MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Multiplicative Background-Foreground Estimation Under Uncontrolled Illumination using Intrinsic Images

Fatih Porikli

TR2005-012 June 2005

Abstract

Instead fo the conventional background and foreground definition, we propose a novel method that decomposes a scene into time-varying background and foreground intrinsic images. The multiplication of these images reconstructs the scene. First, we form a set of previous images into a temporal scale and compute their spatial gradients. By taking advantage of the sparseness of the filter outputs, we estimate the background by median filtering the gradients, and compute the corresponding foreground using the background. We also propose a robust method to threshold foregrounds to obtain a change detection mask of the moving pixels. We show that a different set of filters can detect the static and moving lines. Computationally, the proposed method is comparable with the state of the art, and our simulations prove the effectiveness of the intrinsic background foreground decomposition even under sudden and severe illumination changes.

Multiplicative Background-Foreground Estimation Under Uncontrolled Illumination using Intrinsic Images

Fatih Porikli
Mitsubishi Electric Research Laboratories
fatih@merl.com

Abstract

Instead of the conventional background and foreground definition, we propose a novel method that decomposes a scene into time-varying background and foreground intrinsic images. The multiplication of these images reconstructs the scene. First, we form a set of previous images into a temporal scale and compute their spatial gradients. By taking advantage of the sparseness of the filter outputs, we estimate the background by median filtering the gradients, and compute the corresponding foreground using the background. We also propose a robust method to threshold foregrounds to obtain a change detection mask of the moving pixels. We show that a different set of filters can detect the static and moving lines. Computationally, the proposed method is comparable with the state of the art, and our simulations prove the effectiveness of the intrinsic background/foreground decomposition even under sudden and severe illumination changes.

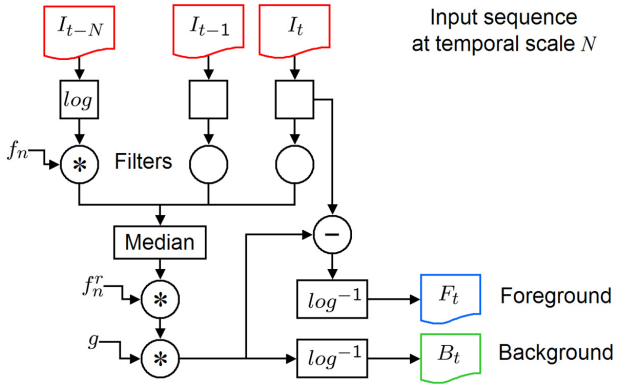

Figure 1: An intrinsic background that is independent from moving objects and static cast shadows is estimated using a set of N previous images.

1. Introduction

Background/foreground detection is an essential component of the most video surveillance systems involving object detection and tracking. Such systems require both robustness against sudden illumination changes and computational feasibility at the same time. However, existing approaches either make strict assumptions on the composition of the scene, or fail to handle abrupt illumination changes (e.g. turning off a light source), or demand high computational power, which restrict them to be a part of a real-time tracking system.

Due to these shortcomings of the previous approaches, we propose a novel background/foreground decomposition method based on the idea of multiplicative intrinsic images. Having the observation that an image is the product of the characteristics of the scene that its depicts, Barrow and Tenenbaum [1] introduced the term "intrinsic image" to refer to a mid-level decomposition of an image as a product of two images.

We extend the idea of intrinsic images to include the motion characteristics of a scene by assuming that an image can also be decomposed into a multiplication of a static part and a dynamic part. We form a set of previous images into a temporal scale and compute the spatial gradients of these images. By taking advantage of the sparseness of the filter outputs, we estimate background as a median filtered gradients and compute the corresponding foreground. We also propose a robust method to threshold foregrounds to obtain a detection mask. Our method is computationally comparable with the state of the art and our results prove the effectiveness of the intrinsic background/foreground images even under abrupt and severe illumination changes. A flow diagram of the proposed algorithm is given in Fig. 1.

The rest of this paper is organized as follows. In the next section, we discuss previous work on intrinsic images and background generation. In section 3, we explain the computation of the multiplicative background and foreground images. In section 4, we present simulation results and discuss the performance of the proposed method under different temporal scales and illumination conditions.

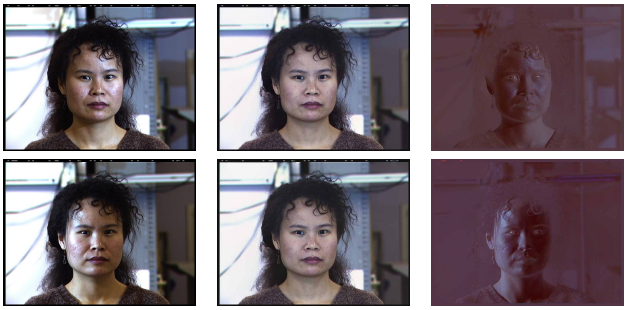

Figure 2: Samples of intrinsic reflectance and illumination decomposition computed by 24 images under different lighting conditions from the CMU face database (RGB color space is used).

