term. Notations are defined in Section IV.

A. Background and foreground model parameters

The *background* parameter estimation and update procedure is automated, based on the work in [4], which presents reasonable results, and it is computationally more effective than the standard EM algorithm.

The *foreground* model parameters (Section IV) correspond to a-priori knowledge about the scene, e.g. the expected size of the appearing objects and the contrast. These features exploit basically low-level information and are quite general, therefore the method is able to consider a large variety of moving objects in a scene. In our experiments, we set these

TABLE II

FOREGROUND PARAMETER SETTINGS

Parameter	Definition and setting strategy
m	the size of the neighborhood window V_s in pixels considered in the process. It depends on the expected size of the objects in the scene, used $m = 1/3\sqrt{T_B}$, where T_B is the approximate average territory of the objects' bounding boxes
$\kappa_{\min}$	control parameter for the minimum required number of pre-classified foreground pixels in the neighborhood. If the ratio of the pixels and the size of the neighborhood is smaller than $\kappa_{\min}$, the foreground probability will be low there, due to the sigmoid function of eq. (17). Small $\kappa_{\min}$ increases the number of detected foreground pixels and can be used if the objects are of compact shape like in the sequence 'Highway'. Otherwise small $\kappa_{\min}$ causes high false foreground detection rate. Applying $\kappa_{\min} = 0.1$ for vehicle monitoring and $\kappa_{\min} = 0.25$ for pedestrians (including cyclists, baby carriages etc.) proved to be good.
τ	the threshold parameter which defines the maximum distance in the feature space between pixels generated by one Gaussian process. We use outdoors in high contrast, $\tau = 0.2 \cdot d_{\max}$, indoors $\tau = 0.1 \cdot d_{\max}$, where $d_{\max}$ is the maximum occurring distance in the feature space.

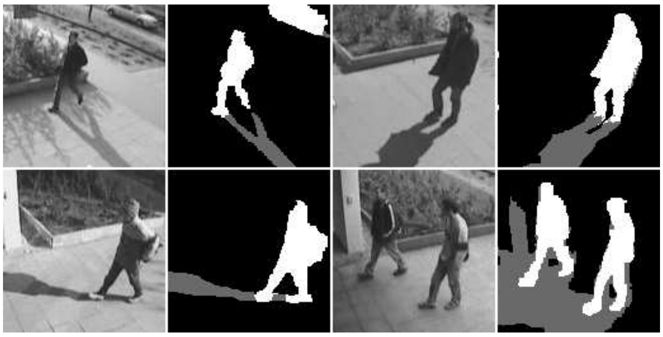


Fig. 6. Different periods of the day in the 'Entrance' sequence, segmentation results. Above left: in the morning ('am'), right: at noon, below left: in the afternoon ('pm'), right: wet weather.

parameters empirically. Table II shows a detailed overview on the foreground parameters and how to set them. Notes on parameter ζ are given in Section VII and in Fig. 15.

B. Shadow parameters

The changes in the global illumination significantly alter the shadow properties (Fig. 6). Moreover, changes can be performed rapidly: indoors due to switch on/off different light sources, while outdoors due to the appearance of clouds.

Regarding the shadow parameter settings, we discriminate parameter initialization and re-estimation. From a practical point of view, initialization may be supervised with marking shadowed regions in a few video frames by hand, once after switching on the system. Based on the training data, we can calculate maximum likelihood estimates of the shadow parameters. On the other hand, there is usually no opportunity for continuous user interaction in an automated surveillance environment, thus the system must adopt the illumination changes raising a claim to an automatic re-estimation procedure.

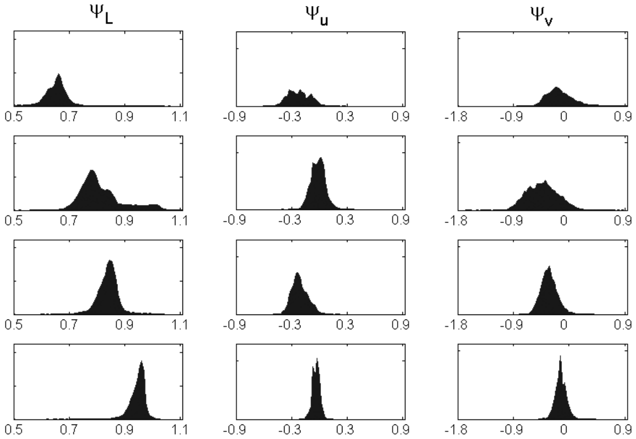


Fig. 7. **Shadow $\overline{\psi}$ statistics** on four sequences recorded by the 'Entrance' camera of our University campus. Histograms of the occurring ψ_L, ψ_u and ψ_v values of shadowed points. Rows correspond to video shots from different parts of the day. We can observe, the peak of the ψ_L histogram strongly depends on the illumination conditions, while the change in the other two shadow parameters is much smaller.

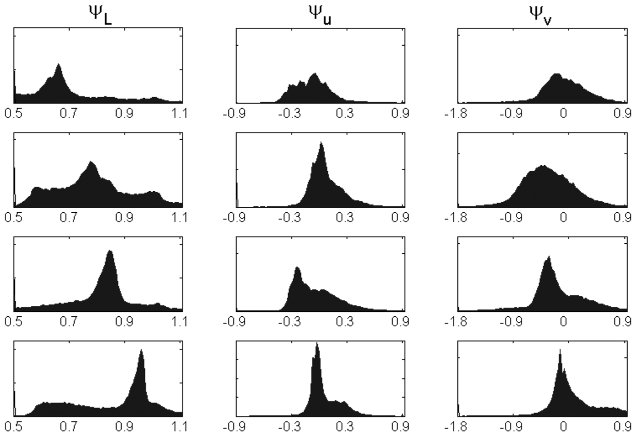


Fig. 8. $\overline{\psi}$ **statistics for all non-background pixels**. Histograms of the occurring ψ_L, ψ_u and ψ_v values of the non-background pixels in the same sequences as in Figure 7.

