A PROBABILISTIC METHOD FOR FOREGROUND AND SHADOW SEGMENTATION

Yang Wang[1], *Tele Tan*[2], *and Kia-Fock Loe*[3]

[1,2]Institute for Infocomm Research, Singapore 119613, {ywang[1], teletan[2]}@i2r.a-star.edu.sg.
[3]Department of Computer Science, National University of Singapore, Singapore 117543, loekf@comp.nus.edu.sg.

ABSTRACT

This paper presents a probabilistic method for foreground segmentation that distinguishes moving objects from their cast shadows in monocular indoor image sequences. The models of background, shadow, and edge information are set up and adaptively updated. A Bayesian framework is proposed to describe the relationships among the segmentation label, background, intensity, and edge information. A Markov random field is used to boost the spatial connectivity of the segmented regions. The solution is obtained by maximizing the posterior probability density of the segmentation field.

1. INTRODUCTION

Detecting dynamic objects in image sequences is very important in such areas as surveillance and object-based coding. Accurate and efficient background removal is critical in these systems. Background subtraction based on intensity or color is a commonly used technique to identify foreground elements. The background model is built from the data and objects are segmented if they appear significantly different from the background.

To deal with illumination or object changes in the background, many researchers have abandoned nonadaptive methods of backgrounding. Friedman and Russell [2] classify each pixel by a probabilistic model of how that pixel looks when it is part of different classes and use an incremental EM algorithm to learn the pixel model. Stauffer and Grimson [6] model each pixel as a mixture of Gaussians and update the model in an adaptive way. The Gaussian distributions are then evaluated to determine which are most likely to result from a background process.

Besides the nonstationariness of the background, camouflage and shadow are two classic problems of subtraction. If regions of the foreground have similar colors as the background, they can be erroneously removed. In addition, shadows cast on the background can be erroneously labeled as foreground. For monocular color video sequences, false segmentation caused by shadows can be minimized by computing differences in a color space that is less sensitive to intensity change [8]. Moreover, edge information can be utilized to improve the quality and reliability of the results [3]. Stauder et al. [5] assumes that static edges caused by the background texture remain in regions covered by shadows and that penumbras exist at the boundary of shadows. However, this is sometimes not true due to the properties of the imaging process [4]. Mikic et al. instead approximate the change of the camera response for the shadowed region by a diagonal matrix.

A unified framework of foreground segmentation, which combines the background, intensity, and edge information, is proposed in this paper. A generalized model is built for the appearance change under shadow. A Markov random field is employed to encourage the formation of continuous segmentation regions. The solution is obtained by maximizing the posterior probability density of the segmentation field. Experiments show that our method greatly improves the accuracy of segmentation.

2. MODEL

Given the image sequence $\{g_k\}$ ($k \in \mathbf{N}$), $s_k(\mathbf{x})$ is denoted as the segmentation label for a point $\mathbf{x}$ within the image g_k at time k. $s_k(\mathbf{x})$ equals 1 for a background pixel, 2 for shadow, and 3 for foreground. Static shadows are considered to be part of the background. The entire segmentation field is expressed compactly as s_k.

In order to segment the foreground regions in a video sequence, the system should first model the background and shadow in the video scene. The edge information also helps detect the changes in the scene.

2.1. Background

Each pixel of an image acquired by the camera contains noise components. Assume that independent Gaussian noise corrupts each pixel in the scene, so that the observation model for the background becomes

$$g_k(\mathbf{x}) = \mu_{b,k}(\mathbf{x}) + n_k(\mathbf{x}), \text{ if } s_k(\mathbf{x}) = 1, \quad (1)$$

where $\mu_{b,k}(\mathbf{x})$ is the intensity mean of a single pixel $\mathbf{x}$ within the background, and $n_k(\mathbf{x})$ is the independent zero-mean additive noise with variance $\sigma^2_{b,k}(\mathbf{x})$ at time k. The parameter vector $(\mu_{b,k}(\mathbf{x}), \sigma^2_{b,k}(\mathbf{x}))^T$ is denoted as $\boldsymbol{\theta}_{b,k}(\mathbf{x})$, and the entire background is expressed as $\boldsymbol{\theta}_{b,k}$.

