A STOCHASTIC LEARNING ALGORITHM FOR PIXEL-LEVEL BACKGROUND MODELS

Nick Mould and Joseph P. Havlicek

University of Oklahoma
School of Electrical and Computer Engineering
Norman, OK, USA

ABSTRACT

A new stochastic learning algorithm for use in nonparametric pixel-level background models is presented in this paper. For the first time, we propose the use of kernel density estimation (KDE) techniques in the model update step to identify outliers within the pixel-level sample collections and replace them with with recently observed background image features. A neighborhood diffusion process that improves on recently reported scene model learning techniques is presented, wherein information sharing between similarly structured adjacent background models is encouraged to promote spatial consistency within localized image regions. We demonstrate the superiority of the proposed algorithm by comparison with the state-of-the-art ViBe system using the well known percentage correct classification (PCC) statistic and a new figure of merit, probability correct classification (PrCC), presented here for the first time.

Index Terms— video segmentation, scene modeling, background modeling

1. INTRODUCTION

In this paper, we present a new stochastic learning algorithm for nonparametric pixel-level background models that overcomes several important problems in video segmentation. The update policy employs data driven nondeterministic techniques to integrate pixel and neighborhood level image features in the grayscale colorspace. At the pixel level, recently observed grayscale image values that have been classified as background structure in the segmentation procedure are integrated into the model by replacing the outlying samples in the corresponding pixel-level background models. At the neighborhood level, new background values are propagated to a randomly chosen neighboring sample collection, where the probability of selection is directly proportional to a measurement of the similarity between the shapes of the central and adjacent sample collection distributions. We demonstrate the effectiveness of our proposed algorithm against the state-of-the-art visual background extraction (ViBe) system [1]. In all of our test cases, we observed significant improvement in terms of percentage correct classification (PCC) and a new metric *probability correct classification* (PrCC) presented here for the first time. In addition, the proposed algorithm achieves excellent performance against the well known and challenging *persistent ghost problem*.

Statistical background modeling techniques have generally been divided into parametric and nonparametric representations. Parametric statistical models first began to appear in the mid to late 1990s when the computational complexity required for parameter estimation became widely available. These types of models were specifically attractive because they required a minimal amount of memory storage to achieve success against a broad spectrum of the initially observed challenges in scene modeling. In [2], the distributions of background values at each pixel location were modeled using multivariate Gaussians in the YUV colorspace, resulting in good segmentations of foreground and background components in the presence of indoor and outdoor lighting variations. Later, Gaussian mixture models (GMMs) were used to characterize multimodal distributions of pixel level features arising in natural scenes composed of dynamic foreground and background components [3–5]. Due to the overwhelming success of parametric scene modeling techniques, significantly improved versions of the GMM [6] and reapplication of well known median filtering techniques [7] continue to be reported in the literature.

Nonparametric background models were proposed for characterizing the temporal distributions of image features in video when computer memory became abundant and the computational resources required for estimating probabilities from sample collections by kernel density estimation (KDE) became available. In practice, both parametric and nonparametric models have been implemented predominantly at the pixel level, resulting in a loss of spatial information when compared to early background modeling techniques. In the earliest pixel-level models, post processing was performed on the detected foreground pixels using connected components labeling followed by morphological region growing/decaying algorithms [8] to refine the final segmentations.

Recently, the lack of spatial awareness in pixel-level video segmentation algorithms has led to a resurgence in the use of neighborhood-level information in many recently reported scene modeling algorithms. In [9], a nonparametric joint domain-range model was used to model the distribution of background image features in the spatial-RGB feature space and KDE was used to estimate the probabilities of observed pixel values prior to segmentation. In other cases, spatial information has been incorporated into pixel-level models by representation through the use of the microstructural textural response [10], spatial derivatives [11] and the local binary pattern (LBP) feature [12]. In [1], Barnich and Van Droogenbroeck proposed the first completely stochastic nonparametric scene model, where background information is shared between neighboring pixel-level models in the model maintenance step. The ViBe system proposed in [1] has been theoretically and experimentally verified to achieve superior performance against a wide range of well known scene modeling algorithms [1, 13]. Table 1 presents an abbreviated list of the prominent scene modeling techniques to which ViBe has been compared favorably.

