# MRF-based Background Initialisation for Improved Foreground Detection in Cluttered Surveillance Videos

Vikas Reddy, Conrad Sanderson, Andres Sanin, Brian C. Lovell

NICTA, PO Box 6020, St Lucia, QLD 4067, Australia *
The University of Queensland, School of ITEE, QLD 4072, Australia

## Abstract

*Robust foreground object segmentation via background modelling is a difficult problem in cluttered environments, where obtaining a clear view of the background to model is almost impossible. In this paper, we propose a method capable of robustly estimating the background and detecting regions of interest in such environments. In particular, we propose to extend the background initialisation component of a recent patch-based foreground detection algorithm with an elaborate technique based on Markov Random Fields, where the optimal labelling solution is computed using iterated conditional modes. Rather than relying purely on local temporal statistics, the proposed technique takes into account the spatial continuity of the entire background. Experiments with several tracking algorithms on the CAVIAR dataset indicate that the proposed method leads to considerable improvements in object tracking accuracy, when compared to methods based on Gaussian mixture models and feature histograms.*

## 1. Introduction

One of the low-level tasks in most intelligent video surveillance applications (such as person tracking and identification) is to segment objects of interest from an image sequence. Typical segmentation approaches employ the idea of comparing each frame against a model of the background, followed by selecting the outliers (i.e., pixels or areas that do not fit the model). However, most methods presume the training image sequence used to model the background is free from foreground objects. This assumption is often not true in the case of uncontrolled environments such as train stations and motorways, where directly obtaining a clear background is almost impossible. Furthermore, in outdoor video surveillance a strong illumination change can render the existing background model ineffective (e.g., due to introduction of shadows [15]), thereby forcing us to compute a new background model. In such circumstances, it becomes inevitable to reinitialise the background model using cluttered sequences (i.e., where parts of the background are occluded). Robust background initialisation in these scenarios can result in improved segmentation of foreground objects, which in turn can lead to more accurate tracking.

The majority of the algorithms described in the literature, such as [9, 11, 13, 18], do not have a robust strategy to handle cluttered sequences. Specifically, they fail when the background in the training sequence is exposed for a shorter duration than foreground objects. This is due to the model being initialised by relying solely on the temporal statistics of the image data, which is easily affected by the inclusion of foreground objects in the training sequence.

To alleviate this problem, a few algorithms have been proposed to initialise the background image from cluttered image sequences. Typical examples include median filtering, finding pixel intervals of stable intensity in the image sequence [19], building a codebook for the background model [9], agglomerative clustering [6] and minimising an energy function using an $\alpha$*–expansion* algorithm [4]. However, none of them evaluate the foreground segmentation accuracy using their estimated background model.

In this paper, we propose to replace the background model initialisation component of a recently introduced foreground segmentation method [13] and show that the performance can be considerably improved in cluttered environments. The proposed background initialisation is carried out in a Markov Random Field (MRF) framework, where the optimal labelling solution is computed using iterated conditional modes. The spatial continuity of the background is also considered in addition to the temporal statistics of the training sequence. This strategy is particularly robust to training sequences containing foreground objects exposed for longer duration than the background over a given time interval.

Experiments on the CAVIAR dataset, where most of the sequences contain occluded backgrounds, show that the proposed framework (MRF + multi-stage classifier) yields considerably better results in terms of tracking accuracy than the baseline multi-stage classifier method [13] as well as methods based on Gaussian mixture models [17] and feature histograms [10].

*Published in: Lecture Notes in Computer Science (LNCS), Vol. 6494, pp. 547-559, 2011. http://dx.doi.org/10.1007/978-3-642-19318-7_43

We continue as follows. The overall foreground segmentation framework is described in Section 2, followed by the details of the proposed MRF-based background initialisation method in Section 3. Performance evaluations and comparisons with three other algorithms are given in Section 4, followed by the main findings in Section 5.