For the above reasons, we use supervised initialization, and focus on the parameter adaption process in the following. The presented method is built into a 24-hour surveillance system of our university campus. We validate our algorithm via four manually evaluated ground truth sequences captured by the same camera under different illumination conditions (Fig. 6). According to section III-B, the shadow parameters are 6 scalars: 3-3 components of $\overline{\mu}_\psi$ respectively $\overline{\sigma}_\psi$ vectors. Fig. 7 shows the one-dimensional histograms for the occurring ψ_L, ψ_u and ψ_v values of shadowed points for each video shot. We can observe that while the variation of parameters $\overline{\sigma}_\psi$, $\mu_{\psi,u}$ and $\mu_{\psi,v}$ are low, $\mu_{\psi,L}$ varies in time significantly. Therefore, we update the parameters in two different ways.

1) Re-estimation of parameters $[\mu_{\psi,u}, \sigma_{\psi,u}]$ and $[\mu_{\psi,v}, \sigma_{\psi,v}]$: The procedure is similar to which was used in [22]. We show it regarding the u component only, since the v component is updated in the same way.

We re-estimate the parameters at fixed time-intervals $\mathcal{T}$.

Denote $\mu_{\psi,u}[t], \sigma_{\psi,u}[t]$ the parameters at time t. W_t is the set containing the observed ψ_u values collected over the pixels detected as shadow between time t and $t + \mathcal{T}$:

$$W_t = \{\psi_u^{[\phi]}(s) | \phi = t, \ldots, t + \mathcal{T} - 1,\ \omega_s^{[\phi]} = \text{sh},\ s \in S\}$$

where upper index $[\phi]$ refers to time, $\#W_t$ is the number of the elements, M_t and D_t are the empirical mean and the standard deviation values of W_t. We update the parameters:

$$\mu_{\psi,u}[t + \mathcal{T}] = (1 - \xi_t)\mu_{\psi,u}[t] + \xi_t M_t$$

$$\sigma^2_{\psi,u}[t + \mathcal{T}] = (1 - \xi_t)\sigma^2_{\psi,u}[t] + \xi_t D_t^2$$

Parameter ξ_t is a weighting term ($0 \le \xi_t \le 1$) depending on $\#W_t$, namely greater number of detected shadow points increase ξ_t and the influence of the M_t respectively D_t^2 term. We use $\mathcal{T} = 60$ sec.

2) Re-estimation of parameters $[\mu_{\psi,L}, \sigma_{\psi,L}]$*:* Parameter $\mu_{\psi,L}$ corresponds to the average background luminance darkening factor of the shadow. Except from window-less rooms with constant lightning, $\mu_{\psi,L}$ is strongly condition dependent. Outdoors, it can vary between 0.6 in direct sunlight and 0.95 in overcast weather. The simple re-estimation from the previous section does not work in this case, since the illumination properties between time t and $t + \mathcal{T}$ may rapidly change a lot, which would result in absolutely false detected shadow values in set W_t presenting false M_t and D_t parameters for the re-estimation procedure.

For this reason, we derive the actual $\mu_{\psi,L}$ from the statistics of all non-background ψ_L-s (where the background filtering should be done by a good approximation only, we use the Stauffer-Grimson algorithm). In Fig. 8 we can observe that the peaks of the 'non-background' ψ_L-histograms are approximately in the same location as they were in Fig. 7. The video shots corresponding to the first and second rows were recorded around noon where the shadows were relatively small, however, the peak is still in the right place in the histogram.

These experiments encourage us to identify $\mu_{\psi,L}$ with the location of the peak on the 'non-background' ψ_L-histograms for the scene.

The description of the update-algorithm of $\mu_{\psi,L}$ is as follows. We define a data structure which contains a ψ_L value with its timestamp: $[\psi_L, t]$. We store the 'latest' occurring $[\psi_L, t]$ pairs of the non-background points in a set Q, and update the histogram h_L of the ψ_L values in Q continuously. The key point is the management of set Q. We define MAX and MIN parameters which control the size of Q. The queue management algorithm, which is introduced in Fig. 9, follows four intentions:

- Q contains always the latest available ψ_L values.
- The algorithm keeps the size of Q between prescribed bounds MAX and MIN ensuring the topicality and relevancy of the data contained.
- The actual size of Q is around MAX in case of cluttered scenarios.
- In the case of few or no motion in the scene, the size of Q decreases until MIN. This fact increases the influence of the forthcoming elements, and causes quicker adaptation, since it is faster to modify the shape of a smaller histogram.

Algorithm 2: updating the $\mu_{\psi,L}$ shadow parameter

```
1) For each frame $t$ we determine:
       $\Psi_t = \{ [\psi_L^{[t]}(s), t] \mid s \in S,\ \omega_s^{[t]} \neq \text{bg}\}$
2) We append $\Psi_t$ to $Q$.
3) We may remove elements from $Q$:
     • if $\#Q <$ MIN, we keep all the elements.
     • if $\#Q \geq$ MIN we find the eldest timestamp $t_e$ in $Q$
       and remove all the elements from $Q$ with time stamp $t_e$.
4) If $\#Q >$ MAX after step 3: in order of their timestamp
   we remove further ('old') elements from $\#Q$ till we reach
   $\#Q \leq$ MAX.
5) We update the histogram $h_L$ regarding $Q$ and apply:
       $\mu_{\psi,L}^{[t+1]} = \text{argmax}\{h_L\}$
```

Fig. 9. Updating algorithm for parameter $\mu_{\psi,L}$.