2. Related Work

2.1. Intrinsic Images

Weiss [18] developed a maximum likelihood (ML) estimation framework to estimate a single reflectance image and multiple illumination images from a fixed view point image sequence that has under significant lighting condition variation. Based on the Weiss's method, Matsushita *et al.* [11] extended Weiss's algorithm to derive time-varying reflectance images and corresponding illumination images from a sequence of images. They also proposed utilizing an eigenspace that captures the illumination variations. We show sample reflectance and corresponding illumination images computed using the CMU face database in Fig. 2.

Instead using a set of images, Finlayson *et al.* [3] recently devised a method for the recovery of an illumination invariant image, which is similar to the reflectance image, from a single color image. They assume the input image contains both non-shadowed surfaces and shadows cast on those surfaces. They calculate an angle for an "invariant direction" in a log-chromaticity space by minimizing the entropy of color distribution. Tappen *et al.* [16] proposed an algorithm that uses multiple cues to recover shading and reflectance intrinsic images from a single image. Using both color information and a classifier trained to recognize gray-scale patterns, each image derivative is classified as being caused by shading or a change in the surfaces reflectance. Deterministic approaches are proposed by Kilger [9] and *et al.* [10] that exploit gray level, local and static features. In statistical approaches, *et al.* [15] and *et al.* [6] proposed a non-parametric approaches independently that use color, global and dynamic features for enhancing object detection.

2.2. Change Detection

Existing background generation methods can be classified as either single-layer or multi-layer approaches.

Single-layer methods construct a model for the color distribution of each pixel based on the past observations. A simple approach assumes that the past observations fits into a certain function such as a uniform distribution, and estimates a mean value by averaging past values to determine the current value of the background. Often a variance score that indicates the conformity of the estimated mean is adapted as the other higher order statistics. The variance score is also employed as a threshold to determine a set of foreground pixels that are inconsistent to the background model.

In [19], a single Gaussian is considered to model the statistical distribution of a background pixel and alpha-blending is used. The background is updated with the current frame according to a preset weight α such as $B_t = (1 - \alpha)B_{t-1} + \alpha I_t$. The α parameter acts as a learning factor; it adjusts how fast the background should be blended to the new frame. Although it is computationally preferable, this method is sensitive to the selection of the learning factor. Depending to the value of α, either the foreground objects may prematurely blended into the background (Fig. 3), or the model becomes unresponsive to the observations. Preset models and moving average operations generally cause so called ghost regions in the background that neither have the true background color nor the foreground object color. Regardless of a pixel belongs to whether foreground or background, this approach blends the current observation in the background model. Although the contamination of the background may be improved by increasing the number of frames, i.e. number of observations, it also limits the adaptability of the model to the illumination changes.

An alternative solution is the selection of the color values that are statistically more frequent. The voting method has advantages over the single model approach; it does not blur the background and it allows adaptation of the sudden changes depending on the color statistics. The major drawback of voting mechanism is its computational load.

The Kalman filter has been extensively used in background adaptation [8, 14, 17]. A version of the Kalman filter that operates directly on the data subspace is presented in [20]. In [13], a similar autoregressive model was proposed to capture the properties of dynamic scenes. The Kalman filter provides estimates for the state of a discrete-time process that obeys the linear stochastic difference equation. In our case, the state corresponds to the color of the background. The various parameters of the filter such as the transition matrix, the process noise covariance and the measurement noise covariance may change at each time step but are generally assumed to be constant. By using larger

covariance values, the background adapt quicker to the illumination changes, however, it becomes more sensitive to the noise and moving objects in the scene. Another drawback of the Kalman filter is its inability to represent multiple modalities, i.e. a background region depicts a swaying tree.

Stauffer and Grimson suggested to model the background with a mixture of Gaussian models [4]. Rather than explicitly modeling the values of all the pixels as one particular type of distribution, the background is constructed by a pixel-wise mixture of Gaussian distributions to support multiple backgrounds. Based on the persistence and the variance of each of the Gaussians, a mixture background is determined. Current observations that do not fit the background distributions are considered foreground until there is a Gaussian that includes them with sufficient evidence supporting it. Stauffer's background update method make use of an expectation maximization (EM) based framework, and contains two significant parameters; a learning constant and a parameter that controls the proportion of the data that should be accounted for by the background. The mixture of Gaussians method is the basis for a large number of related techniques [5]. A non-parametric approach was used in [2] where the use of Gaussian kernels for modeling the density at a particular pixel was proposed. Mittal and Pragios [12] integrated optical flow in the modeling of the dynamic characteristics.