## 2. Foreground Segmentation Framework

We build on the patch-based multi-stage foreground segmentation method proposed in [13], which has four major components:

1. Division of a given image into overlapping blocks (patches), followed by generating a low-dimensional 2D Discrete Cosine Transform (DCT) based descriptor for each block [8].
2. Classification of each block into foreground or background based on a background model, where each block is sequentially processed by up to three classifiers. As soon as one of the classifiers deems that the block is part of the background, the remaining classifiers are not consulted. In sequential order of processing, the three classifiers are:
   (a) a probability measurement according to a location specific multivariate Gaussian model of the background (i.e., one Gaussian for each block location);
   (b) an illumination robust similarity measurement through a cosine distance metric;
   (c) a temporal correlation check where blocks and decisions from the previous image are taken into account.
3. Model reinitialisation to address scenarios where a sudden and significant scene change can make the current model inaccurate.
4. Probabilistic generation of the foreground mask, where the classification decisions for all blocks are integrated. The overlapping nature of the analysis is exploited to produce smooth contours and to minimise the number of errors (both false positives and false negatives).

Parts 2(a) and 2(b) require a location specific Gaussian model, which can be characterised by a mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In an attempt to allow the training sequence to contain moving foreground objects, a rudimentary Gaussian selection strategy is employed in [13]. Specifically, for each block location a two-component Gaussian mixture model (GMM) is trained, followed by taking the absolute difference of the weights of the two Gaussians. If the difference is greater than $0.5$, the Gaussian with the dominant weight is retained. The reasoning is that the less prominent Gaussian is modelling moving foreground objects and/or other outliers. If the difference is less than $0.5$, it is assumed that no foreground objects are present and all available data for that particular block location is used to estimate the parameters of the single Gaussian.

There are several problems with the above parameter selection approach. It is assumed that foreground objects are either continuously moving in the sequence or that no object stays in one location for more than 25% of the length of the training sequence. This is not guaranteed to occur in uncontrolled environments such as railway stations. The decision to retain the dominant Gaussian solely relies on local temporal statistics and ignores rich local spatial correlations that naturally exist within a scene.

To address the above problems, we propose to estimate the parameters of the background model via a Markov Random Field (MRF) framework, where in addition to temporal information, spatial continuity of the entire background is considered. The details of the MRF-based algorithm are given in the following section.

## 3. Proposed Background Initialisation

Let the resolution of the image sequence $I$ be $\mathcal{W} \times \mathcal{H}$, with $\phi$ colour channels. The proposed algorithm has three main stages: **(1)** division of each frame into non-overlapping blocks and collection of possible background blocks over a given time interval, **(2)** partial background reconstruction using unambiguous blocks, **(3)** ambiguity resolution through exploitation of spatial correlations across neighbouring blocks. An example of the algorithm in action is shown in Fig. 1. The details of the three stages are given below.

In stage 1, each frame is viewed as an instance of an undirected graph, where the nodes of the graph are blocks of size $N \times N \times \phi$ pixels[1]. We denote the nodes of the graph by $\mathcal{N}(i,j)$ for $i = 0,1,2,\cdots,(\mathcal{W}/N)-1$, $j = 0,1,2,\cdots,(\mathcal{H}/N)-1$. Let $I_f$ be the $f$-th frame of the training image sequence and let its corresponding node labels be denoted by $\mathcal{L}_f(i,j)$, and $f = 1,2,\cdots,F$, where $F$ is the total number of frames. For convenience, each node label $\mathcal{L}_f(i,j)$ is vectorised into an $\phi N^2$ dimensional vector $\mathbf{l}_f(i,j)$. In comparison to pixel-based processing, block-based processing is more robust against noise and captures better contextual spatial continuity of the background.

At each node $(i,j)$, a representative set $\mathcal{R}(i,j)$ is maintained. It contains only unique representative labels, $\mathbf{r}_k(i,j)$ for $k = 1,2,\cdots,S$ (with $S \leq F$) that were obtained along its temporal line. To determine uniqueness, the similarity of

[1] For implementation purposes, each block location and its instances at every frame are treated as a node and its labels, respectively.

labels is calculated as described in Section 3.1. Let weight $W_k$ denote the number of occurrences of $\mathbf{r}_k$ in the sequence, i.e., the number of labels at location $(i,j)$ which are deemed to be the same as $\mathbf{r}_k(i,j)$.