Parameter $\sigma_{\psi,L}$ is updated similarly to $\sigma_{\psi,u}$ but only in the time periods when $\mu_{\psi,L}$ does not change significantly.

Note that the above update process may fail in scenarios free of shadows. However, that case occurs mostly under artificial illumination conditions, where the shadow detector module can be switched off using a priori knowledge.

VI. MRF OPTIMIZATION

The MAP estimator in eq. (2) is realized by combining a conditional independent random field of signals and an unconditional Potts model [37]. The optimal segmentation corresponds to the global labeling, $\widehat{\Omega}$, defined by

$$\widehat{\Omega} = \text{argmin}_{\Omega} \sum_{s \in S} \underbrace{- \log P(\overline{x}_s | \omega_s)}_{\epsilon_{\omega_s}(s)} + \sum_{r,s \in S} \Theta(\omega_r, \omega_s) \qquad (19)$$

where the minimum is searched over all the possible segmentations (Ω) of a given input frame. The first part of eq. (19) contains the sum of the local class-energy terms regarding the pixels of the image (see eq. (3) and eq. (18)). The second part is responsible for the smooth segmentation: $\Theta(\omega_r, \omega_s) = 0$ if s and r are not neighboring pixels, otherwise:

$$\Theta(\omega_r, \omega_s) = \begin{cases} -\beta & \text{if } \omega_r = \omega_s \\ +\beta & \text{if } \omega_r \neq \omega_s \end{cases}$$

In applications using the Potts-MRF models, the quality of the segmentation depends both on the appropriate probabilistic model of the classes, and on the optimization technique which finds a good global labeling with respect to eq. (19). The latter factor is a key issue, since finding the global optimum is NP hard [44]. On the other hand, stochastic optimizers using simulated annealing (SA) [20][45] and graph cut techniques [44][46] have proved to be practically efficient offering a ground to validate different energy models.

The results shown in Section VII have been generated by a SA algorithm which uses the Metropolis criteria [47] for accepting new states[2], while the cooling strategy changes the temperature after a fixed number of iterations. The relaxation parameters are set by trial and error taking aim at the maximal quality. Comparing the proposed model to reference MRF methods is done using the same parameter settings.

After verifying our model with the above stochastic optimizer, we have also tested some quicker techniques for practical purposes. We have found the deterministic Modified Metropolis (MMD) [36] relaxation algorithm similarly efficient but significantly faster for this task: processing 320×240 images runs with 1 fps. We note that a coarse but quick MRF optimization method is the ICM algorithm [48]. If we use ICM with our model, the running speed is 3 fps, in exchange for some degradation in the segmentation results.

VII. Results

The goal of this section is to demonstrate the benefit of using the introduced contributions of the paper: the novel foreground calculus, the shadow model and the benefit of the textural features. The demonstration is done in two ways: in Fig. 10–15 we show segmented images by the proposed and previous methods, while regarding three sequences we perform numerical evaluation.

A. Test sequences

We have validated our method on several test sequences. Here, we show results regarding the following 7 videos:

- ‘Laboratory’ test sequence from the benchmark set [35]. This shot contains a simple environment where previous methods [12] have already produced accurate results.
- ‘Highway’ video [35]. This sequence contains dark shadows, but homogenous background without illumination artifacts. In contrast with [21] our method reaches the appropriate results without post processing, which is strongly environment-dependent.
- ‘Corridor’ indoor surveillance video. Although, it is on the face of a simple office environment the bright objects and background elements often saturate the image sensors and it is hard to accurately separate the white shirts of the people from the white walls in the background.
- 4 surveillance video sequences captured by the ‘Entrance’ (outdoor) camera of our university campus in different lightning conditions. (Fig 6). These sequences contain difficult illumination and reflection effects and suffer from sensor saturation (dark objects and shadows). Here, the presented model improves the segmentation results significantly versus previous methods.

B. Demonstration of the improvements via segmented images

In the introduction we gave an overview on the state-of-the art methods (Table I) indicating their way of (i) shadow detection (ii) foreground modeling (iii) textural analysis.

1) Comparison of shadow models: Results of different shadow detectors are demonstrated in Fig. 11. For the sake of comparison we have implemented in the same framework an illumination invariant (‘II’) method based on [29], and a constant ratio model (‘CR’), similarly to [21]. We have observed that the results of the previous and the proposed methods are similar in simple environments, but our improvements become significant in the surveillance scenes:

- In the *‘Laboratory’* sequence, the ‘II’ approach is reasonable, while the ‘CR’ and the proposed method are similarly accurate.
- Regarding the *‘Highway’* video, although the ‘II’ and ‘CR’ find the objects without shadows approximately, the results are much noisier than it is with our model.
- On the *‘Entrance am’* surveillance video, the ‘II’ method fails completely: shadows are not removed, while the foreground component is also noisy due to the lack of luminance features in the model. The ‘CR’ model also produces poor results: due to the long shadows and various field objects the constant ratio model becomes inaccurate. Our model handles these artifacts robustly.

The improvements of the proposed method versus the ‘CR’ model can be also observed in Fig. 14 ($2^{\rm nd}$ and $5^{\rm th}$ row).

2) Comparison of foreground models: In this paper we have proposed a basically new approach regarding foreground modeling, which needs neither high frame rate, in contrast to [3][11][12], nor high level object descriptors [15]. Other previous models [21][22] that have used the uniform calculus expressing foreground may generate any colors in a given domain with the same probability. As it is shown in Fig. 12, 13 and 14 ($3^{\rm rd}$ and $5^{\rm th}$ rows), the uniform model is often a coarse approximation, and our method is able to improve the results significantly. Moreover, we have observed that our model is robust with respect to fine changes in the threshold parameter ζ (Fig. 15, $3^{\rm rd}$ row). On the other hand, the uniform model is highly sensitive to set ζ appropriately, even in scenarios which can be segmented properly with an adequate uniform value (Fig. 15, $2^{\rm nd}$ row).