0-7803-7750-8/03/$17.00 ©2003 IEEE

For a static background, a sequence of background images of the scene could be recorded and the intensity mean and variance of each pixel can be calculated.

For a nonstationary background, the update method is based on the idea of Stauffer and Grimson [6]. The recent history of each pixel, $\{g_i(\mathbf{x})\}_{1\le i\le k}$, is modeled as a mixture of Gaussians. Each time the Gaussian with the most supporting evidence and the least variance is chosen as the background model.

2.2. Shadow

Given the intensity mean of a background point, we use a linear transformation to describe the change of intensity for the same point when shadowed in the video frame at time k.

$$g_k(\mathbf{x}) = \mu_{s,k}(\mathbf{x}) + n_k(\mathbf{x}), \text{ if } s_k(\mathbf{x}) = 2, \quad (2a)$$

$$\mu_{s,k}(\mathbf{x}) = a_k \mu_{b,k}(\mathbf{x}) + c_k. \quad (2b)$$

When a_k equals 1, the edge information will not change if the area is shadowed by the foreground. Moreover, if the image input is multi-channel (R, G, B), the chromaticity will remain unchanged under such a linear transformation when c_k is zero. Therefore, the shadow model can be viewed as the generalization of the previous assumptions. With this model for the appearance change, we can easily derive the rules for estimating means and variances for the points under shadow.

2.3. Edge

For the kth frame g_k, $\mathbf{e}_{g,k}(\mathbf{x})$ is denoted as the frame edge vector at site $\mathbf{x}$. $\mathbf{e}_{g,k}(\mathbf{x}) = (e^h_{g,k}(\mathbf{x}), e^v_{g,k}(\mathbf{x}))^T$, where $e^h_{g,k}(\mathbf{x})$ is the intensity difference of the two horizontally neighboring points within the frame, and $e^v_{g,k}(\mathbf{x})$ is the vertical difference. The entire filed is expressed as $\mathbf{e}_{g,k}$.

Similarly, we can define the background edge vector $\mathbf{e}_{b,k}(\mathbf{x})$ at site $\mathbf{x}$, $\mathbf{e}_{b,k}(\mathbf{x}) = (e^h_{b,k}(\mathbf{x}), e^v_{b,k}(\mathbf{x}))^T$. From the background model, we know that $\mathbf{e}_{b,k}(\mathbf{x})$ is of bivariate normal distribution with mean $\boldsymbol{\mu}_{e,k}(\mathbf{x})$ and covariance matrix $\boldsymbol{\Sigma}_{e,k}(\mathbf{x})$ for each site $\mathbf{x}$. $\boldsymbol{\mu}_{e,k}(\mathbf{x})$ and $\boldsymbol{\Sigma}_{e,k}(\mathbf{x})$ are determined by the intensity means and variances of the four neighboring background points. The parameter vector $(\boldsymbol{\mu}_{e,k}(\mathbf{x}), \boldsymbol{\Sigma}_{e,k}(\mathbf{x}))^T$ is denoted as $\boldsymbol{\theta}_{e,k}(\mathbf{x})$, and the entire field at time k is expressed as $\boldsymbol{\theta}_{e,k}$. The background edge information $\boldsymbol{\theta}_{e,k}$ can be calculated from $\boldsymbol{\theta}_{b,k}$. The edge model can be used to locate changes in the structure of the scenes as edges appear, disappear, or change direction.

3. ALGORITHM

Given the current frame g_k, frame edge field $\mathbf{e}_{g,k}$, background $\boldsymbol{\theta}_{b,k}$, and background edge field $\boldsymbol{\theta}_{e,k}$, we wish to compute the maximum *a posteriori* (MAP) estimation of the segmentation field s_k. Naturally the background (or background edge) information is independent on the label field. Using the Bayes' rule and ignoring the constants with respect to the unknowns,

$$\begin{aligned}\hat{s}_k &= \arg\max_{s_k} p(s_k \mid \boldsymbol{\theta}_{b,k}, \boldsymbol{\theta}_{e,k}, g_k, \mathbf{e}_{g,k}) \\ &= \arg\max_{s_k} p(g_k, \mathbf{e}_{g,k} \mid s_k, \boldsymbol{\theta}_{b,k}, \boldsymbol{\theta}_{e,k})\, p(s_k, \boldsymbol{\theta}_{b,k}, \boldsymbol{\theta}_{e,k}) \\ &= \arg\max_{s_k} p(g_k, \mathbf{e}_{g,k} \mid \boldsymbol{\theta}_{b,k}, \boldsymbol{\theta}_{e,k}, s_k)\, p(s_k). \quad (3)\end{aligned}$$