It is assumed that one element of $\mathcal{R}(i,j)$ corresponds to the background. To ensure labels corresponding to moving objects are not stored, label $\mathbf{b}_f(i,j)$ will be registered as $\mathbf{r}_{k+1}(i,j)$ only if it appears in at least $f_{min}$ consecutive frames, where $f_{min}$ ranges from 2 to 5.

In stage 2, representative sets $\mathcal{R}(i,j)$ having just one label are used to initialise the corresponding node locations $\mathcal{B}(i,j)$ in the background $\mathcal{B}$.

In stage 3, the remainder of the background is estimated iteratively. An optimal labelling solution is calculated by considering the likelihood of each of its labels along with the *a priori* knowledge of the local spatial neighbourhood modelled as an MRF. Iterated conditional mode (ICM), a deterministic relaxation technique, performs the optimisation.

The MRF framework is described in Section 3.2. The strategy for selecting the location of an empty background node to initialise a label is described in Section 3.3. The procedure for calculating the energy potentials, a prerequisite in determining the *a priori* probability, is described in Section 3.4. In Section 3.5, the background model (used by the foreground segmentation algorithm overviewed in Section 2) is modified using the estimated background frame.

### 3.1. Similarity Criteria for Labels

Two labels $\mathbf{l}_f(i,j)$ and $\mathbf{r}_k(i,j)$ are similar if the following two constraints are satisfied:

$$\frac{\left(\mathbf{r}_k(i,j) - \mu_{r_k}(i,j)\right)' \left(\mathbf{l}_f(i,j) - \mu_{b_f}(i,j)\right)}{\sigma_{r_k}\sigma_{b_t}} > \mathcal{T}_1 \quad (1)$$

and

$$\frac{1}{\phi N^2} \sum\nolimits_{n=0}^{\phi N^2 - 1} |d_{k_n}(i,j)| < \mathcal{T}_2 \quad (2)$$

where $\mu_{r_k}, \mu_{l_f}$ and $\sigma_{r_k}, \sigma_{l_f}$ are the mean and standard deviation of the elements of labels $\mathbf{r}_k$ and $\mathbf{l}_f$ respectively, while $\mathbf{d}_k(i,j) = \mathbf{l}_f(i,j) - \mathbf{r}_k(i,j)$.

Eqns. (1) and (2) respectively evaluate the correlation coefficient and the mean of absolute differences (MAD) between the two labels. The former constraint ensures that labels have similar texture/pattern while the latter one ensures that they are close in $\phi N^2$ dimensional space. In contrast, we note that in [6] the similarity criteria is based just on the sum of squared distances between the two blocks.

$\mathcal{T}_1$ is selected empirically (typically 0.8), to ensure that two visually identical labels are not treated as being different due to image noise. $\mathcal{T}_2$ is proportional to image noise.

### 3.2. Markov Random Field (MRF) Framework

MRF has been widely employed in solving problems in image processing that can be formulated as labelling problems [3, 16].

Let $\mathbf{X}$ be a 2D random field, where each random variate $X_{(i,j)}$ $(\forall\, i,j)$ takes values in discrete *state space* $\Lambda$. Let $\omega \in \Omega$ be a *configuration* of the variates in $\mathbf{X}$, and let $\Omega$ be the set of all such configurations. The joint probability distribution of $\mathbf{X}$ is considered Markov if

$$p(\mathbf{X} = \omega) > 0,\ \forall\ \omega \in \Omega \quad (3)$$

and

$$p\left(X_{(i,j)} | X_{(a,b)}, (i,j) \neq (a,b)\right) = p\left(X_{(i,j)} | X_{\mathcal{N}_{(i,j)}}\right) \quad (4)$$

where $X_{\mathcal{N}_{(i,j)}}$ refers to the local *neighbourhood system* of $X_{(i,j)}$.