3) Microstructural features: Complementing the pixel-level feature vector with the microstructural component enhances the segmentation result if the background or the foreground is textured. To demonstrate the additional information, Fig. 10 shows a synthetic example. Consider Fig. 10a as a frame of a sequence where the bright rectangle in the middle corresponds to the foreground (image v. shows an enlarged part of it). The background consists of four equal rectangular regions, each of them has a particular texture, which are enlarged in i-iv. images. Similarly to the real-world case, the observed pixel values are affected by Gaussian noise. Below, we can see results of background subtraction. First (image b), the feature vector only consists of the gray value of the pixel. Secondly (image c), we complete it with horizontal and vertical edge detectors similarly to [12]. Finally (image d), we use the kernel set of Fig. 3, with the proposed kernel selection strategy, providing the best results.

In Fig 14, the $4^{\rm th}$ and $5^{\rm th}$ rows show the segmentation results with and without the textural components, improvements are

[2] A state is a candidate for the optimal segmentation.

observable in the fine details, especially near the legs of the people in the magnified regions.

C. Numerical evaluation

The quantitative evaluations are done through manually generated ground truth sequences. Since the goal is foreground detection, the crossover between shadow and background does not count for errors.

Denote the number of correctly identified foreground pixels of the evaluation sequence by TP (*true positive*). Similarly, we introduce FP for misclassified non-foreground points, and FN for misclassified foreground points.

The evaluation metrics consists of the *Recall* rate and the *Precision* of the detection.

$$\text{Recall} = \frac{TP}{TP+FN} \qquad \text{Precision} = \frac{TP}{TP+FP}$$

For numerical validation, we used 100 frames from the ‘Entrance pm’ sequence and 50-50 frames from the ‘Highway’ and ‘Entrance am’ video shots.

Advantages of using Markov Random Fields versus morphology based approaches were examined previously [12][19], therefore, we focus on the state-of-the-art MRF models. The evaluation of the improvements is done by exchanging our new model elements one by one for the latest similar solutions in the literature, and we compare the segmentation results.

Regarding shadow detection, the ‘CR’ model is the reference, and we compare the foreground model to the ‘uniform’ calculus again.

In Table III, we compare the shadow and foreground model to the reference methods. The results confirm that our shadow calculus improves the precision rate, since it decreases the number of false negatively detected shadow pixels significantly. Due to the proposed foreground model, the recall rate increases through detecting several background/shadow colored foreground parts. If we ignore both improvements both evaluation parameters decrease (#1 in Table III).

VIII. Conclusion

The present paper has introduced a general model for foreground segmentation without any restrictions on a-priori probabilities, image quality, objects’ shapes and speed. The frame rate of the source videos might also be low or unstable, and the method is able to adapt to the changes in lighting conditions. We have contributed to the state-of-the-art in three areas: (1) we have introduced a more accurate, adaptive shadow model; (2) we have developed a novel description for the foreground based on spatial statistics of the neighboring pixel values; (3) We have shown how different microstructure responses can be used in the proposed framework as additional feature components improving the results.

We have compared each contribution of our model to previous solutions in the literature, and observed its superiority. The proposed method now works in a real-life surveillance system (see Fig. 6) and its efficiency has been validated.

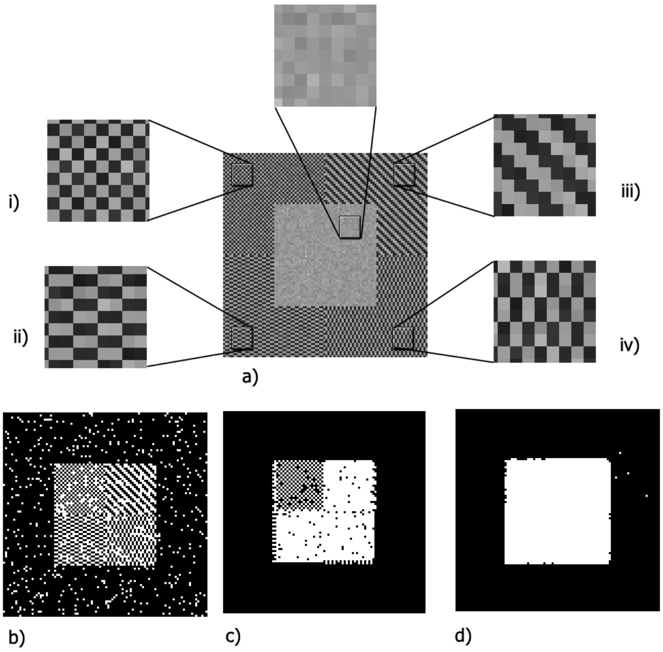


Fig. 10. Synthetic example to demonstrate the benefits of the microstructural features. a) input frame, i-v) enlarged parts of the input, b-d) result of foreground detection based on: (b) gray levels (c) gray levels with vertical and horizontal edge features [12] (d) proposed model with adaptive kernel