The likelihood model $p(g_k, \mathbf{e}_{g,k} \mid \boldsymbol{\theta}_{b,k}, \boldsymbol{\theta}_{e,k}, s_k)$ and the prior model $p(s_k)$ should be defined for the video sequence.

3.1. Likelihood

Assuming conditional independence between spatially distinct observations, we factorize the likelihood model as

$$\begin{aligned}&p(g_k, \mathbf{e}_{g,k} \mid \boldsymbol{\theta}_{b,k}, \boldsymbol{\theta}_{e,k}, s_k) \\ &= \prod_{\mathbf{x}\in\mathbf{X}} p(g_k(\mathbf{x}), \mathbf{e}_{g,k}(\mathbf{x}) \mid \boldsymbol{\theta}_{b,k}(\mathbf{x}), \boldsymbol{\theta}_{e,k}(\mathbf{x}), s_k(\mathbf{x})), \quad (4)\end{aligned}$$

where $\mathbf{X}$ is the spatial domain of the video scene. Given the segmentation label, background, and background edge information, we further assume that the image intensity and image edge are conditionally independent on each other at each site. Thus, the likelihood can be factorized as the product of the intensity likelihood and edge likelihood.

$$\begin{aligned}&p(g_k(\mathbf{x}), \mathbf{e}_{g,k}(\mathbf{x}) \mid \boldsymbol{\theta}_{b,k}(\mathbf{x}), \boldsymbol{\theta}_{e,k}(\mathbf{x}), s_k(\mathbf{x})) \\ &= p(g_k(\mathbf{x}) \mid \boldsymbol{\theta}_{b,k}(\mathbf{x}), s_k(\mathbf{x}))\, p(\mathbf{e}_{g,k}(\mathbf{x}) \mid \boldsymbol{\theta}_{e,k}(\mathbf{x}), s_k(\mathbf{x})). \quad (5)\end{aligned}$$

When site $\mathbf{x}$ is in the background, we can calculate the intensity likelihood using the background model.

$$\begin{aligned}&p(g_k(\mathbf{x}) \mid \boldsymbol{\theta}_{b,k}(\mathbf{x}), s_k(\mathbf{x}) = 1) \\ &= N(g_k(\mathbf{x}); \mu_{b,k}(\mathbf{x}), \sigma^2_{b,k}(\mathbf{x})), \quad (6)\end{aligned}$$

where $N(\mathbf{z}; \mathbf{m}, \boldsymbol{\Sigma})$ is a Gaussian density with argument $\mathbf{z}$, mean $\mathbf{m}$, and covariance $\boldsymbol{\Sigma}$.

When site $\mathbf{x}$ is shadowed, the density can be calculated by the shadow model.

$$\begin{aligned}&p(g_k(\mathbf{x}) \mid \boldsymbol{\theta}_{b,k}(\mathbf{x}), s_k(\mathbf{x}) = 2) \\ &= N(g_k(\mathbf{x}); a_k \mu_{b,k}(\mathbf{x}) + c_k, \sigma^2_{b,k}(\mathbf{x})). \quad (7)\end{aligned}$$

When site $\mathbf{x}$ is in the foreground, the background has no influence on the image intensity information. Uniform distribution is assumed for the pixel. The conditional probability density becomes

$$\begin{aligned}&p(g_k(\mathbf{x}) \mid \boldsymbol{\theta}_{b,k}(\mathbf{x}), s_k(\mathbf{x}) = 3) \\ &= p(g_k(\mathbf{x}) \mid s_k(\mathbf{x}) = 3) = \frac{1}{y_{\max}}. \quad (8)\end{aligned}$$

Here $[0, y_{\max}]$ is the intensity range for every point $\mathbf{x}$ in the scene.