The mixture methods are adaptable to illumination changes and they do not cause ghost regions. Furthermore, they can handle multiple backgrounds. However, one can claim that their performance deteriorates when the scene to be described is dynamic and exhibits non-stationary properties in time as illustrated in Fig. 3. Besides, a true "mixture" proposes blending of different models, which eventually comprise incorrect information for regardless of current observation fits or not. Another drawback of this multiple model solutions is the computational load of constructing and maintaining the background models. For the higher number of models in the mixture, such methods become computationally very demanding to be practical.

3. Intrinsic Background

A scene can be described as a composition of a static reflectance and a varying illumination field as stated in [18];

$$I_t = R \cdot L_t \tag{1}$$

where R contain the reflectance values of the scene, while the time dependent L_t contains the illumination intensities. This formulation is equivalent to $i_t = r + l_t$ in the log domain (We denote variables in log domain in lower-case). Since illumination images, L_t, represent the distribution of the incident light onto the scene while reflectance image, R, depict the surface reflectance properties of the scene, this representation becomes useful to analyze and manipulate the reflectance-lighting properties of the captured scene.

We propose a similar decomposition that enables evaluation of the motion properties and detection of moving objects in the scene is the foreground-background description

$$I_t = B_t \cdot C_t \tag{2}$$

where the backgrounds B_t represent the static and the changing foregrounds C_t stand for the relatively dynamic constituents of the scene. In log domain this relation becomes $i_t = b_t + c_t$. Note that this is different from the common background definition which states the relation as a cumulative inference, i.e. foreground is a residual of the observation when the background is removed, or in terms of layers of colors values i.e. background and foreground masks.

For a real world scene, the static constituent of a scene, i.e. color of a building, as well as its the dynamic constituent, i.e. moving objects, changes with time. A typical example is the variation of day light as a result of the interference of clouds in an outdoors setup and switching on and off the lighting sources in an indoors setup. While a time invariant background image B may reasonably describe the static scene texture without including moving objects, the estimated foreground images C_t tend to contain considerable amount of scene texture and shadows especially due to the lighting in such scenarios. Therefore, unlike the Weiss's method that implicitly assumes the reflectance image has to be independent from the illumination changes, we consider time-varying intrinsic images as it is first posed in Matsushita's work.

First, we select a support set of N images $\{I_{t-N}, .., I_{t-1}, I_t\}$ from the input sequence. These N previous images may be selected depending on the motion characteristics of the objects if it is known as a priori information. It is easy to see that the minimization of the overlapping regions between a moving object's appearances within the consecutive images will decrease the contamination of the static and changing constituents in the decomposition since our method uses the temporal median operators. Thus, we construct our support set as $\{I_{t-kN}, .., I_{t-k}, I_t\}$ where k is the sampling period that can be adjusted depending on the motion characteristics. In general, assignment of larger sampling periods provides sufficiently discriminating support set images, however, the sampling period can be set to a small value in case the objects move relatively faster with respect to the chosen frame rate. We observed that the decomposition process is not highly sensitive to the value of the sampling period k; we set $k = 100$ and obtained satisfactory results for the test sequences in our experiments. It should also noted that the support set images can be sampled at variable sampling rates, i.e., k change depending on the average motion.

