Background Modeling Using Adaptive Cluster Density Estimation for Automatic Human Detection

Harish Bhaskar[1], Lyudmila Mihaylova[1] and Simon Maskell[2]
[1]Lancaster University, United Kingdom [2]QinetiQ, Malvern, United Kingdom
(h.bhaskar, mila.mihaylova)@lancaster.ac.uk, s.maskell@signal.qinetiq.com

Abstract: Detection is an inherent part of every advanced automatic tracking system. In this work we focus on automatic detection of humans by enhanced background subtraction. *Background subtraction* (BS) refers to the process of segmenting moving regions from video sensor data and is usually performed at pixel level. In its standard form this technique involves building a model of the background and extracting regions of the foreground. In this paper, we propose a *cluster-based* BS technique using a mixture of Gaussians. An adaptive mechanism is developed that allows automated learning of the model parameters. The efficiency of the designed technique is demonstrated in comparison with a pixel-based BS [ZdH06].

1 Introduction & Related Work

Motion detection is critical to many automated visual applications. A high degree of sensitivity and robustness is often desired from detection mechanisms. The simplest way of accomplishing detection is through building a representation of the scene background and comparing each new frame with this representation. This procedure is known as *background subtraction*. Some of the popular techniques for BS include mixture of Gaussians [ZdH06], kernel density estimation [EHD00], colour and gradient cues [JSS02], high level region analysis [KKBM99], Kalman filter [ZS03], hidden Markov models [SRP$^+$01], and Markov random fields [PR01]. The general idea behind some of the aforementioned techniques is to represent each pixel of an scene using a probability density function (PDF). A pixel from a new image is classified as background depending on how well described the pixel is by its density function. However, these techniques are bounded by limitations such as explicitly handling dynamic changes of the background, e.g., gradual or sudden (as in moving clouds); motion changes including camera oscillations and high frequency background objects (tree branches, sea waves, etc.) and changes in the background geometry (such as parked cars) [CGPP05]. In this paper we propose an automated detection algorithm using cluster density estimation based on a Gaussian mixture model (GMM) and self adaptative parameters. The rest of the paper is organised as follows. Section 2 presents the proposed detection technique, Section 3 gives results over real video sequences, and the last Section contains conclusions.

2 The Proposed Technique

The fundamental problem of *cluster* background subtraction involves a decision whether a *cluster of pixels* belongs to the *background* (*bG*) or *foreground* (*fG*) object based on the

ratio of probability density functions:

$$\frac{p(bG|\boldsymbol{c}_k^i)}{p(fG|\boldsymbol{c}_k^i)} = \frac{p(\boldsymbol{c}_k^i|bG)p(bG)}{p(\boldsymbol{c}_k^i|fG)p(fG)}, \tag{1}$$

where, the vector $\boldsymbol{c}_k^i = (c_{1,k}^i, \ldots, c_{\ell,k}^i)$ characterises the i-th cluster ($0 \leq i \leq q$) at time instant k (and current image), containing ℓ number of pixels such that $[Im]_k = \left[\boldsymbol{c}_k^1, \ldots, \boldsymbol{c}_k^q\right]$ is the whole image; $p(bG|\boldsymbol{c}_k^i)$ is the probability density function (PDF) of the background, subtracted based on a certain feature (e.g., colour, edges) of the cluster $\boldsymbol{c}_k^i$; $p(fG|\boldsymbol{c}_k^i)$ is the PDF of the foreground on the same cluster $\boldsymbol{c}_k^i$; $p(\boldsymbol{c}_k^i|bG)$ refers to the PDF model of the background and $p(\boldsymbol{c}_k^i|fG)$ is the appearance model of the foreground object. In our cluster BS technique the decision that any cluster belongs to a background is made if:

$$p(\boldsymbol{c}_k^i|bG) > threshold \left(= \frac{p(\boldsymbol{c}_k^i|fG)p(fG)}{p(bG)}\right). \tag{2}$$

Since the threshold is a scalar, the decision in (2) is made based on the average of the distributions of all pixels within the cluster $\boldsymbol{c}_k^i$. Most of the existing BS techniques such as [EHD00, ZdH06] take this decision at pixel level in contract to the proposed here algorithm at cluster level. The appearance of the foreground, characterised by $p(\boldsymbol{c}_k^i|fG)$ is assumed uniform. The background model represented as $p(\boldsymbol{c}_k^i|bG)$ is estimated from a training set $\Re$ which is a rolling collection of images over a specific update time T. The time T is crucial since its update determines the model ability to adapt to illumination changes and to handle appearances and disappearances of objects in a scene. If the frame rate is known, the time period T can be adapted: $T = \frac{N}{fps}$, e.g., as a ratio between the number N of frames obtained through the online process and the frame rate, fps, frames per second. At time instant k we have $\Re_k = \left\{\boldsymbol{c}_k^i, ..., \boldsymbol{c}_{k-T}^i\right\}$.