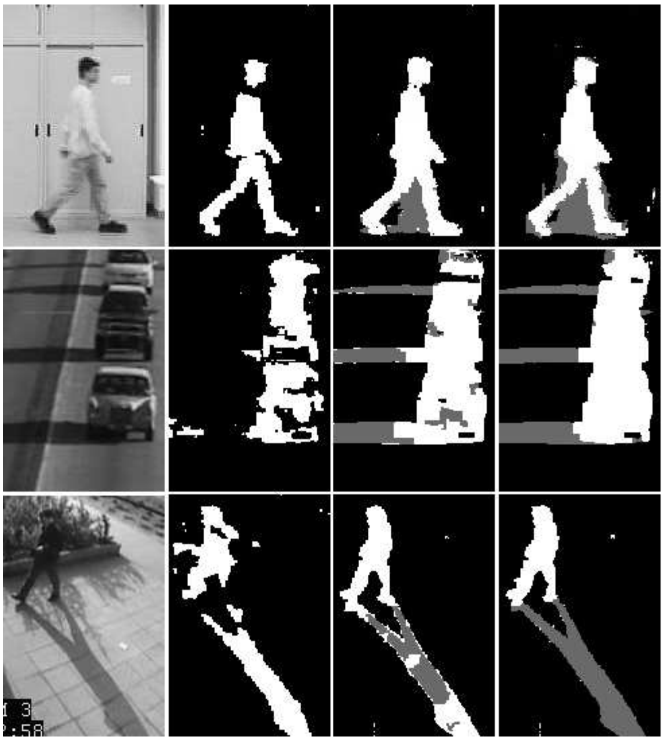


Fig. 11. *Shadow model validation:* Comparison of different shadow models in 3 video sequences (From above: ‘Laboratory’,‘Highway’,‘Entrance am’) . Col. 1: video image, Col. 2: $C_1C_2C_3$ space based illumination invariants [29]. Col. 3: ‘constant ratio model’ by [21] (without object-based postprocessing) Col 4: Proposed model

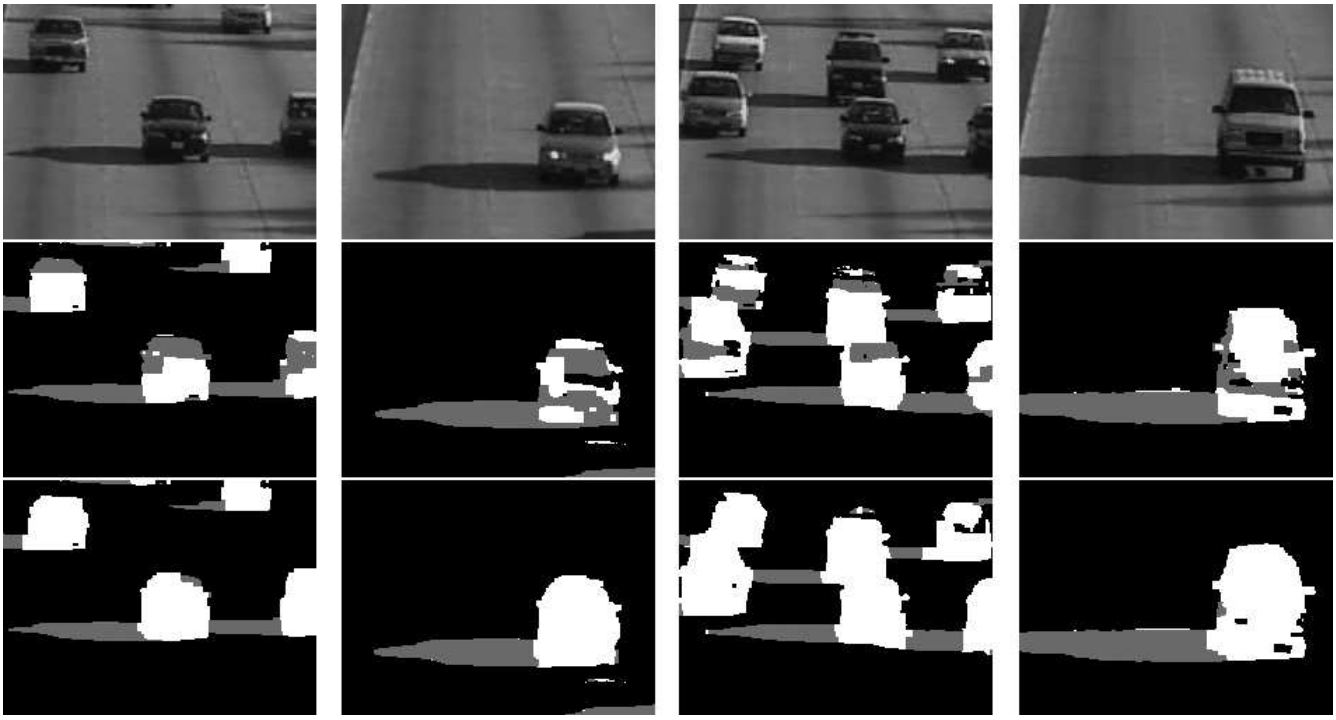


Fig. 12. *Foreground model validation:* Segmentation results on the 'Highway' sequence. Row 1: video image; Row 2: results by uniform foreground model; Row 3: Results by the proposed model

TABLE III

VALIDATION OF THE MODEL ELEMENTS. RESULTS WITH (#1) 'CONSTANT RATIO' SHADOW MODEL WITH THE 'UNIFORM' FOREGROUND MODEL (#2) 'CONSTANT RATIO' SHADOW MODEL WITH THE PROPOSED FOREGROUND MODEL (#3) 'UNIFORM' FOREGROUND MODEL WITH THE PROPOSED SHADOW MODEL, (#4) RESULTS WITH OUR PROPOSED SHADOW AND FOREGROUND MODEL

Video	Recall				Precision			
	#1	#2	#3	**#4**	#1	#2	#3	**#4**
Entrance pm	0.89	0.97	0.85	**0.96**	0.66	0.62	0.85	**0.83**
Entrance am	0.85	0.92	0.86	**0.93**	0.62	0.63	0.82	**0.81**
Highway	0.82	0.84	0.86	**0.90**	0.73	0.72	0.80	**0.80**