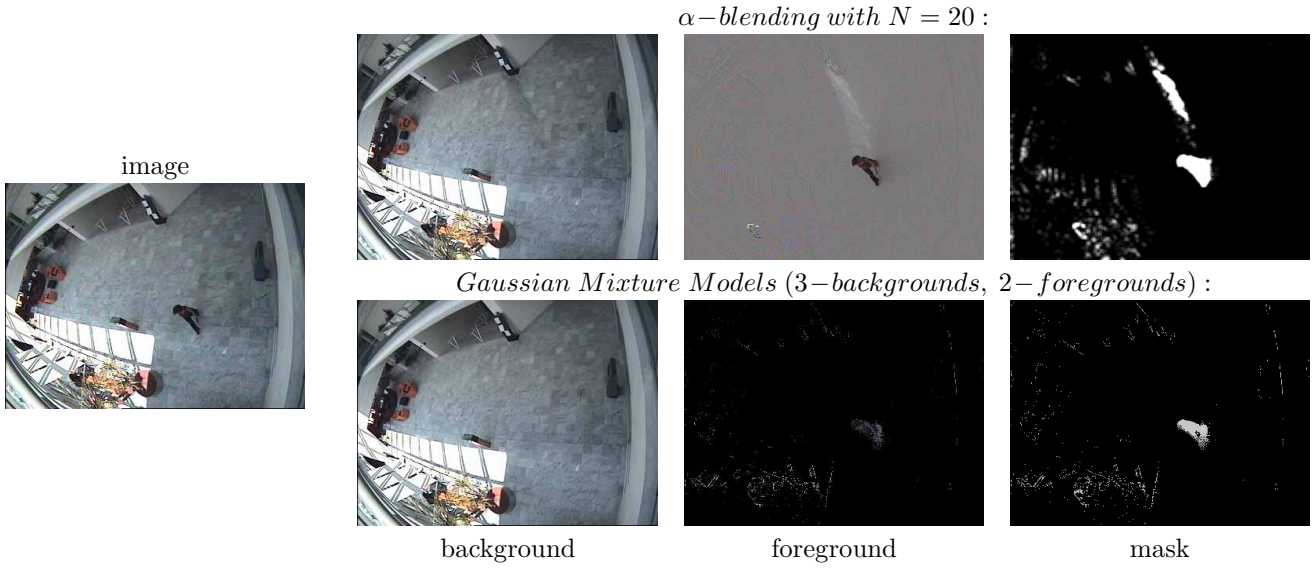

Figure 3: Detection results for single and multi layer approaches. As expected, α-blending contaminates the background severely. The multi-layer approach improves the contamination, however it deletes the stationary objects as in the lower left part of the scene, and it is sensitive to the image noise that is induced by compression and camera jitter.

Then, we apply spatial derivative filters f_n to the images to compute the intensity gradients $f_n * i_t$ where $*$ represents the convolution operator. Since the filter outputs that are applied to c_t are Laplacian distributed and independent over the space and time, the ML estimate of the static constituent in the transform domain $\hat{b}_{tn}$ is obtained as

$$\hat{b}_{tn} = \text{median}_t\{f_n * i_t\}, \tag{3}$$

which is a result of the fact that when the derivative filters are applied to the natural images, the filter outputs tend to be sparse [6],[7]. We impose two derivative filters $f_0 = [1 -1]$, $f_1 = [1 -1]^T$. The dynamic constituent in the transform domain c_{tn} are then computed by using the estimated static constituents $\hat{b}_{tn}$ as

$$\hat{c}_{tn} = (f_n * i_t) - \hat{r}_{tn} \tag{4}$$

Finally, the time-varying background images b_t and foreground images c_t is recovered by solving the systems of the following linear equations [18]

$$\hat{b}_t = g * \left(\sum_n f_n^r * \hat{b}_{tn} \right) \tag{5}$$

$$\hat{c}_t = g * \left(\sum_n f_n^r * \hat{c}_{tn} \right) \tag{6}$$

where f_n^r is the reversed filter of f_n, and g is a filter which satisfies the following equation in the transform domain

$$G = (F_n \cdot F_n^r)^{-1} \tag{7}$$

where the transform domain is obtained by the Fourier transformation. The filter g is independent of the image sequence, thus it can be computed in advance. The final backgrounds and foregrounds are calculated by taking the inverse logarithm; $B_t = e^{\hat{b}_t}, C_t = e^{\hat{c}_t}$.

We define a binary change detection mask image M_t that corresponds to the foreground pixels in the current image I_t. We set a varying threshold using the variance of the difference between the current background and foreground images $D_t(x, y) = B_t(x, y) - C_t(x, y)$. It can be shown that the distribution of the difference forms a Gaussian function. Thus, we compute the mean μ_t and variance σ_t^2 of the difference, and assign the 95% percentile as the threshold $\tau = 2.5\sigma$, which is based on the assumption that pixels having high temporal gradients are significantly less than the static pixels. In other words, moving and cast shadows are not counted as moving object pixels since at such pixels the temporal gradient should be minimum for a static camera setup. The change detection mask is obtained by

$$M_t(x, y) = \begin{cases} 1 & |D(x, y) - \mu_t| > 2.5\sigma_t \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Figure 4 shows the estimated backgrounds, foregrounds, and masks for a traffic scene.

4. Experiments

We tested the intrinsic background/foreground method with several benchmark sequences. Figures 4, 5, 6 show the