Unfortunately, the theoretical factorisation of the joint probability distribution of the MRF turns out to be intractable. To simplify and provide computationally efficient factorisation, Hammersley-Clifford theorem [2] states that an MRF can equivalently be characterised by a Gibbs distribution. Thus

$$p(\mathbf{X} = \omega) = e^{-U(\omega)/T} \ / \left(\sum\nolimits_{\omega} e^{-U(\omega)/T}\right) \quad (5)$$

where the denominator is a normalisation constant known as the *partition function*, $T$ is a constant used to moderate the peaks of the distribution and $U(\omega)$ is an *energy function* which is the sum of *clique/energy potentials* $V_c$ over all possible cliques $C$:

$$U(\omega) = \sum\nolimits_{c \in C} V_c(\omega) \quad (6)$$

The value of $V_c(\omega)$ depends on the local configuration of clique $c$.

In our framework, information from two disparate sources is combined using Bayes' rule. The local visual observations at each node to be labelled yield label likelihoods. The resulting label likelihoods are combined with *a priori* spatial knowledge of the neighbourhood represented as an MRF.

Let each input image $I_f$ be treated as a realisation of the random field $\mathcal{B}$. For each node $\mathcal{B}(i,j)$, the representative set $\mathcal{R}(i,j)$ containing unique labels is treated as its *state space* with each $\mathbf{r}_k(i,j)$ as its plausible label[2].

Using Bayes' rule, the posterior probability for every label at each node is derived from the *a priori* probabilities and the observation-dependent likelihoods given by:

$$P(\mathbf{r}_k) = l(\mathbf{r}_k) p(\mathbf{r}_k) \quad (7)$$

[2] To simplify the notations, index term $(i,j)$ has been omitted from here onwards.

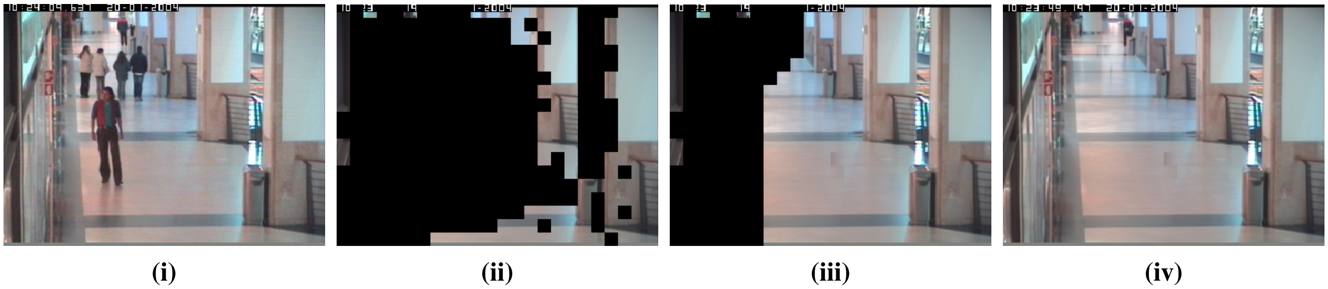

**Figure 1.** Example of background estimation from an image sequence cluttered with foreground objects: **(i)** example frame, **(ii)** partial background initialisation (after stage 2), **(iii)** remaining background estimation in progress (stage 3), **(iv)** estimated background.

The product is comprised of likelihood $l(\mathbf{r}_k)$ of each label $\mathbf{r}_k$ of set $\mathcal{R}$ and its *a priori* probability density $p(\mathbf{r}_k)$, conditioned on its local neighbourhood. In the derivation of likelihood function it is assumed that at each node the observation components $\mathbf{r}_k$ are conditionally independent and have the same known conditional density function dependent only on that node. At a given node, the label that yields maximum *a posteriori* (MAP) probability is chosen as the best continuation of the background at that node.

To optimise the MRF-based function defined in Eqn. (7), ICM is used since it is computationally efficient and avoids large scale effects[3] [3]. ICM maximises local conditional probabilities iteratively until convergence is achieved. In ICM an initial estimate of the labels is typically obtained by maximising the likelihood function. However, in our framework an initial estimate consists of partial reconstruction of the background at nodes having just one label which is assumed to be the background. Using the available background information, the remaining unknown background is estimated progressively (see Section 3.3).