For each point $\mathbf{x}$, denote the set of its four nearest neighboring points by $M_{\mathbf{x}}$. Considering the spatial connectivity of the image, we assume that the four neighboring points have the same segmentation labels. Thus the edge likelihood can be approximated by

$$p(\mathbf{e}_{g,k}(\mathbf{x}) \mid \boldsymbol{\theta}_{e,k}(\mathbf{x}), s_k(\mathbf{x}))$$

$$\approx p(\mathbf{e}_{g,k}(\mathbf{x}) \mid \boldsymbol{\theta}_{e,k}(\mathbf{x}), s_k(\mathbf{y}) = s_k(\mathbf{x}), \forall \mathbf{y} \in M_x). \quad (9)$$

Similarly, the edge likelihood can be derived from the models in section 2.

$$p(\mathbf{e}_{g,k}(\mathbf{x}) \mid \boldsymbol{\theta}_{e,k}(\mathbf{x}), s_k(\mathbf{x}) = 1)$$
$$\approx N(\mathbf{e}_{g,k}(\mathbf{x}); \boldsymbol{\mu}_{e,k}(\mathbf{x}), \boldsymbol{\Sigma}_{e,k}(\mathbf{x})). \quad (10)$$

$$p(\mathbf{e}_{g,k}(\mathbf{x}) \mid \boldsymbol{\theta}_{e,k}(\mathbf{x}), s_k(\mathbf{x}) = 2)$$
$$\approx N(\mathbf{e}_{g,k}(\mathbf{x}); a_k \boldsymbol{\mu}_{e,k}(\mathbf{x}), \boldsymbol{\Sigma}_{e,k}(\mathbf{x})). \quad (11)$$

$$p(\mathbf{e}_{g,k}(\mathbf{x}) \mid \boldsymbol{\theta}_{e,k}(\mathbf{x}), s_k(\mathbf{x}) = 3)$$
$$\approx \left(\frac{1}{y_{\max}} - \frac{|e^h_{g,k}(\mathbf{x})|}{y^2_{\max}}\right)\left(\frac{1}{y_{\max}} - \frac{|e^v_{g,k}(\mathbf{x})|}{y^2_{\max}}\right)$$
$$\approx N(e^h_{g,k}(\mathbf{x}); 0, \frac{y^2_{\max}}{6})\ N(e^v_{g,k}(\mathbf{x}); 0, \frac{y^2_{\max}}{6}). \quad (12)$$

The approximation in the last step of (12) is achieved by moment matching.

3.2. Prior

The prior model $p(s_k)$ represents the prior probability of the segmentation field. We model s_k as a Markov random field to form spatial constraints from the neighborhood. The distribution is given by a Gibbs density that has the following form [7]:

$$p(s_k) \propto \exp\{-\sum_{c \in C} V(s_k(\mathbf{x}) \mid \mathbf{x} \in c)\}, \quad (13)$$

where C is the set of all cliques c, and V is the clique potential function at time k. A clique is a set of points that are neighbors of each other. The clique potential depends only on the pixels that belong to clique c. Spatial connectivity is imposed by the following two-pixel clique potential.

$$V(s_k(\mathbf{x}), s_k(\mathbf{y})) = \frac{1}{\|\mathbf{x} - \mathbf{y}\|_\infty}(1 - \delta(s_k(\mathbf{x}) - s_k(\mathbf{y}))), \quad (14)$$

where $\delta(\cdot)$ is the Dirac delta function, and $\|\cdot\|_\infty$ denotes the max-norm. Thus two neighboring pixels are more likely to belong to the same class than to different classes. The constraint becomes stronger with decreasing distance between the neighboring sites.

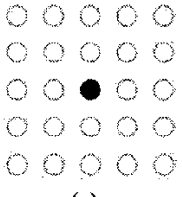
3.3. Optimization

Combining the above models, the Bayesian MAP estimate is obtained by minimizing the objective function.