Our proposed algorithm and the ViBe system are both nonparametric pixel-level background models and they both perform neigh-

This work was supported in part by the U.S. Army Research Laboratory and the U.S. Army Research Office under grant W911NF-08-1-0293.

Table 1. Prominent Background Modeling Techniques

Author(s)	Model Description	Feature Vector	Feature Vector Localization
Donohoe, Hush and Ahmed [14]	Temporal Low-Pass	Grayscale	Pixel
McKenna, Jabri, Duric, *et al.* [15]	Multivariate Normal	RGB/Sobel	Neighborhood
Oliver, Rosario and Pentland [16]	PCA	Grayscale	Frame
Stauffer and Grimson [4, 5]	GMM	Grayscale/RGB	Pixel
Elgammal [17, 18]	Nonparametric	Grayscale/RGB	Pixel
Cucchiara, Piccardi and Prati [19]	Median Filter	RGB	Pixel
Zivkovic [6, 20]	GMM	Grayscale/RGB	Pixel
Kim, Thanarat, Chalidabbhognse, *et al.* [21]	Codebook	RGB	Pixel

borhood level information sharing to promote spatial consistency throughout the image lattice. The difference between the two algorithms is the model update policy. In ViBe, new background samples are integrated into the corresponding models by random replacement using a uniformly distributed random variable, and information is swapped between neighboring models by randomly selecting a neighbor and then randomly replacing a sample within that neighboring sample collection. We have observed that random neighbor selection produces undesirable results in cases where the adjacent sample collections have differently shaped distributions, as is the case with pixels on different sides of an edge. In addition, the random replacement of samples within the distribution has a tendency to corrupt the model in cases where the replaced value was in a high density region within the model. The proposed method overcomes the limitations of ViBe by (1) replacing the outliers within the distributions and (2) discouraging the sharing of information between incompatible sample collections.

2. PROPOSED SCENE MODEL

We employ a pixel level nonparametric model to characterize the temporal distributions of background image features according to [1, 17, 18]

$$M(\mathbf{p}) = \{\phi_1, \phi_2, \phi_3, \ldots, \phi_N\}, \tag{1}$$

where M is a nonparametric model of the background scene represented by a collection of N previously observed values in the grayscale intensity feature space and $\mathbf{p} = (x_1, x_2)$ is the spatial coordinate of a single pixel. In terms of versatility, nonparametric models are unique in that they are well suited to the representation of multimodal statistical distributions where the number of modes is unknown and likely to change over time.

We performed a blind initialization of the model over N frames, by assigning each grayscale value directly according to

$$\begin{aligned} M(\mathbf{p}) &= \{\phi_1, \phi_2, \phi_3, \ldots, \phi_N\} \\ &= \{I_{k-(N-1)}(\mathbf{p}), \ldots, I_k(\mathbf{p})\}, \end{aligned} \tag{2}$$

where I_k represents a single video frame at time k. Because descriptive information about the foreground and background structures is not generally available during the initialization process, and because the presence of moving foreground objects is both likely to occur and unlikely to be detected accurately, we elected to use a naive initialization strategy. With this approach, the effects of a moving object are spread over several spatial locations rather than concentrated at a single location as in the case of the single frame bootstrapping techniques.

Segmentation was performed by thresholding the estimated background probabilities of each observed pixel value $I_k(\mathbf{p})$ within the unsegmented frame I_k according to

$$L_k(\mathbf{p}) = \begin{cases} \text{Foreground}, & P(I_k(\mathbf{p})) < T \\ \text{Background}, & \text{Otherwise} \end{cases}, \tag{3}$$

where T is a fixed threshold and $P(I_k(\mathbf{p}))$ is the background probability of a single observed pixel estimated by

$$P(I_k(\mathbf{p})) = \frac{1}{N} \sum_{i=1}^{N} K(I_k(\mathbf{p}), \phi_i^{\mathbf{p}}). \tag{4}$$

In (4), $\phi_i^{\mathbf{p}}$ represents the i'th sample from the background model M at pixel location $\mathbf{p}$, and K is a uniform spherical cutoff kernel of radius R defined by [1]

$$K(a, b) = \begin{cases} 1, & |a - b| \leq R \\ 0, & \text{Otherwise} \end{cases} \tag{5}$$

where $a, b \in \mathbb{R}$.