Every cluster $\boldsymbol{c}_k^i$, $(0 \leq i \leq q)$ at time instant k is generated using a colour clustering mechanism of the nearest neighbour approach [vdHDdR04], although other techniques can be used. The aim of the clustering process is to separate data based on certain similarities. Clustering is carried out based on the *hue, value, saturation* (HSV) colour model due to its inherent ability to cope with illumination changes. Constraints such as spatial distance, hue difference and brightness changes are imposed on the model. For each pixel feature x from the image $[Im]_k$, the Euclidean norm is calculated between the pixel feature and its neighbourhood of n_B connected pixels x_{nB}. If the value d of the Euclidean norm is smaller than a predefined threshold ϵ, i.e., $d \leq \epsilon$, then the pixels from regions sharing similar colour features are clustered together. The results, however, can vary in accordance with the features chosen and the type of clustering scheme. A GMM containing M components is then used to represent the density distribution

$$\widetilde{p}(\boldsymbol{c}_k^i|\Re_k, bG + fG) = \sum_{m=1}^{M} \widetilde{\pi}_{m,k} \mathcal{N}(\boldsymbol{c}_k^i; \widetilde{\boldsymbol{\mu}}_k, \widetilde{\sigma}_{m,k}^2 I), \tag{3}$$

both of the background and foreground where $\widetilde{\boldsymbol{\mu}}_{1,k}, ..., \widetilde{\boldsymbol{\mu}}_{M,k}$ and $\widetilde{\sigma}_{1,k}^2, ..., \widetilde{\sigma}_{M,k}^2$ are the estimates of the mean vectors and of the variances that describe the Gaussian components; I is the identity matrix. The estimated mixing weights $\widetilde{\pi}_m$ sum up to one. Given the

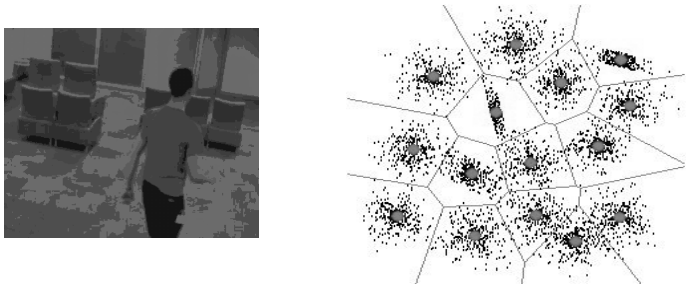

Figure 1: a) An image with clusters b) Clusters with their centres

new cluster $\boldsymbol{c}_k^i$ at time instant k, the update equations for the cluster parameters can be calculated as follows:

$$\widetilde{\pi}_{m,k+1} = \widetilde{\pi}_{m,k} + \frac{1}{T_k}(o_{m,k} - \widetilde{\pi}_{m,k}), \widetilde{\boldsymbol{\mu}}_{m,k+1} = \widetilde{\boldsymbol{\mu}}_{m,k} + o_{m,k}\left(\frac{1}{T_k\widetilde{\pi}_{m,k}}\right)\boldsymbol{\delta}_{m,k}, \quad (4)$$

$$\widetilde{\sigma}^2_{m,k+1} = \widetilde{\sigma}^2_{m,k} + o_{m,k}(\frac{1}{T_k\widetilde{\pi}_{m,k}})(\boldsymbol{\delta}'_{m,k}\boldsymbol{\delta}_{m,k} - \sigma^2_{m,k}), \quad (5)$$