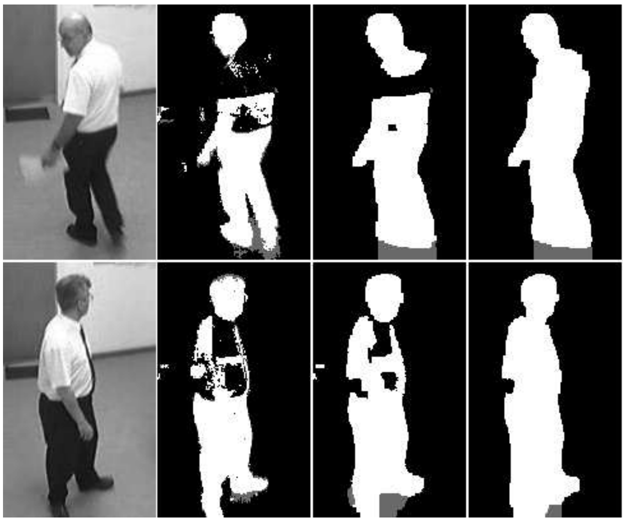


Fig. 13. *Foreground model validation* regarding the 'Corridor' sequence. Col. 1: video image, Col. 2: Result of the preliminary detector. Col. 3: Result with uniform foreground calculus Col 4: Proposed foreground model

ACKNOWLEDGMENT

The authors would like to thank Zoltan Kato, Levente Kovács and Zoltán Szlávik for their kind remarks, and the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

[1] S. C. Zhu and A. L. Yuille, "A flexible object recognition and modeling system," *Int'l Journal of Computer Vision*, vol. 20, no. 3, 1996.

[2] L. Havasi, Z. Szlávik, and T. Szirányi, "Higher order symmetry for non-linear classification of human walk detection," *Pattern Recognition Letters*, vol. 27, pp. 822–829, 2006.

[3] Y. Sheikh and M. Shah, "Bayesian modeling of dynamic scenes for object detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1778–1792, 2005.

[4] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[5] Y. Zhou, Y. Gong, and H. Tao, "Background segmentation using spatial-temporal multi-resolution MRF," in *Workshop on Motion and Video Computing*. IEEE, 2005, pp. 8–13.

[6] A. Licsár, L. Czúni, and T. Szirányi, "Adaptive stabilization of vibration on archive films," *Lecture Notes in Computer Science, CAIP'2003*, vol. LNCS 2756, pp. 230–237, 2003.

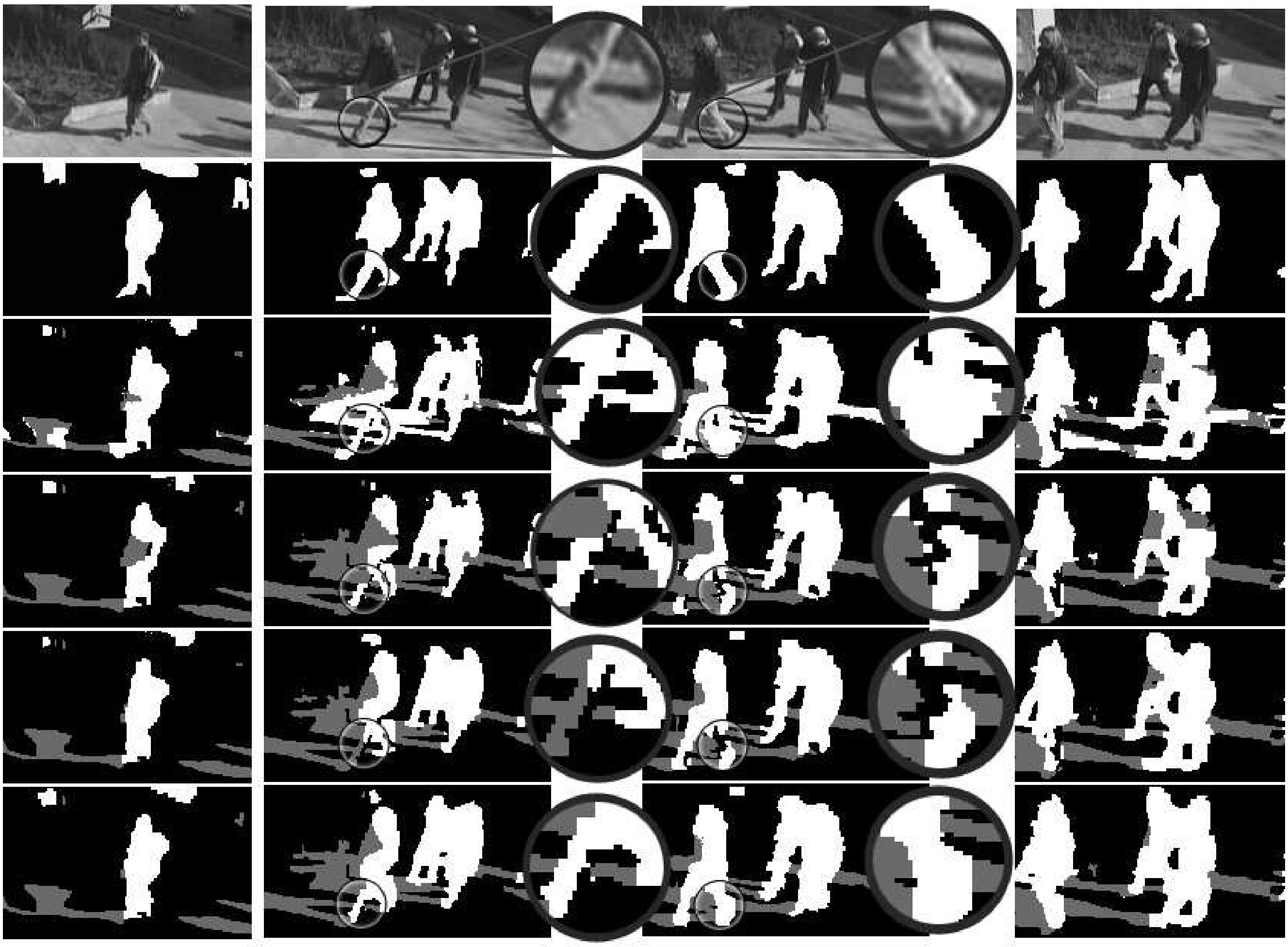


Fig. 14. *Validation of all improvements* in the segmentation regarding 'Entrance pm' video sequence Row 1. Video frames, Row 2. Ground truth Row 3. Segmentation with the 'constant ratio' shadow model [21], Row 4. Our shadow model with 'uniform foreground' calculus [22] Row 5. The proposed model without microstructural features Row 6. Segmentation results with our final model.