3. MODEL UPDATE POLICY

Because nonparametric models characterize statistical distributions with fixed size sample collections, learning is generally conducted by replacement of the oldest value within the sample collection [17]. In [1], the authors propose a technique whereby the sample to be replaced is chosen by a uniformly distributed random variable, arguing that this tactic produces a uniform decay of the sample collection over time. We propose a more conservative approach based on replacement of the most significant outliers in the collection. This replacement strategy is similar to the online k-means algorithm used in parametric models [4] in that low probability regions within the model are more likely to be discarded and replaced with more recent observations.

We define the outlier in each background model to be the least probable value and identify it by estimating the probability of each sample with respect to the entire sample collection using KDE according to

$$\text{Outlier Index} = \underset{i=1,\ldots,N}{\arg\min} \frac{1}{N} \sum_{j=1}^{N} K(\phi_i^{\mathbf{p}}, \phi_j^{\mathbf{p}}), \tag{6}$$

where $\phi_i^{\mathbf{p}}$ and $\phi_j^{\mathbf{p}}$ are samples from the model $M(\mathbf{p})$ and K is a spherical cutoff kernel. In (6), the radius of the kernel is computed from the data using a technique introduced by Elgammal in [17], where the bandwidth is set to the median absolute deviation measured between all of the possible unique sample pairs and where

pairs composed of identical samples are excluded. In the case where no unique outlier exists, the sample to be replaced is selected at random from the collection of minumum probability values identified by (6).

We propose a neighborhood information sharing policy that is similar to [1] except that we specifically discourage the sharing of information between incompatible sample collections. For each pixel level background model, we form a probability mass function by assigning a weight to each of the eight-connected neighboring background models based on a measurement of the similarity between the central $M(\mathbf{p})$ and neighboring $M(\mathbf{q})$ distributions. Here, $\mathbf{q}$ represents the spatial location of a single neighboring background distribution in the collection of pixels that are eight-connected neighbors of the central pixel $\mathbf{p}$ indicated by $\Lambda(\mathbf{p})$. The similarity metric w is computed by measuring the L^2 norm between histograms of the two sample distributions and then exponentiating the result according to

$$w_i(\mathbf{p},\mathbf{q}) = \exp\left[-\left(\sum_{i=1}^{256}[h(M(\mathbf{p}))_i - h(M(\mathbf{q}))_i]^2\right)^{(1/2)}\right], \tag{7}$$

where $h(\cdot)$ is a function that takes a collection of values and produces a 256 bin histogram and $\mathbf{q} \in \Lambda(\mathbf{p})$. The neighboring distribution that the new background value will be inserted into is selected by drawing at random from the distribution defined by the normalized neighborhood similarity weights $\{w_i\}_{i\in|\Lambda(\mathbf{p})|}$. Once a neighboring distribution is selected, the value is integrated into the model using the outlier replacement strategy given in (6).

4. RESULTS & DISCUSSION

We demonstrate the effectiveness of our proposed algorithm on a surveillance video provided by the performance evaluation in tracking and surveillance (PETS) workshop [22] and the Beach People sequence from the University of California San Diego (UCSD) background subtraction dataset [23]. We processed 200 frames of the PETS sequence and 250 frames of the Beach People sequence. In both cases, ground truth data was manually created.