where $\boldsymbol{\delta}_{m,k} = \boldsymbol{c}_k^i - \widetilde{\boldsymbol{\mu}}_{m,k}$, $'$ denotes the transpose operation, and $o_{m,k}$ refers to the ownership of the new cluster and defines the closeness of this cluster to a particular GMM component. The ownership of any new cluster is set to 1 for "close" components (with the largest $\widetilde{\pi}_{m,k}$), and the others are set to zero. A cluster is close to a component iff the Mahalanobis distance between the Gaussian mixture component and the cluster centre is, e.g., less than 3. If there exist no "close" components, a component is generated with $\widetilde{\pi}_{m+1,k} = \frac{1}{T_k}$, with an initial mean $\widetilde{\boldsymbol{\mu}}_0$ and variance $\widetilde{\sigma}^2_0$. The model presents clustering of components and the background is approximated with the B largest components,

$$\widetilde{p}(\boldsymbol{c}_k^i|\Re_k, bG) \sim \sum_{m=1}^{B} \widetilde{\pi}_{m,k}\mathcal{N}(\widetilde{\boldsymbol{\mu}}_k, \widetilde{\sigma}^2_m I), \quad B = \underset{b}{\operatorname{argmin}}(\sum_{m=1}^{b} \widetilde{\pi}_{m,k} > (1 - c_f)), \quad (6)$$

where b is a variable defining the number of considered clusters, c_f is the proportion of the data that belong to foreground objects without influencing the background model. The proportionality between the pixels belonging to the foreground to the pixels from the background is assumed constant in most adaptive models [ZdH06]. This assumption does not hold true when videos of objects are captured from a close proximity. In such circumstances, the proportion of pixels belonging to the objects of interest, i.e., the foreground pixels, are much higher than the background pixels. This ratio defining the percentage of foreground and background pixels can be updated from the training set as follows:

$$c_f = \frac{\widetilde{p}(\boldsymbol{c}_k^i|\Re_k, fG)}{\widetilde{p}(\boldsymbol{c}_k^i|\Re_k, bG)}. \quad (7)$$

3 Results and Analysis

The proposed clustering model is compared to the pixel-based BS GMM developed in [ZdH06]. Figures 2 and 3 illustrate the outcome of BS on two video sequences. The first sequence of frames is the original sequence, followed by the results of the pixel level

BS GMM and the results of the proposed model, respectively. It can be clearly seen that the pixel-based model is contaminated with a large amount of clutter or false alarms in detection. The detection of the proposed cluster-based GMM is considerably better. The

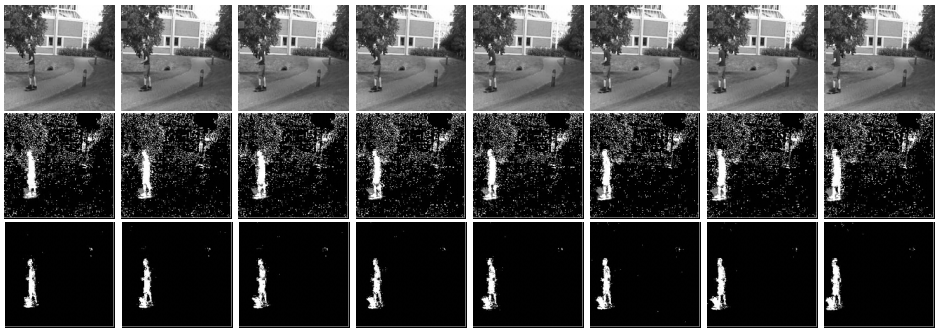

Figure 2: Results from the pixel-based BS [ZdH06] and proposed cluster BS on sequence 1

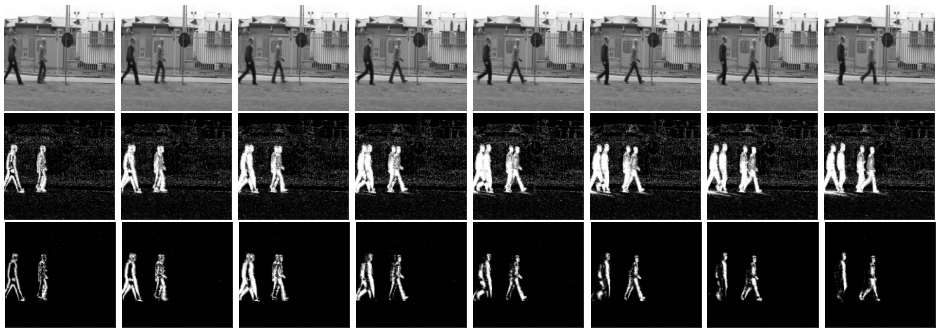

Figure 3: Results from the pixel-based BS [ZdH06] and proposed cluster BS model on sequence 2

techniques are also compared using quantitative measurements such as recall and precision. Recall and precision measures quantify how well an algorithm matches the ground truth [CK04]. *Recall* [DG06] is calculated as the ratio of the number of foreground pixels correctly identified to the number of foreground pixels in the ground truth and precision is computed as the ratio of the number of foreground pixels correctly identified to the number of foreground pixels detected. Figure 4 shows that the proposed algorithm has