[7] M. Heikkila and M. Pietikainen, "A texture-based method for modeling the background and detecting moving objects," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 657–662, 2006.

[8] J. Zhong and S. Sclaroff, "Segmenting foreground objects from a dynamic textured background via a robust Kalman filter," in *Proc. IEEE International Conference on Computer Vision*, 2003, pp. 44–50.

[9] S. Chaudhuri and D. Taur, "High-resolution slow-motion sequencing: how to generate a slow-motion sequence from a bit stream," *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 16–24, 2005.

[10] J. Kato, T. Watanabe, S. Joga, L. Ying, and H. Hase, "An HMM/MRF-based stochastic framework for robust vehicle tracking," *IEEE Trans. on Intelligent Transportation Systems*, vol. 5, no. 3, pp. 142–154, 2004.

[11] J. Rittscher, J. Kato, S. Joga, and A. Blake, "An HMM-based segmentation method for traffic monitoring," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1291–1296, 2002.

[12] Y. Wang, K.-F. Loe, and J.-K. Wu, "A dynamic conditional random field model for foreground and shadow segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 279–289, 2006.

[13] R. Cutler and L. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 781–796, August 2000.

[14] L. Czúni and T. Szirányi, "Motion segmentation and tracking with edge relaxation and optimization using fully parallel methods in the cellular nonlinear network architecture," *Real-Time Imaging*, vol. 7, no. 1, pp. 77–95, 2001.

[15] A. Yilmaz, X. Li, and M. Shah, "Object contour tracking using level sets," in *Proc. Asian Conference on Computer Vision, (ACCV 2004)*, Jaju Islands, Korea, 2004.

[16] A. Cavallaro, E. Salvador, and T. Ebrahimi, "Detecting shadows in image sequences," in *Proc. of European Conference on Visual Media Production*, March 2004, pp. 167–174.

[17] R. Cucchiara, C. Grana, G. Neri, M. Piccardi, and A. Prati, "The Sakbot system for moving object detection and tracking," in *Video-Based Surveillance Systems-Computer Vision and Distributed Processing*, 2001, pp. 145–157.

[18] K. Siala, M. Chakchouk, F. Chaieb, and O. Besbes, "Moving shadow detection with support vector domain description in the color ratios space," in *Proc. International Conference on Pattern Recognition*, vol. 4, 2004, pp. 384–387.

[19] Cs. Benedek and T. Szirányi, "Markovian framework for foreground-background-shadow separation of real world video scenes," in *Proc. Asian Conference on Computer Vision (ACCV)*, vol. LNCS 3851. Hyderabad, India: Springer, Jan. 2006, pp. 898–907.

[20] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 721–741, 1984.

[21] I. Mikic, P. Cosman, G. Kogut, and M. M. Trivedi, "Moving shadow and object detection in traffic scenes," in *Proc. International Conference on Pattern Recognition*, 2000.

[22] Y. Wang and T. Tan, "Adaptive foreground and shadow detection in image sequences," in *Proc. International Conference on Pattern Recognition*, 2002, pp. 983–986.

[23] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr, "Interactive image segmentation using an adaptive GMMRF model," in *Proc. European Conference on Computer Vision*. Springer, 2004, pp. 456–468.

[24] Z. Kato, T. C. Pong, and G. Q. Song, "Multicue MRF image segmentation: Combining texture and color," in *Proc. of International Conference on Pattern Recognition*, Quebec, Canada, Aug. 2002, pp. 660–663.

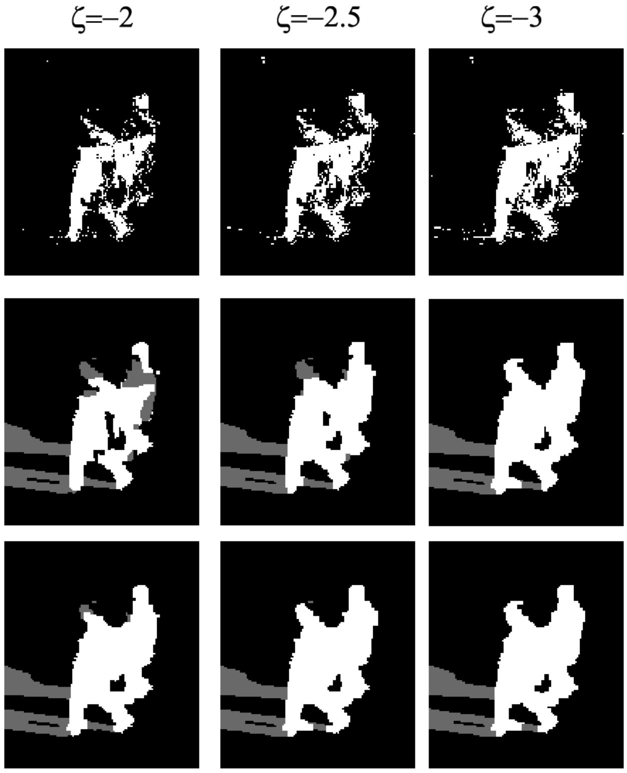


Fig. 15. *Effect of changing the ζ foreground threshold parameter*. Row 1: preliminary masks (F), Row 2: results with uniform foreground calculus using $\epsilon_{\rm fg}(s) = \zeta$, Row 3. results with the proposed model. Note: for the uniform model, $\zeta = -2.5$ is the optimal value with respect to the whole video sequence.