Table 2 shows the results of the proposed algorithm compared to ViBe in terms of percentage correct classification (PCC) and probability of correct classification (PrCC). Percentage correct classification (PCC) is computed according to

$$\text{PCC} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}, \tag{8}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. To better identify the differences in the two scene models, we propose the probability of correct classification (PrCC) performance metric and use it to evaluate each algorithm prior to application of the final segmentation threshold. We argue that the pixel-level foreground and background probabilities allow for a richer analysis of the scene models when compared to the alternative binary classification results that have been traditionally used to evaluate video segmentation systems. The PrCC is computed according to

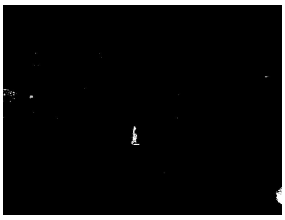
$$\text{PrCC} = \frac{\text{TP}_{\text{prob}}+\text{TN}_{\text{prob}}}{\text{TP}_{\text{prob}}+\text{TN}_{\text{prob}}+\text{FP}_{\text{prob}}+\text{FN}_{\text{prob}}} \tag{9}$$

where TP_{prob} is the sum of the foreground probabilities at the ground truth foreground pixel locations, TN_{prob} is the sum of the background probabilities at the ground truth background locations, FP_{prob} is the sum of the foreground probabilities at the ground truth background locations and FN_{prob} is the sum of the background probabilities at the ground truth foreground location.

Table 2. Performance results for the proposed algorithm and ViBe using Percentage Correct Classification (PCC) and Probability Correct Classification (PrCC).

Sequence	PCC		PrCC	
	ViBe	Proposed	ViBe	Proposed
PETS	99.3%	**99.5%**	93.5%	**99.0%**
BeachPeople	93.5%	**95.1%**	85.6%	**87.3%**

(a)



(b)

Fig. 2. Final segmentation results depicting frame 450 of the PETS sequence after application of the segmentation threshold to the probability images shown in Fig.1 (g, h). Subfigure (a) corresponds to the foreground probability image in Fig.1 (g), and (b) corresponds to the foreground probability image in Fig.1 (h).

Fig. 1 depicts selected frames from the two test videos, where the effectiveness of our algorithm can be observed by comparing foreground probability images from ViBe and the proposed algorithm. Clearly, the model update policy proposed in this paper produces a significant reduction in the number of potential false positives along the edges of stationary background structures. The ghost problem occurs in situations where stationary objects begin to move and uncover previously unobserved background structure resulting in the erroneous detection of foreground object. Fig. 2 shows the final segmentation results from the two frames shown in Fig. 1, where a persistent ghost in the original location of the moving person (Fig. 2(a)) is eliminated by the proposed model update technique (Fig. 2(b)). Our proposed scene model is effective against the ghost problem because the false foreground detections in the ghost region are quickly identified as outliers and replaced with more appropriate background pixel values in the model update step.

5. REFERENCES

[1] O. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algortihm for video sequences," *IEEE Trans. Image Processing*, vol. 20, no. 6, pp. 1709–1724, June 2011.

[2] C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal., Machine Intel.*, vol. 19, no. 7, pp. 780–785, July 1997.

[3] N. Friedman and S. Russell, "Image segmentation in video sequences: A probabilistic approach," in *Proc. Thirteenth Conf. Uncertainty in Artificial Intelligence*, Aug. 1-3 1997.

