ADAPTIVE EIGEN-BACKGROUNDS FOR OBJECT DETECTION

J. Rymel, J. Renno, D. Greenhill, J. Orwell and G.A. Jones

Digital Imaging Research Centre
School of Computing and Information Systems,
Kingston University,
Kingston upon Thames, Surrey, KT1 2EE , UK
{j.rymel,j.renno,d.greenhill,j.orwell,g.jones}@kingston.ac.uk

ABSTRACT

Most tracking algorithms detect moving objects by comparing incoming images against a reference frame. Crucially, this reference image must adapt continuously to the current lighting conditions if objects are to be accurately differentiated. In this work, a novel *appearance model* method is presented based on the *eigen-background* approach[1]. The image can be efficiently represented by a set of appearance models with few significant dimensions. Rather than accumulating the necessarily enormous training set to generate the eigen model, the described technique builds and adapts the eigen-model online evolving both the parameters and number of significant dimension. For each incoming image, a reference frame may be efficiently hypothesized from a subsample of the incoming pixels. A comparative evaluation that measures segmentation accuracy using large amounts of manually derived *ground truth* is presented.

1. INTRODUCTION

Most motion detection and tracking systems detect moving objects by comparing incoming images against a reference frame representing all the static structure within the scene[2, 3, 4]. Moving objects are extracted as connected components of pixels which *significantly* differ from this reference frame. These objects are typically tracked across the image (and hence ground plane) in order to generate a temporal description of events within a scene. Although a number of applications of the technique have been described for indoor scenes, the most popular application domain has been visual surveillance of public spaces. Crucially, this reference image must continuously adapt to the current lighting conditions if moving objects are to be accurately differentiated. In the recent past the problem of background estimation has become an increasing interesting research area.

Outdoor environments set unique challenges for constructing of accurate reference images including changes in *environmental scene illumination* including slow variations from dawn-to-dusk, highly disruptive glare of reflected sunlight, and the more rapid variations caused by weather *e.g.* rain and snow on roads; *embedded scene elements* such as swaying trees or traffic lights generate significant variations over arbitrarily large regions; *camera oscillation* induced by winds on poorly damped camera mountings result in highly disruptive periodic variations; *event induced scene variations* are luminance changes of the background caused by moving objects themselves *e.g.* shadows and the sweep of headlights on rain-soaked surfaces; *stationary events* are typically vehicles that park. Moreover, stationary objects must at some point be incorporated within the background scene representation if subsequent events passing between the camera and this stationary object are to be detected.

Popular background estimation techniques model pixels individually. Stauffer and Grimson[4] build and iteratively update a multi-modal model of the greylevel *probability density functions* (PDF) of the static background at each pixel using a mixture of Gaussians. However this fails to take into account the substantial degree of correlation between neighbouring pixels. *Active appearance models* (AAM)[5] are able to exploit the very high degree of correlation of pixel luminance and capture the inherent variability with relatively few *shape parameters*. Pentland *et al* [2] use this technique to build *Eigen-Backgrounds*. Black *et al* [6] utilise robust statistics to identify outliers within the training data, and improve the specificity of the models.

In this paper, the work of Oliver and Pentland[2] is extended by proposing a novel background estimation technique that learns the covariation of greylevels within the incoming images using principal component analysis to generate the *eigen-backgrounds*. Rather than accumulating the necessarily enormous training set, our technique builds and adapts the eigen-model online. The number of significant modes as well as the mean and covariance of the model as continuously adapted to match the environmental conditions. Our technique takes advantage of the over-constraint available to efficently use sub-sampling when generating background hypothesises.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

2. DESCRIPTION OF ALGORITHM

In order to identify moving objects within the image stream, it is necessary to build a stochastic representation of the static areas. This is done by dividing the image into a grid of neighbourhoods and continuously learning the statistical variations within each. For every incoming image a background estimate is *hypothesised* for each neighbourhood to facilitate comparison with the incoming frame. The incoming frame is the used to *update* the statistical model for each neighbourhood. Detection is achieved by thresholding greylevel differences against estimates of the greaylevel variance at each pixel.

2.1. Background hypothesising

Any neighbourhood in a greylevel background image may be represented as an attribute vector $\mathbf{i} = (i_1, \ldots, i_N)$ where i_n represents the greylevel of the n^{th} pixel in an image neighbourhood containing N pixels. This static background image may be formulated as an *appearance model* [2] where any image instance may be represented as a linear deformation from an average background image $\hat{\mathbf{i}}$

$$\mathbf{i} = \widehat{\mathbf{i}} + \mathbf{P}\mathbf{b} \qquad \mathbf{b} = \mathbf{P}^T \delta\mathbf{i} \tag{1}$$

where the $N \times M$ matrix $\mathbf{P}$ represents the orthonormal shape matrix (the inverse is equivalent to the transpose P^T) usually derived from *principal component analysis* (PCA) of a large training set $\mathcal{T}$ of background images, and the vector $\mathbf{b}$ of dimension M represents the M most significant shape parameters.