[25] Cs. Benedek and T. Szirányi, “A Markov random field model for foreground-background separation,” in *Proc. Joint Hungarian-Austrian Conference on Image Processing and Pattern Recognition (HACIPPR)*, Veszprém, Hungary, May 2005.

[26] D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew, “On the removal of shadows from images,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 59–68, 2006.

[27] C. Fredembach and G. D. Finlayson, “Hamiltonian path based shadow removal,” in *Proc. British Machine Vision Conference*, 2005, pp. 970–980.

[28] N. Paragios and V. Ramesh, “A MRF-based real-time approach for subway monitoring,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 1034–1040.

[29] E. Salvador, A. Cavallaro, and T. Ebrahimi, “Cast shadow segmentation using invariant color features,” *Computer Vision and Image Understanding*, no. 2, pp. 238–259, 2004.

[30] F. Porikli and J. Thornton, “Shadow flow: a recursive method to learn moving cast shadows,” in *Proc. IEEE International Conference on Computer Vision*, vol. 1, 2005, pp. 891–898.

[31] N. Martel-Brisson and A. Zaccarin, “Moving cast shadow detection from a Gaussian mixture shadow model,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, June 2005, pp. 643–648.

[32] Y. Haeghen, J. Naeyaert, I. Lemahieu, and W. Philips, “An imaging system with calibrated color image acquisition for use in dermatology,” *IEEE Transactions on Medical Imaging*, vol. 19, no. 7, pp. 722–730, 2000.

[33] M. G. A. Thomson, R. J. Paltridge, T. Yates, and S. Westland, “Color spaces for discrimination and categorization in natural scenes,” in *Proc. Congress of the International Colour Association*, June 2002, pp. 877–880.

[34] L. Li and M. Leung, “Integrating intensity and texture differences for robust change detection,” *IEEE Trans. on Image Processing*, vol. 11, no. 2, pp. 105–112, 2002.

[35] A. Prati, I. Mikic, M. M. Trivedi, and R. Cucchiara, “Detecting moving shadows: algorithms and evaluation,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 918–923, 2003.

[36] Z. Kato, J. Zerubia, and M. Berthod, “Satellite image classification using a modified metropolis dynamics,” in *Proc. International Conference on Acoustics, Speech and Signal Processing*, March 1992, pp. 573–576.

[37] R. Potts, “Some generalized order-disorder transformation,” in *Proceedings of the Cambridge Philosophical Society*, no. 48, 1952, p. 106.

[38] D. S. Lee, “Effective Gaussian mixture learning for video background subtraction,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 827–832, 2005.

[39] D. A. Forsyth, “A novel algorithm for color constancy,” *International Journal of Computer Vision*, vol. 5, no. 1, pp. 5–36, 1990.

[40] E. A. Khan and E. Reinhard, “Evaluation of color spaces for edge classification in outdoor scenes,” in *Proc. of International Conference on Image Processing*, vol. 3. Genoa, Italy: IEEE, Sept. 2005, pp. 952–955.

[41] W. Feller, *An introduction to probability theory and its applications*, 2nd ed. John Wiley & Sons, 1966, vol. 1.

[42] W. K. Pratt, *Digital Image Processing*, 2nd ed. John Wiley & Sons, 1991, no. ISBN 0-471-85766-1.

[43] R. Haralick, “Digital step edges from zero crossing of second directional derivatives,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 6, no. 1, pp. 58–68, 1984.

[44] Y. Boykov, O. Veksler, and R. Zabih, “Fast approximate energy minimization via graph cuts,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[45] E. Aarts and J. Korst, *Simulated Annealing and Boltzman Machines*. New York: John Wiley & Sons, 1990.

[46] Y. Boykov and V. Kolmogorov, “An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004.

[47] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, “Equation of state calculations by fast computing machines,” *Journal of Chem. Physics*, vol. 21, pp. 1087–1092, 1953.

[48] J. Besag, “On the statistical analysis of dirty images,” *Journal of Royal Statistics Society*, vol. 48, pp. 259–302, 1986.

Csaba Benedek received the M.Sc. degree in computer sciences from the Budapest University of Technology and Economics in 2004. Currently, he is pursuing the Ph.D. degree at the Pázmány Péter Catholic University, Budapest. Meanwhile, he is member of the Distributed Events Analysis Research Group at the Computer and Automation Research Institute, Hungarian Academy of Sciences. As a visitor, he has recently worked with the ARIANA Project at INRIA Sophia-Antipolis, France. His research interests include Bayesian image segmentation, change detection, video surveillance and aerial image processing.

Tamás Szirányi received the Ph.D. degree in electronics and computer engineering in 1991 and the D.Sci. degree in 2001 from the Hungarian Academy of Sciences, Budapest. He was appointed to a Full Professor position an 2001 at Veszprém University, Hungary, and in 2004, at the Pázmány Péter Catholic University, Budapest. He is currently a scientific advisor at the Computer and Automation Research Institute, Hungarian Academy of Sciences, where he is the head of the Distributed Events Analysis Research Group. His research activities include texture and motion segmentation, surveillance systems for panoramic and multiple camera systems, measuring and testing the image quality, digital film restoration, Markov Random Fields and stochastic optimization, image rendering and coding.

Dr. Szirányi was the founder and first president (1997 to 2002) of the Hungarian Image Processing and Pattern Recognition Society. He is an Associate Editor of IEEE Transactions on Image Processing. He was honored with the Master Professor award in 2001.