At every node, the likelihood of each of its labels $\mathbf{r}_k$ $(k = 1, 2, \cdots, S)$ is calculated using corresponding weights $W_k$. The higher the occurrences of a label, the more is its likelihood to be part of the background. Empirically, the likelihood function is modelled by a simple weighted function, given by:

$$l(\mathbf{r}_k) = W_{c_k} / \sum_{k=1}^{S} W_{c_k} \qquad (8)$$

where $W_{c_k} = \min(W_{max}, W_k)$. Capping the weight is necessary in circumstances where the image sequence has a stationary foreground object visible for an exceedingly long period.

The spatial neighbourhood modelled as Gibbs distribution (Eqn. (5)) is encoded into an *a priori* probability density. The formulation of the clique potential $V_c(\omega)$ referred in Eqn. (6) is described in the Section 3.4. Using Eqns. (5) and (6) the calculated clique potentials $V_c(\omega)$ are transformed into *a priori* probabilities. For a given label, the smaller the value of energy function, the greater is its probability in being the best match with respect to its neighbours.

In our evaluation of the posterior probability given by Eqn. (7), more emphasis is given to the local spatial context term than the likelihood function which is based on mere temporal statistics. Thus, taking log of Eqn. (7) and assigning a weight to the prior, we get:

$$\log\left(P(\mathbf{r}_k)\right) = \log\left(l(\mathbf{r}_k)\right) + \eta \log\left(p(\mathbf{r}_k)\right) \qquad (9)$$

where $\eta$ has been empirically set to number of neighbouring nodes used in clique potential calculation (typically $\eta$ = 3).

## 3.3. Node Initialisation

Nodes containing a single label in their representative set are directly initialised with that label in the background (see Fig. 1(ii)). However, in rare situations there's a possibility that all sets may contain more than 1 label (no trivial nodes). In such cases, the label having the largest weight from the representative sets of the 4 corner nodes is selected as an initial seed. We assume at least 1 of the corner regions corresponds to a static region. The rest of the nodes are initialised based on constraints as explained below. In our framework, the local *neighbourhood system* [7] of a node and the corresponding cliques are defined as shown in Fig. 2. The background at an empty node will be assigned only if at least 2 neighbouring nodes of its 4-connected neighbours adjacent to each other and the diagonal node located between them are already assigned with background labels. For instance, in Fig. 2, we can assign a label to node $X$ if at least nodes $B$, $D$ (adjacent 4-connected neighbours) and $A$ (diagonal node) have already been assigned with labels. In other words, label assignment at node $X$ is *conditionally independent* of all other nodes given these 3 neighbouring nodes.

Let us assume that all nodes except $X$ are labelled. To label node $X$ the procedure is as follows. In Fig. 2, four cliques involving $X$ exist. For each candidate label at node $X$, the energy potential for each of the four cliques is evaluated independently given by Eqn. (10) and summed together to obtain its energy value. The label that yields the least value is likely to be assigned as the background.

[3] An undesired characteristic where a single label is wrongly assigned to most of the nodes of the random field.

Mandating that the background should be available in at least 3 neighbouring nodes located in three different directions with respect to node $X$ ensures that the best match is obtained after evaluating the continuity of the pixels in all possible orientations.

In cases where not all the three neighbours are available, to assign a label at node $X$ we use one of its 4-connected neighbours whose node has already been assigned with a label. Under these contexts, the clique is defined as two adjacent nodes either in the horizontal or vertical direction.

After initialising all the empty nodes an accurate estimate of the background is typically obtained. Nonetheless, in certain circumstances an incorrect label assignment at a node may cause an error to occur and propagate to its neighbourhood. The problem is successfully redressed by the application of ICM. In subsequent iterations, in order to avoid redundant calculations, the label process is carried out only at nodes where a change in the label of one of their 8-connected neighbours occurred in the previous iteration.