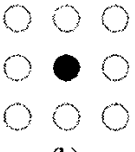
$$F_k(s_k) = \sum_{\mathbf{x} \in \mathbf{X}} U_{1,k}(\mathbf{x}, s_k(\mathbf{x})) + \sum_{\mathbf{x} \in \mathbf{X}} U_{2,k}(\mathbf{x}, s_k(\mathbf{x})) + \lambda \sum_{\{\mathbf{x},\mathbf{y}\} \in C} V(s_k(\mathbf{x}), s_k(\mathbf{y})), \quad (15)$$

where $U_{1,k}(\mathbf{x}, s_k(\mathbf{x})) = -\ln p(g_k(\mathbf{x}) \mid \boldsymbol{\theta}_{b,k}(\mathbf{x}), s_k(\mathbf{x}))$, and $U_{2,k}(\mathbf{x}, s_k(\mathbf{x})) = -\ln p(\mathbf{e}_{g,k}(\mathbf{x}) \mid \boldsymbol{\theta}_{e,k}(\mathbf{x}), s_k(\mathbf{x}))$. The parameter λ and initial values of a_1 and c_1 are manually determined. λ reflects the importance of spatial connectivity. Each time after the segmentation of the kth frame, the linear transformation of the shadow model can be adaptively updated from the set $\{(g_k(\mathbf{x}), b_k(\mathbf{x})) \mid \hat{s}_k(\mathbf{x}) = 2\}$ when shadow area is detected. Moreover, the size of the neighborhood will affect the segmentation results (see figure 1 and section 4).

The objective function does not have a unique minimum since it is nonconvex in terms of $s_k(\mathbf{x})$. Obviously, there is no simple method of performing the optimization. To arrive at a sub-optimal estimate, we use a local technique known as iterated conditional modes (ICM) algorithm [1]. The ICM scheme employs the greedy strategy in the iterative minimization. Given the observed data and the other labels, the algorithm sequentially updates the label by locally minimizing the objective function at each site.

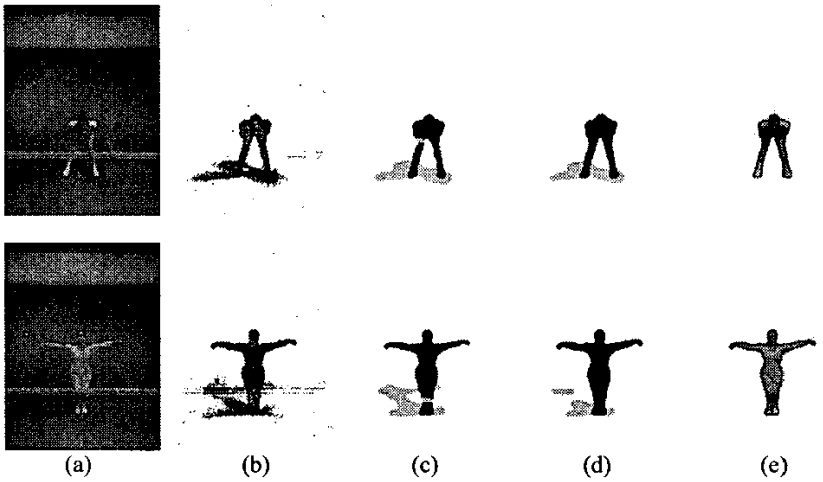


(a)



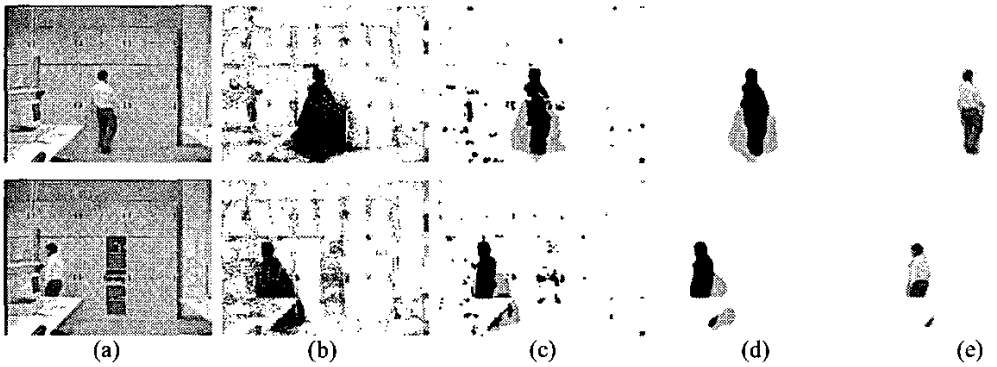
(b)