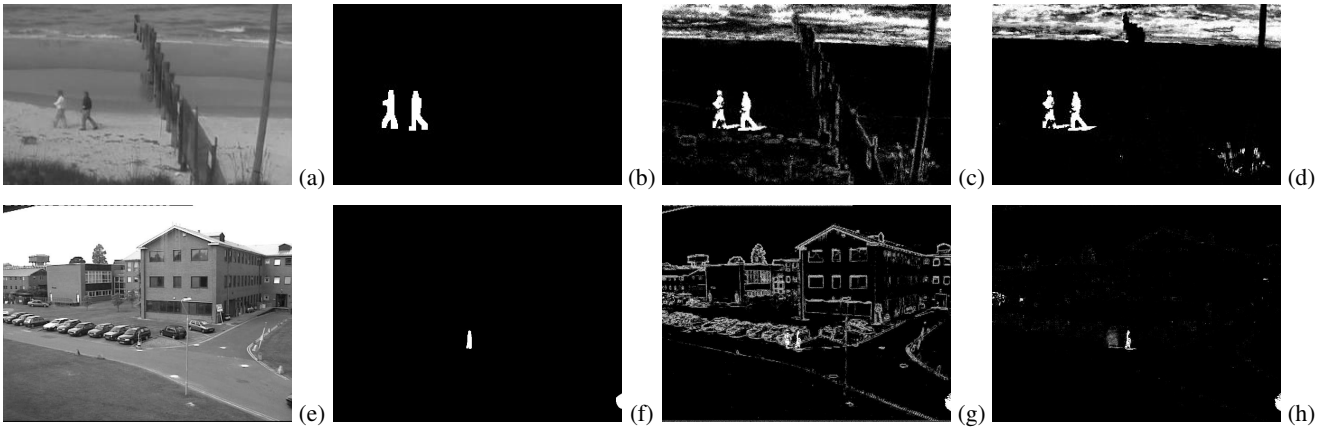


Fig. 1. Selected video frames depicting frame 230 of the Beach People sequence and frame 450 of the PETS sequence. (a) and (e) show the original grayscale images, (b) and (f) show the ground truth segmentation images, (c) and (g) show the ViBe foreground probability images and (d) and (h) show the foreground probability images from the proposed algorithm.

[4] C. Stauffer and W.E.L. Grimson, "Adaptive mixture models for real-time tracking," in *Proc. IEEE Int'l. Conf. on Comp. Vision, Pattern Recog.*, Fort Collins, CO, USA, June 23-25 1999, vol. 2.

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

[6] Z. Zivkovic and F. van der Heijden, "Recursive unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal., Machine Intel.*, vol. 26, no. 5, pp. 651–656, May 2004.

[7] A. Briassouli and N. Ahuja, "Extraction and analysis of multiple periodic motions in video sequences," *IEEE Trans. Pattern Anal., Machine Intel.*, vol. 29, no. 7, pp. 1244–1261, July 2007.

[8] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proc. IEEE Int'l. Conf. Computer Vision*, Kerkyra, Sep. 20-27 1999, pp. 255–261.

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

[10] C. Benedek and T. Sziranyi, "Bayesian foreground and shadow detection in uncertain frame rate surveillance videos," *IEEE Trans. Image Processing*, vol. 17, no. 4, pp. 608–621, Apr. 2008.

[11] A. Adam, R. Kimmel, and E. Rivlin, "On scene segmentation and histograms-based curve evolution," *IEEE Trans. Pattern Anal., Machine Intel.*, vol. 31, no. 9, pp. 1708–1714, Sep. 2009.

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

[13] S. Brutzer, B. Hoferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *Proc. IEEE Int'l. Conf. on Comp. Vision, Pattern Recog.*, Colorado Springs, CO, June 2011, pp. 1937–1944.

[14] G.W. Donohoe, D.R. Hush, and N. Ahmed, "Change detection for target detection and classification in video sequences," in *Proc. IEEE Int'l. Conf. Acoustic Speech and Signal Processing*, 1988, pp. 1084–1087.

[15] S.J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," in *Computer Vision and Image Understanding*, Oct 2000, vol. 80(1), pp. 42–56.

[16] N.M. Oliver, B. Rosario, and A.P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal., Machine Intel.*, vol. 22, no. 8, pp. 831–843, Aug. 2000.

[17] A.M. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in *Proc. European Conf. Computer Vision*, 2000, vol. 1843, pp. 751–767.

[18] A. Elgammal, R. Duraiswami, D. Harwood, and L. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proc. IEEE*, vol. 90, no. 7, pp. 1151–1163, July 2002.

[19] R. Cucchiara, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Trans. Pattern Anal., Machine Intel.*, vol. 25, no. 10, pp. 1337–1342, Oct. 2003.

[20] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proc. IEEE Int'l. Conf. on Pattern Recognition*, 2004, vol. 2, pp. 28–31.

[21] K. Kim, T. Thanarat, H. Chalidabbhognse, D. Harwood, and L. Davis, "Real time foreground-background segmentation using codebook model," *Real-Time Imaging*, vol. 11, pp. 172–185, 2005.

[22] Computer Vision Pattern Recognition (CVPR), "Performance evaluation of tracking and surveillance (PETS)," *Website: ftp://ftp.pets.rdg.ac.uk/PETS2001/*, 2001.

[23] Vijay Mahadevan and Nuno Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Trans. Pattern Anal., Machine Intel.*, vol. 32, no. 1, pp. 171–177, 2010.