### 3.4. Calculation of the Energy Potential

In Fig. 2, it is assumed that all nodes except $X$ are assigned with the background labels. The algorithm needs to assign an optimal label at node $X$. Let node $X$ have $S$ labels in its state space $\mathcal{R}$ for $k = 1, 2, \cdots, S$, where one of them represents the true background. Choosing the best label is accomplished by analysing the spectral response of every possible clique constituting the unknown node $X$. For the decomposition we chose the Discrete Cosine Transform (DCT) [8] in a similar manner to [12].

We consider the top left clique consisting of nodes $A$, $B$, $D$ and $X$. Nodes $A$, $B$ and $C$ are assigned with background labels. Node $X$ is assigned with one of $S$ candidate labels. For each colour channel $z$, we take the 2D DCT of the resulting clique. The transform coefficients are stored in matrix $\mathbf{T}_k^z$ of size $M \times M$ ($M = 2N$) with its elements referred to as $T_k^z(v, u)$. The term $T_k^z(0, 0)$ (reflecting the sum of pixels at each node) is forced to 0 since we are interested in analysing the spatial variations of pixel values.

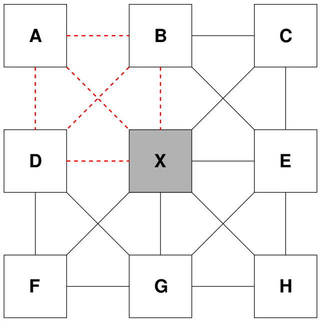

**Figure 2.** The local neighbourhood system and its four cliques. Each clique is comprised of 4 nodes (blocks). To demonstrate one of the cliques, the the top-left clique has dashed red links.

Similarly, for other labels present in the state space of node $X$, we compute their corresponding 2D DCT as mentioned above. A graphical example of the procedure is shown in Fig. 3.

Assuming that pixels close together have similar intensities, when the correct label is placed at node $X$, the resulting transformation has a smooth response (less high frequency components) when compared to other candidate labels.

The energy potential for each label is calculated after summing potentials obtained across the $\phi$ colour channels, as given below:

$$V_c(\omega_k) = \sum\nolimits_{z=1}^{\phi} \left( \sum\nolimits_{v=1}^{M} \sum\nolimits_{u=1}^{M} |T_k^z(v, u)| \right) \quad (10)$$

where $\omega_k$ is the local configuration involving label $k$. The potentials over the other three cliques in Fig. 2 are calculated in a similar manner.

### 3.5. Modified Background Model for Foreground Segmentation

The foreground detection framework described in Section 2 uses a background model comprised of location specific multivariate Gaussians. The background image reconstructed through the MRF-based process is used as follows. First, the dual-Gaussian training strategy used in Section 2 is run on a given training sequence, obtaining the mean vectors and diagonal covariance matrices for each location. The mean vectors are then replaced by rerunning step 1 of the segmentation framework on the estimated background image. The covariance matrices are retained as is. Preliminary experiments indicated that when stationary backgrounds were occluded by foreground objects for a long duration, the variances computed in step 1 were similar to the variances of the true background.

## 4. Experiments

The proposed framework (MRF + multi-stage classifier) was evaluated with segmentation methods based on the baseline multi-stage classifier [13], Gaussian mixture models (GMMs) [17] and feature histograms [10]. In our experiments the same parameter settings were used across all sequences (i.e., they were not optimised for any particular sequence). The block size was set to 16 × 16. The values of $\mathcal{T}_1$ and $\mathcal{T}_2$ (see Eqns. 1 and 2) were set to 0.8 and 3 respectively, while $W_{max}$ (see Eqn. 8) and $T$ (Eqn. 5) were set to 150 and 1024 respectively. The algorithm was implemented in C++ with the aid of the Armadillo library [14].