Fig. 1. Neighborhood systems



(a) (b) (c) (d) (e)

Fig. 2. Segmentation results of the "aerobic" sequence



(a) (b) (c) (d) (e)

Fig. 3. Segmentation results of the "laboratory" sequence

4. RESULTS AND DISCUSSION

Test results of two indoor sequences, the "aerobic" sequence and the "laboratory" sequence, are shown in figure 2 and 3, respectively. Figure 2a shows two frames of the "aerobic" sequence, 2b the segmentation result of the simple background subtraction method, 2c the result of our algorithm without using edge information, 2d the result of the proposed method, and 2e the foreground detected by our method. The gray regions indicate moving cast shadows. In figure 2c and 2d, the 24-pixel neighborhood is used. Figure 3a shows two frames of the "laboratory" sequence, 3b the result of simple background subtraction, 3c the result of the proposed method using the 8-pixel neighborhood, 3d the result of the proposed method using the 48-pixel neighborhood, and 3e the foreground detected by our method.

Comparing with the results of simple background subtraction, the accuracy of the calculated object location is greatly improved by the proposed approach. The moving cast shadows are exactly removed from the foreground. The flickering background pixels, which are detected as foreground by simple background subtraction, are correctly classified by our algorithm. The open cabinet in the second "laboratory" image is classified as background after a period of background updating.

The camouflage region separates the person into two parts in figure 2c. This problem is successfully overcome in figure 2d by incorporating the edge information. Figure 3c and 3d shows the influence of the neighborhood size. Strong spatial constraints should be employed when the noise in the scene is heavy.

5. CONCLUSION

In this paper we have presented a probabilistic approach for foreground segmentation and shadow detection in indoor image sequences. In our work, we could identify two sources of spatial information when detecting objects and shadows. The first is the edge information, the differences help locate changes in the scene. The second source of information is spatial connectivity, objects and shadows usually form continuous regions. On the other hand, the temporal information helps adaptively update the models from previous segmentation results.

Experimental results show that our method successfully deals with changing background, camouflage and shadows in video sequences. Moreover, the algorithm can be easily implemented for color image sequences. How to automatically initialize the parameters in the model is our future study.

6. ACKNOWLEDGEMENTS

The authors acknowledge James Davis et al. and Andrea Prati et al. for providing the test data on their websites.

7. REFERENCES

[1] J. Besag, "On the statistical analysis of dirty pictures," *J. Roy. Stat. Soc. B*, vol. 48, pp. 259-302, 1986.

[2] N. Friedman and S. Russell, "Image segmentation in video sequences: A probabilistic approach," *Proc. 13th Conf. Uncertainty in Artificial Intelligence*, pp. 175-181, 1997.

[3] S. Jabri, Z. Duric, H. Wechsler, and A. Rosenfield, "Detection and location of people in video images using adaptive fusion of color and edge information," *Proc. 15th International Conf. Pattern Recognition*, vol. 4, pp. 627-630, 2000.

[4] I. Mikic, P. C. Cosman, G. T. Kogut, and M. M. Trivedi, "Moving shadow and object detection in traffic scenes," *Proc. 15th International Conf. Pattern Recognition*, vol. 1, pp. 321-324, 2000.

[5] J. Stauder, R. Mech, and J. Ostermann, "Detection of moving cast shadows for object segmentation," *IEEE Trans. Multimedia*, vol. 1, pp. 65-76, 1999.

[6] C. Stauffer, and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 22, pp. 747-757, 2000.

[7] A. M. Tekalp, *Digital Video Processing*, 1995.

[8] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 19, pp. 780-785, 1997.