Although based upon *eigen-backgrounds*[2], there are a number of extensions. For each incoming image from the camera, a reference frame must be hypothesized representing the current static scene (*i.e.* with no moving objects present) under the current lighting conditions. This is achieved efficiently by sub-sampling a subset of pixel locations $\{i_j; j = 1, \kappa M\}$ and their associated greylevels. The samples are chosen using a uniform random number generator from within each neighbourhood. On the assumption that these greylevels are projected from the static scene, a $M \times 1$ appearance parameter vector $\mathbf{b}$ may be generated in the appearance model for each neighbourhood as follows

$$\mathbf{b}' = \left[\mathbf{P}'^T\mathbf{P}'\right]^{-1}\mathbf{P}'^T\left(\mathbf{i}' - \widehat{\mathbf{i}}'\right) \tag{2}$$

where $\mathbf{P}'$, $\mathbf{i}'$ and $\widehat{\mathbf{i}}'$ are the subsampled versions of the *shape* matrix, average neighbourhood and current neighbourhood vectors. The length κM of these matrix and vector quantities (and hence the number of sub-sampled pixel locations in each neighbourhood) is chosen to over-determine the estimation of $\mathbf{b}'$ to ensure some robustness to the selection of some pixels which result from the projection of some non-static scene element. Currently we choose this *sub sample* factor κ to be 4. Values larger yield little increase in robustness yet lead to greater computational cost.

The *number of significant modes* defines the complexity of the model used to hypothesis the background image. If too many modes are used then the computational overhead increases, while too few modes is likely to result in the generation of an inaccurate hypotheses. The solution is to limit the number to that required to accurately model the background. In addition, there are various *truncation strategies* such as Cootes *et al* [5] that can be used to constrain the hypothesis to plausible solution spaces.

2.2. Detection

The differences between the reference and incoming frame are thresholded against a pixel variance image to identify any pixel greylevels that are outside the variance expected at that pixel. These thresholded values are placed in a *detection mask*. There are various different methods of generating a pixel variance image. Pentland's technique [2] simply used an empirically chosen global variance. It is possible to use the variance of the modeled PDM itself to generate a variance for each pixel. Alternatively, it may also possible be sensible to derive the pixel variances from the temporally averaged inter-frame pixel differences.

2.3. Updating the Appearance Model

An appearance model describes the co-variability within an image neighbourhood, and is defined by the mean $\widehat{\mathbf{i}}_t$, the shape matrix $\mathbf{P}_t$ and the eigenvalues matrix $\mathbf{\Lambda}_t$, Since the eigenvalues represent the variance of distribution along the principal axes in the shape space, the covariance matrix can be constructed as a diagonal matrix composed of the ordered eigenvalues. Assuming an appearance model has already been instantiated, it will be updated to take into account the incoming image from the camera. New observations augment the previous covariance in two ways: firstly, by adapting the mean and covariance of the eigen model, and secondly, allowing evolution of the subspace itself by increasing the dimensionality of the subspace.

The first step is to project the new observation $\mathbf{i}_t$ into the eigenspace associated with the previous model to estimate the shape parameters $\mathbf{b}'_t$ using equation 2. Projecting this back into the image space, we compute the error vector $\mathbf{e}_t$ representing the displacement of the new point orthogonal to the original eigenspace *i.e.*

$$\mathbf{e}_t \quad = \quad \left(\mathbf{i}_t - \widehat{\mathbf{i}}_{t-1}\right) - \mathbf{P}_{t-1}\mathbf{b}'_t \tag{3}$$

In the third step, the new mean are is updated simply as

$$\widehat{\mathbf{i}}_t = \frac{1}{\tau + 1}\left(\tau\widehat{\mathbf{i}}_{t-1} + \mathbf{i}_t\right) \tag{4}$$

where $\tau = \min(t, T)$ is a weighting function controlling the influence of current data relative to history. Note that during the initialisation of a new model (*i.e.* $t < T$) the update process is influenced by the incoming data, after which the process updates using a fixed-length temporal filter.