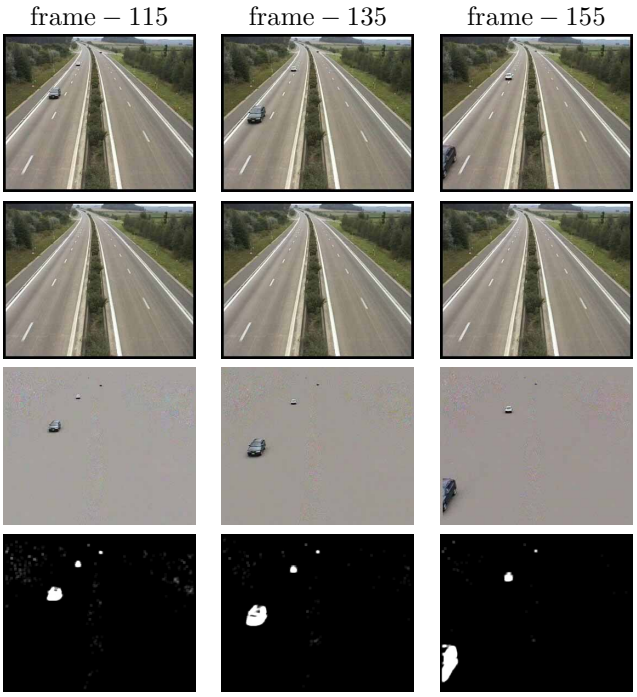

Figure 4: Intrinsic backgrounds (2^{nd}-row), foregrounds (3^{rd}-row), and detection masks (4^{th}-row) for samples from a traffic scene.

estimation results for the *traffic* (generic), *browse3* (PETS 2004), and *green* (MPEG-7) sequences. We used $N = 15$ previous frames that are sampled at $k = 100$ to form the support sets. In these figures, the second row shows the estimated background images B_t, the third row corresponds to the estimated foreground images C_t, and the last row gives the detection masks M_t for the corresponding original images given in the first row. The threshold $\tau = 2.5\sigma$ kept same for all sequences. In Fig. 4 the scene contains moving bushes and clouds in the distance and diffuse shadows under the vehicles. Note that the foreground images are accurate and do not contain the mentioned residues. It is also seen that the diffuse shadows are eliminated in the masking operation. The sequence given in Fig. 5 has camera jitter and encoding artifacts, which has been amplified by the GMM mixture method as demonstrated in Fig. 3. As visible, the results of intrinsic method are significantly less noisy. Figure 6 presents another scene where the color of the sky changes and shadows become more eminent with time. We observed that the proposed method accurately decomposed the given sequences into background and foreground constituents in each case.

To test the sudden severe illumination change response, we simulated a severe brightness reduction by suppressing the color values of each channel. The scene abruptly becomes darker at frame 443 where the average intensity decreases almost 40%, which is an similar to the turning lights off in an indoors setting.

Table 1: Average Processing Times

α- blending (N=10)	11ms
GMM (3 models)	55ms
GMM (5 models)	97ms
Intrinsic Background (N=15)	53ms

Our simulations show that only the intrinsic background/foreground method can accommodate such a sudden and severe illumination change as shown in Fig. 8. The α-blending was unable recover the background since the pixel confidences values were significantly high before the brightness change (lower confidence values just melts moving objects in the background, thus is not preferable in an actual setting). A state of the art Gaussian mixture model with 3 background and 2 foreground models took more than 200 frames to adapt the change. We observed the proposed method not sensitive to the sampling period, and using higher values gives better estimation performances.

The computational cost of the proposed algorithm is also comparable with the existing methods. For a 320×240 color sequence (each channel is treated independently) on a 3Ghz P4 platform, the average processing time of the above algorithms are presented in the table 1.

We also tested different convolution filters. Instead of using the spatial derivation filters, we a applied line detector $f_{line} = [-1\ \ 2\ \ -1]$. The response of the line filters are sparse and has a Laplacian form, which enables us to adapt the proposed method. We show sample detection results of the line filters in Fig. 7. These filters capture the static and moving edges of the scene, which is a valuable information for shape extraction.

5. Summary and Conclusions

We propose a novel method to estimate the background (static regions, shadows cast by buildings, etc) and foreground (moving objects) of a sequence captured by a stationary camera. As opposed to the additive background/foreground models, we decompose a sequence into time-varying multiplicative backgrounds and foregrounds using the intrinsic image approach as presented.

There are several advantages of the proposed method: 1) It is robust to the sudden and severe illumination changes that a scene may undergo. 2) It is not restricted to the model based background assumptions, and it does not require fitting neither background nor foreground models. 3) It is computationally feasible to implement into a real-time system. 4) It is not sensitive to the fine-tuning of its parameters. 5) We show that it is also possible to estimate a

static edge map of a scene. 6) Our results show that the multiplicative background/foreground can recover background image and detect moving pixels accurately.