We used the OpenCV v2.0 [5] implementations for the last two algorithms, in conjunction with morphological post-processing (opening followed by closing using a $3 \times 3$ kernel) in order to improve the quality of the obtained

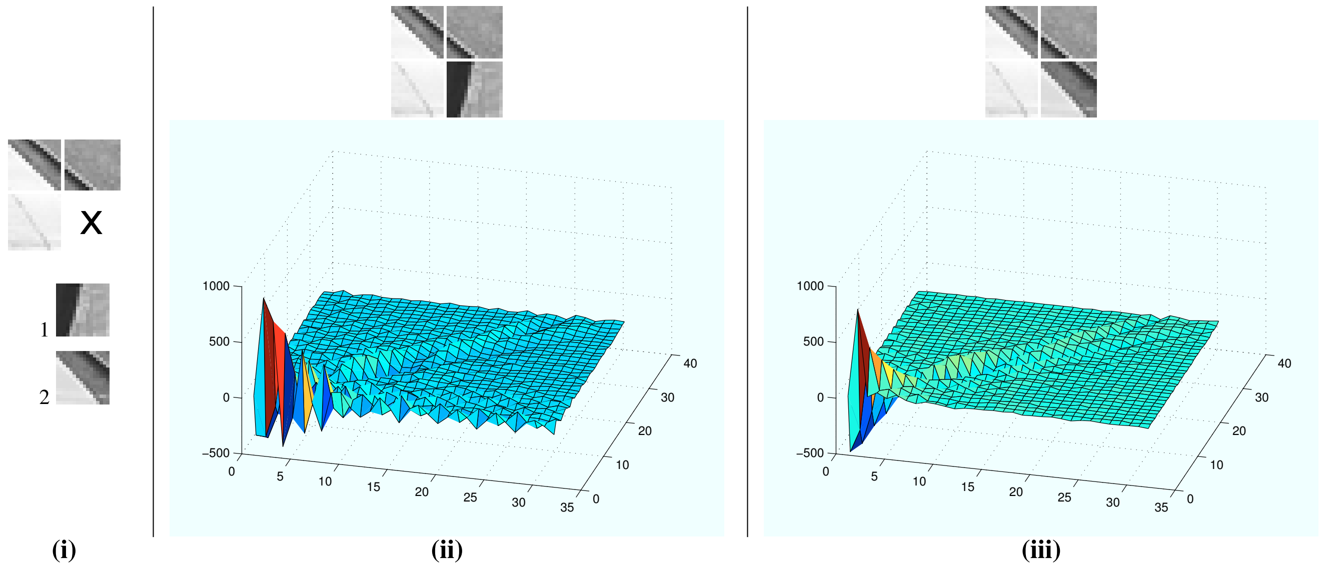

**Figure 3.** An example of the processing done in Section 3.4. **(i)** A clique involving empty node $X$ with two candidate labels in its representative set. **(ii)** A clique and a graphical representation of its DCT coefficient matrix where node $X$ is initialised with candidate label 1. The gaps between the blocks are for ease of interpretation only and are not present during DCT calculation. **(iii)** As per (ii), but using candidate label 2. The smoother spectral distribution for candidate 2 suggests that it is a better fit than candidate 1.

foreground masks [10]. The methods' default parameters were found to be optimal, except for the histogram method, where the built-in morphology operation was disabled as we found that it produced worse results than the above-mentioned opening and closing. We note that the proposed foreground segmentation approach does not require any such ad hoc post-processing.

In our experiments, we studied the influence of the various foreground segmentation algorithms on tracking performance. The foreground masks obtained from the detectors were passed as input to several tracking systems. We used the tracking systems implemented in the video surveillance module of OpenCV v2.0 [5] and the tracking ground truth data that is available for the sequences in the second set of the CAVIAR[4] dataset. We randomly picked 30 sequences from the dataset for our experiments. The tracking performance was measured with two metrics: multiple object tracking accuracy (MOTA) and multiple object tracking precision (MOTP), as proposed by Bernardin and Stiefelhagen [1].

Briefly, MOTP measures the average pixel distance between the ground-truth locations of objects and their locations according to a tracking algorithm. The lower the MOTP, the better. MOTA accounts for object configuration errors, false positives, misses as well as mismatches. The higher the MOTA, the better.

We performed 20 tracking simulations by evaluating four foreground object segmentation algorithms (baseline multi-stage classifier, GMM, feature histogram and the proposed method) in combination with five tracking algorithms (blob matching, mean shift, mean shift with foreground feedback, particle filter, and blob matching with particle filter for occlusion handling). The performance result in each simulation is the average performance of the 30 test sequences. We used the first 200 frames of each sequence for initialising the background model.