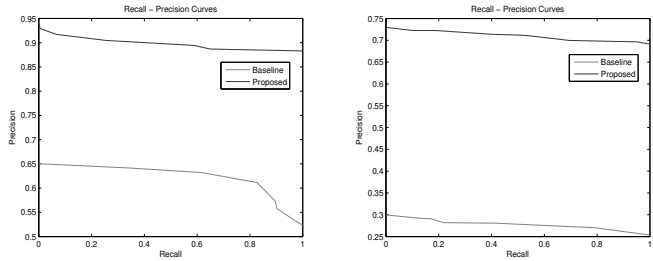

Figure 4: Recall-Precision curves of the pixel-based BS [ZdH06] and proposed cluster BS model, for sequences 1 and 2, respectively.

higher level of precision for the same values of the recall. The precision values directly relate to the number of correctly classified foreground pixels [DG06], and are inversely proportional to the misclassified foreground pixels. It is evident that compared with the pixel-based GMM [ZdH06] the proposed cluster GMM maximises the proportion of correctly classified pixels and minimises the misclassification.

4 Conclusions

In this paper we proposed a *cluster* level background subtraction technique based on a GMM. The model parameters are adapted through a learning process. The proposed model has been compared with a pixel level GMM technique for BS and much superior performance has been reported for the proposed technique. It reduces considerably the clutter and achieves high level of precision. Open issues for future research are learning the appearance model of the foreground, adapting it to moving backgrounds, adapting the number of Gaussian mixture components and embedding the detection process into an automatic tracking system.

Acknowledgements. The authors acknowledge the support of UK MOD Data and Information Fusion Defence Technology Centre under the Tracking Cluster project DIFDTC/CSIPC1/02.

References

[CGPP05] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting Moving Objects, Ghosts and Shadows in Video Streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1337–1342, 2005.

[CK04] S.-C. Cheung and C. Kamath. Robust techniques for background subtraction in urban traffic video. *Video Communications and Image Processing*, 5308(1):881–892, 2004.

[DG06] J. Davis and M. Goadrich. The Relationship Between Precision-Recall and ROC Curves. In *Proc. 23rd Intl. Conf. on Machine Learning*, 2006.

[EHD00] A. Elgammal, D. Harwood, and L. Davis. Non-parametric Model for Background Subtraction. In *Proc. 6th Europ. Conf. on Computer Vision*, June/July 2000.

[JSS02] O. Javed, K. Shafique, and M. Shah. A Hierarchical Approach to Robust Background Subtraction using Color and Gradient Information. In *Proc. of IEEE Workshop on Motion and Video Computing*, 2002.

[KKBM99] K.Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *Proc. Intl. Conf. Comp. Vis.*, pages 255–261, 1999.

[PR01] N. Paragios and V. Ramesh. A MRF-based Real-Time Approach for Subway Monitoring. In *Proc. of IEEE Conf. CVPR*, 2001.

[SRP+01] B. Stenger, V. Ramesh, N. Paragios, F. Coetzec, and J.M. Buhmann. Topology free hidden Markov models: application to background modeling. In *Proc. of the Intl. Conf. on Computer Vision.*, 2001.

[vdHDdR04] F. van der Heijden, R.P.W. Duin, and D. de Ridder. *Classification, Parameter Estimation and State Estimation*. John Wiley and Sons, 2004.

[ZdH06] Z. Zivkovic and F. V. der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pat. Rec. Letters*, 27(7):773–780, 2006.

[ZS03] J. Zhong and S. Sclaroff. Segmenting Foreground Objects from a Dynamic Textured Background via a Robust Kalman Filter. In *Proc. IEEE Intl. Conf. Comp. Vis.*, 2003.