References

[1] H.G. Barrow and J.M. Tenenbaum, "Recovering intrinsic scene characteristics from images", *In Computer Vision Systems, Academic Press*, pp. 3.26, 1978

[2] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction", *In Proc. of European Conf. on Computer Vision*, pp. II:751767, 2000

[3] G.D. Finlayson, S.D. Hordley and M.S. Drew, "Removing Shadows from Images", *In Proc. of European Conf. on Computer Vision* Vol.4, pp. 823.836, 2002

[4] C. Stauffer and W.Grimson, "Adaptive background mixture models for real-time tracking", *In Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 1999

[5] O. Javed, K. Shafique, and M. Shah, "A hierarchical approach to robust background subtraction using color and gradient information", *In MVC*, pp. 2227, 2002

[6] C. Jiang and M.O. Ward, "Shadow identification", *In Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, pp. 606.612, 1992

[7] J. Huang and D. Mumford, "Statistics of natural images and models", *In Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, pp. 541.547, 1999

[8] K. Karmann, A. Brand, "Time-varying image processing and moving object recognition", *Elsevier Science Publish.*, 1990

[9] M. Kilger, "A shadow handler in a video-based real-time traffic monitoring system", *In Proc. of IEEE Workshop on Applications of Computer Vision*, pp. 11.18, 1992

[10] D. Koller, K. Danilidis, and H. Nagel, "Model-based object tracking in monocular image sequences of road traffic scenes", *In Int'l Journal of Computer Vision*, vol. 10, pp. 257.281, 1993

[11] Y. Matsushita, K. Nishino, K. Ikeuchi, and S. Masao, "Illumination normalization with time-dependent intrinsic images for video surveillance", *In Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2004

[12] A. Mittal and N. Paragios, "Motion-Based background subtraction using adaptive kernel density estimation", *In Proc. Int'l Conf. on Computer Vision and Pattern Recognition*, 2004

[13] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh, "Background modeling and subtraction of dynamic scenes", *In Proc. of IEEE Int'l Conf. on Computer Vision*, pp. 13051312, 2003

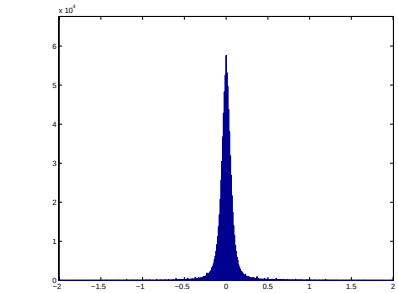

Figure 7: Histogram of the line filter $f_{line} = [-1\ 2\ -1]$ outputs and the corresponding results of the proposed method. The histogram has a Laplacian form. The intrinsic background shows the static texture in the scene (bright lines), and the foreground corresponds to the moving texture, which is a valuable information for shape extraction.