The new eigenvalues $\mathbf{\Lambda}_t$ are estimated by recovering the eigen-decomposition $\mathbf{R}_{t,t-1}\mathbf{\Lambda}_t\mathbf{R}_{t,t-1}^T$ of the following *augmented covariance* Λ_t^+ derived from the original previous covariance matrix $\mathbf{\Lambda}_{t-1}$ and the new data sample $\mathbf{i}_t$[7].

$$\Lambda_t^+ = \frac{1}{\tau+1}\left(\tau\begin{bmatrix}\mathbf{\Lambda}_{t-1} & \mathbf{0}\\ \mathbf{0}^T & 0\end{bmatrix} + \begin{bmatrix}\mathbf{b}_t\mathbf{b}_t^T & \lambda\mathbf{b}_t\\ \lambda\mathbf{b}_t^T & \lambda^2\end{bmatrix}\right) \quad (5)$$

where $\lambda = \mathbf{e}_t^T(\mathbf{i}_t - \widehat{\mathbf{i}}_{t-1})/||\mathbf{e}_t||$. This augmented covariance is created by augmenting $\mathbf{\Lambda}_{t-1}$ in two ways: firstly by including the new image vectors $\mathbf{b}_t$ and $\mathbf{e}_t$, and secondly by allowing for an increase in the number of significant modes. The square matrix $\mathbf{R}_{t,t-1}$ represents the rotation of $\mathbf{\Lambda}_{t-1}$ in the enlarged space onto the new cluster covariance $\mathbf{\Lambda}_t$. Finally the new augmented shape matrix $\mathbf{P}_t^+$ is computed from the old as follows

$$\mathbf{P}_t^+ = [\mathbf{P}_{t-1}, \mathbf{e}_t/||\mathbf{e}_t||]\,\mathbf{R}_{t,t-1} \quad (6)$$

Note that the dimensionality of both the augmented shape matrix $\mathbf{P}_t^+$ and covariance $\mathbf{\Lambda}_t^+$ have been incremented.

2.3.1. Controlling the Shape Space Dimensionality

Updating the appearance model as detailed above involves continually incrementing the dimensionality of the shape matrix. This is desirable in the initial stages where there are insufficient dimensions to accurately model the data. Once the model has exceeded the optimal number of dimensions, M, the augmented covariance Λ_t^+ and shape matrix $\mathbf{P}_t^+$ must be continually truncated to M dimensions. Given the eigen decomposition of Λ_t^+ returns Λ and R ordered on decreasing eigenvalues, $\mathbf{P}_t$ is simply the left most M columns of $\mathbf{P}_t^+$, and Λ_t is the top left $M \times M$ matrix of Λ_t^+.

2.3.2. Creating a new Appearance Model

The hypothesis method assumes that an appropriate model exists to describe the incoming frame. The initial model is constructed by setting the mean of the neighbourhood to the value of the first frame $t = 0$, $\widehat{\mathbf{i}}_\mathbf{0} = \mathbf{i}_\mathbf{0}$. Then the initial 1×1 covariance matrix $\mathbf{\Lambda_0}$ is set to σ^2 where σ is an empirically chosen value that represents the amount of noise within the image. The initial shape matrix $\mathbf{P_0}$ is initialised as

$$\mathbf{P} = \left[\frac{1}{\sqrt{N}}, \ldots \frac{1}{\sqrt{N}}\right]^T \quad (7)$$

where N is the number of pixels. This shape matrix is initialised as a $N \times 1$ vector. As each additional frame is used to update the appearance model, the dimensionality of the subspace is allowed to increase to a maximum M_{MAX}.

3. RESULTS

The following comparative methodology compares the performance of our algorithm against the original eigen-background method[1] and the standard Stauffer and Grimson algorithm[4]. The performance metric compares motion classification output with manually derived *ground truth*. The *detection rate* and the *specificity* represent the percentage of correctly classified foreground and background pixels respectively Ellis[8]. True object pixels are defined as pixels inside a manually positioned rectangular bounding box. The product of these two measures generates a single per-frame performance metric.

Two datasets are proposed that represent different types of environmental variation from which objects are to be detected. The publically available PETS 2001 dataset features a stationary camera located at a high vantage point with a steep lookdown angle. There are a relatively few object occurrences and object interactions in the dataset. The frequency of periods of partial and full occlusion between objects is also quite low. One of the challenging features of this dataset is the small distant objects that have a very low contrast relative to the background. The DIRC dataset is an *in-house dataset* that was also captured in the RGB colour space. However, there are constant lighting variations between the different frames, the quality of the data set is a lot higher and there is less compression artifacts.