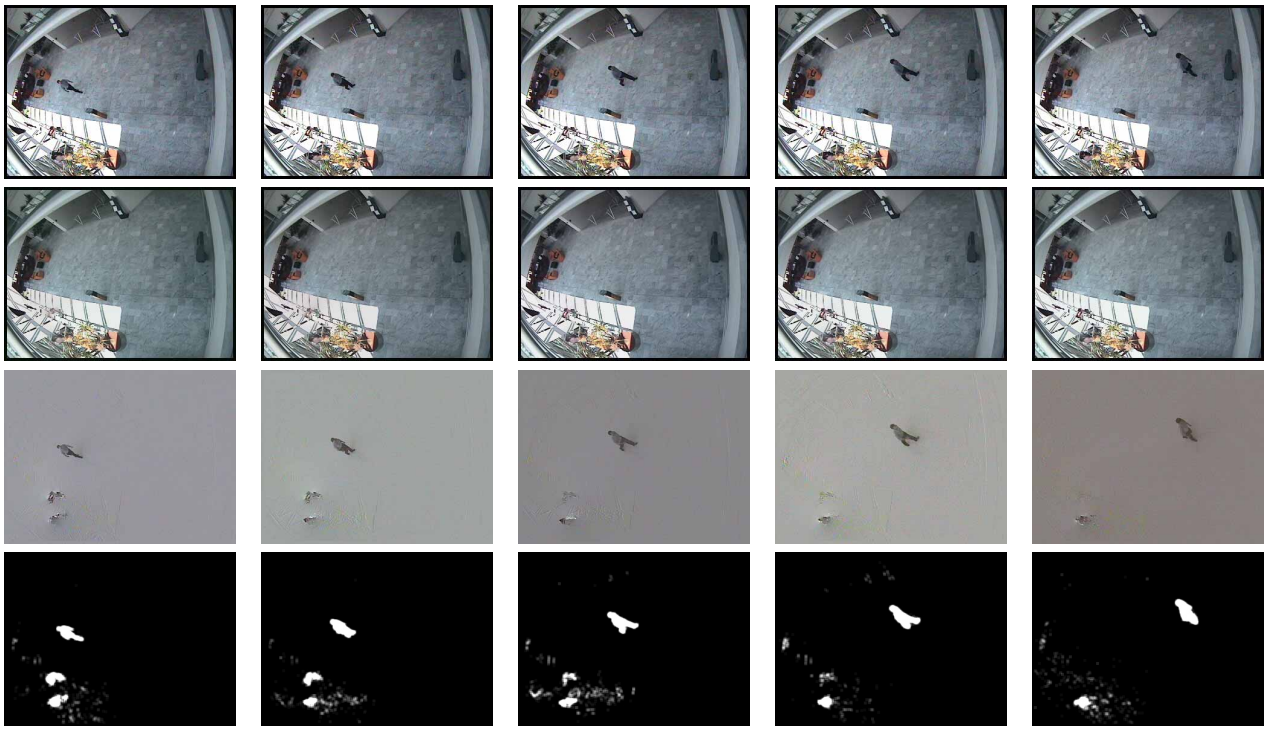

Figure 5: Results from the PETS-2004 *browse-3* sequence: intrinsic backgrounds (2^{nd}-row), foregrounds (3^{rd}-row), and detection masks (4^{th}-row). There are moving people in the lower left of the scene. ($N = 15, k = 10, \tau = 2.5\sigma$)

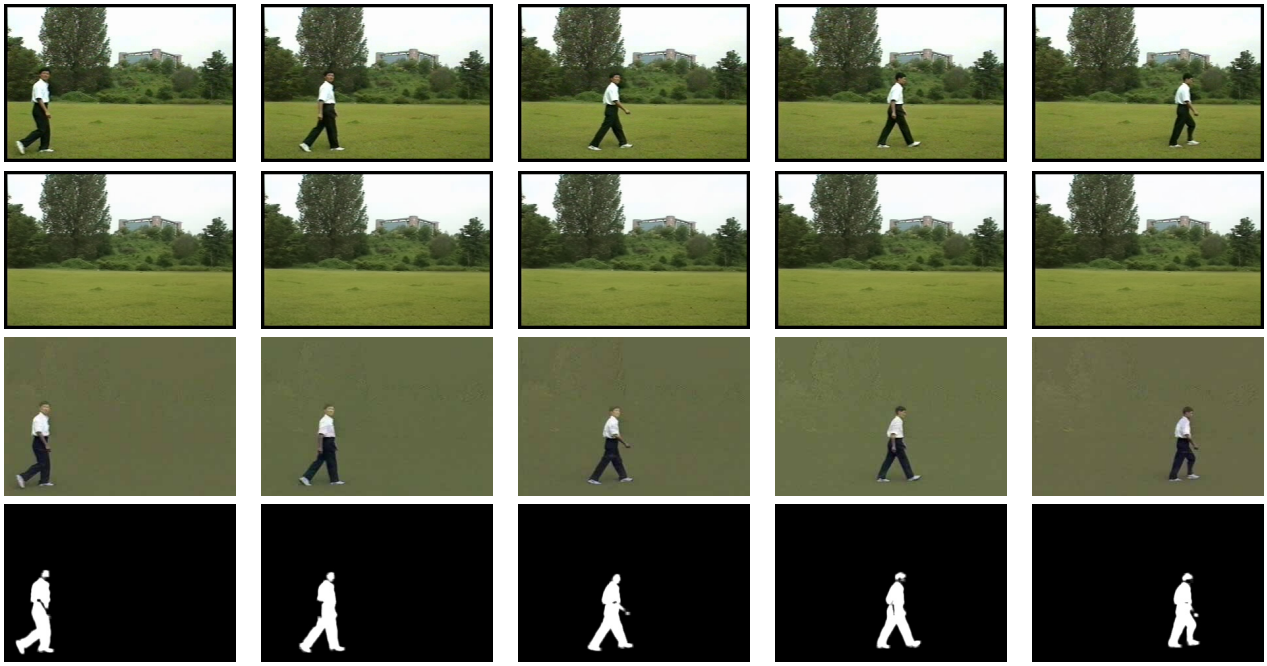

Figure 6: Results from an MPEG-7 sequence: intrinsic backgrounds (2^{nd}-row), foregrounds (3^{rd}-row), and detection masks (4^{th}-row). As visible, the foregrounds and detection masks are accurately extracted. ($N = 15, k = 10, \tau = 2.5\sigma$)

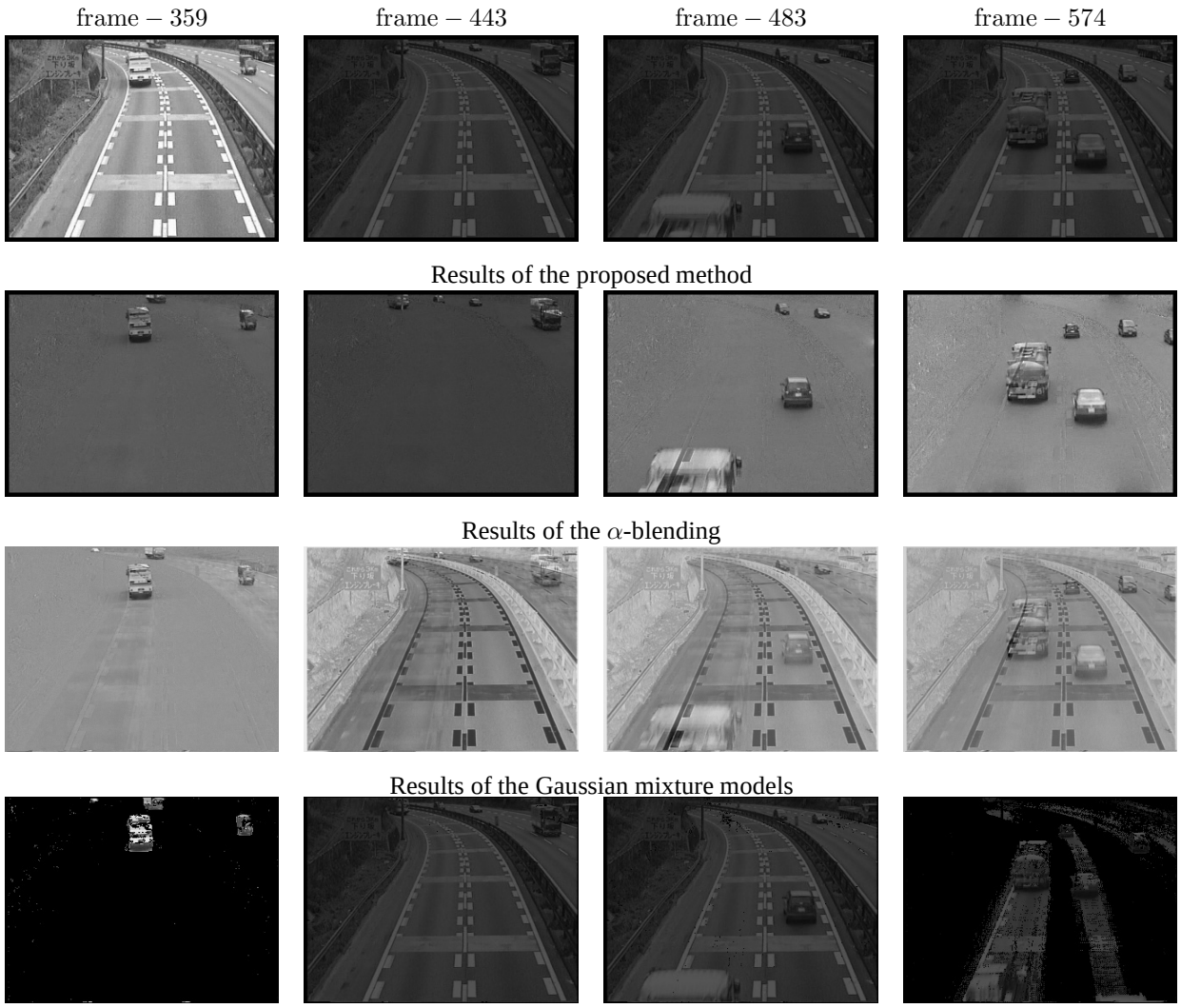

Figure 8: Only the proposed method can accommodate a sudden and severe illumination change as demonstrated in the above sequence. At frame-443, the camera parameters are distorted and scene is captured at a much lower brightness level, which is similar to turning the lights off in an indoors setting. Each row shows the corresponding foregrounds images.

[14] C. Ridder, O. Munkelt, and H. Kirchner, "Adaptive background estimation and foreground detection using Kalman filtering", *In Proc. ICAM*, 1995

[15] J. Stauder, R. Mech, and J. Ostermann, "Detection of moving cast shadows for object segmentation", *In IEEE Transactions on Multimedia*, vol. 1, no. 1, pp. 65.76, Mar. 1999

[16] M. Tappen, W. Freeman, E. Adelson, "Recovering Shading and Reflectance from a single image", In NIPS, 2002

[17] K. Toyama, J. Krumm, B. Brumitt, B. Meyers, "Wallflower: Principles and Practice of Background Maintenance", *In Proc. of Int'l Conf. on Computer Vision*, pp. 255.261, 1999

[18] Y. Weiss, "Deriving intrinsic images from image sequences", *In Proc. of IEEE Int'l Conf. on Computer Vision*, pp. 68.75, Jul., 2001

[19] C.R. Wren, A. Azarbayejani, T.J. Darrell, and A.P. Pentland "Pfinder: Real-time tracking of the human body", *In PAMI*, 19(7):780785, July 1997

[20] J. Zhong and S. Sclaroff, "Segmenting foreground objects from a dynamic, textured background via a robust kalman filter", *In Proc. of IEEE Int'l Conf. on Computer Vision*, pp. 4450, 2003