Fig. 1. (a) PETS Dataset (b) DIRC Dataset

The results from the three algorithms can be seen in Figure 2 and Table 1. Figure 2 plots the performance for a particular frame sequence from the PETS dataset. The *eigen background* algorithm[1] deteriorates over the sequence as this non-adaptive algorithm poorly models this fragment of the dataset. Both Stauffer and Grimson's algorithm and the proposed algorithm adapt quickly to the progressive change in environmental conditions. (Figure 2 also highlights a current artefact of the evaluation methodology where the metric falls to zero when there are no scene objects present.)

Table 1 presents these performance measures averaged over each of the datasets. The comparison was extended to include background estimation by temporal averaging[9]. It should be noted that the original eigen-background algorithm performed marginally better than temporal averaging

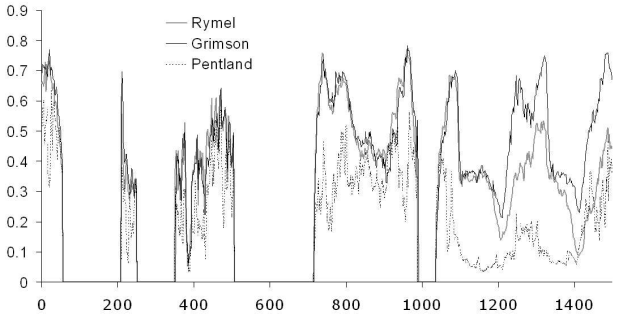

Fig. 2. Plots of the performance results for each algorithm.

(probably because the model is built from the entire training data set and poorly models any specific interval). The proposed algorithm and the Stauffer and Grimson method perform significantly better - largely because they are adaptive. Currently the Stauffer out-performs the proposed algorithm but it should be bourne in mind that the former maintains several *memories* per pixel allowing it to model the more rapid lighting changes. A similar multi-modal extension to our technique may improve accuracy significantly.

	Algorithm			
Data	Proposed Algorithm	Stauffer and Grimson[4]	Wren *et al*[1]	Temporal Averaging
PETS	0.22	0.28	0.13	0.11
DIRC	0.31	0.33	0.18	0.15

Table 1. Performance of the Change Detection Algorithms

4. CONCLUSIONS

In this paper we propose a solution to background estimation using *eigen-backgrounds* for visual surveillance. The advantages of our technique is its ability to take advantage of the covariability between pixel greylevels to accurately and efficiently hypothesize an accurate background estimate with relatively few sub-samples. The computational and storage burden required by eigen-background techniques is avoided by an online incremental eigen analysis algorithm which adapts the subspace itself to the environmental conditions. An important contribution is the ability to adapt the subspace itself. The method compares favorably to the state-of-the-art Stauffer and Grimson algorithm which uses a Mixture of Gaussians providing each pixel with multiple memories. We are currently investigating an extension to our algorithm to enable each neighbourhood to be modelled by multiple clusters in shape space (based on the work of Ong and Gong [10]). It is hoped this approach will increase the accuracy of the hypothesis generation by more closely modelling the non-linear aspects of lighting variations.

5. REFERENCES

[1] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, July 1997.

[2] N.M. Oliver, B. Rosario, and A.P. Pentland, "A Bayesian Computer Vision System for Modeling Human Interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831–843, Aug 2000.

[3] G.A. Jones J. Orwell, P. Remagnino, "From Connected Components to Object Sequences," in *First IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Grenoble, March 31st 2000, pp. 129–136.

[4] C. Stauffer and W.E.L. Grimson, "Learning Patterns of Activity using Real-Time Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, August 2000.

[5] T. F. Cootes, G. J. Edwards, and Christopher J. Taylor, "Active Appearance Models," in *Proceedings of the Fifth European Conference on Computer Vision*, Freiburg, Germany, 1998, pp. 484–498.

[6] Fernando De la Torre and Michael Black, "Robust principal component analysis for computer vision," in *International Conference on Computer Vision*, Vancouver, July 9-12 2001, vol. I, pp. 362–369.

[7] P.M. Hall, D. Marshall, and R.R. Martin, "Incremental Eigenanalysis for Classification," in *Proceedings of the British Machine Vision Conference*. 1998, vol. 1, pp. 286–295, BMVA Press.

[8] T. Ellis, "Performance Metrics and Methods for Tracking in Surveillance," in *Third IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS2002)*, Copenhagen, Denmark, June 2002, pp. 26–31.

[9] A.J. Lipton, H. Fujiyoshi, and R.S. Patil, "Moving target classification and tracking from real-time video," in *Proceedings IEEE Image Understanding Workshop*, 1998, pp. 129–136.

[10] E.-J. Ong and S. Gong, "A Dynamic Human Model using Hybrid 2D-3D Representations in Hierarchical PCA Space," in *Proceedings of the British Machine Vision Conference*, Nottingham, September 1999, pp. 33–42, BMVA Press.