Examples of qualitative results are illustrated in Fig. 4. It can be observed that foreground masks generated using methods based on GMMs [17], feature histograms [10], and the baseline multi-stage classifier [13] have considerable false negatives, which are due to foreground objects being included into the background model. In contrast, the MRF based model initialisation approach results in noticeably better foreground detection.

The quantitative tracking results, presented in Fig. 5, indicate that in all cases the proposed framework led to the best precision and accuracy values. For tracking precision (MOTP), the next best method [13] obtained an average pixel distance of 11.03, while the proposed method reduced the distance to 10.28, indicating an improvement of approximately 7%. For tracking accuracy (MOTA), the next best method obtained an average accuracy value of 0.35, while the proposed method achieved 0.5, representing a considerable improvement of about 43%.

[4] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

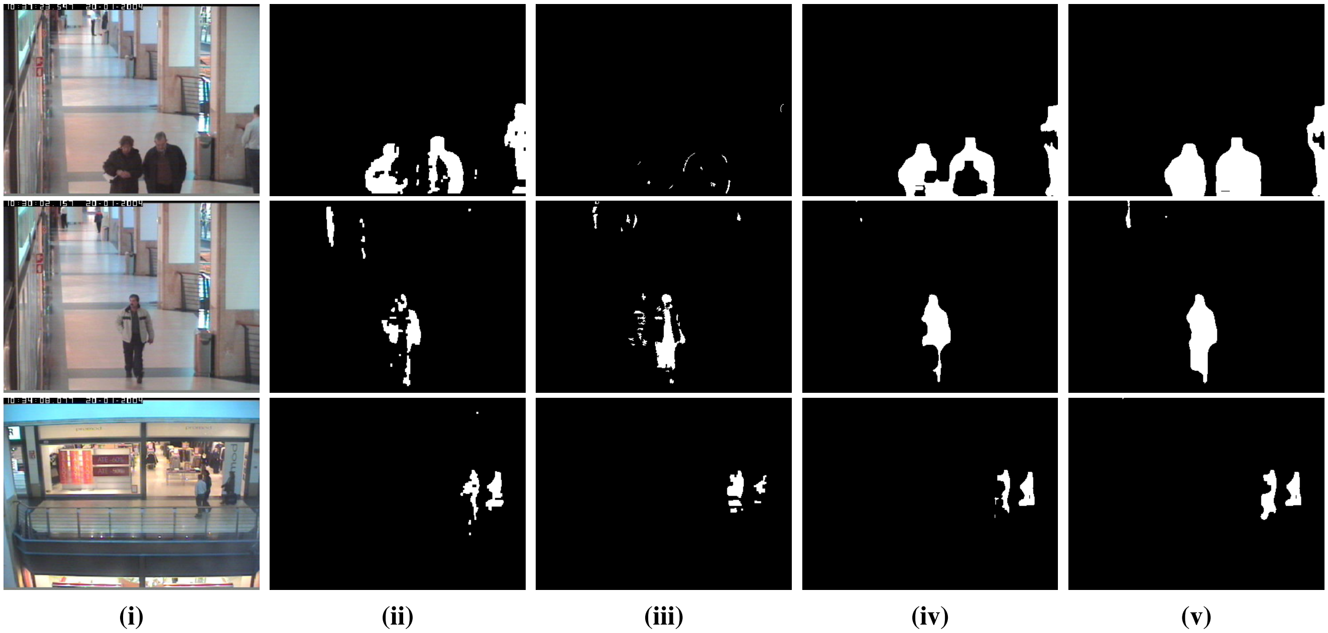

**Figure 4.** **(i)** Example frames from CAVIAR dataset; foreground masks obtained using: **(ii)** GMM based method [17], **(iii)** histogram based method [10], **(iv)** baseline multi-stage classifier [13], **(v)** proposed MRF based framework. We note the masks shown in columns **(ii)** to **(iv)** have considerable amount of false negatives since the foreground objects were included in the background model, while the results of the proposed framework (column **(v)**) have minimal errors.

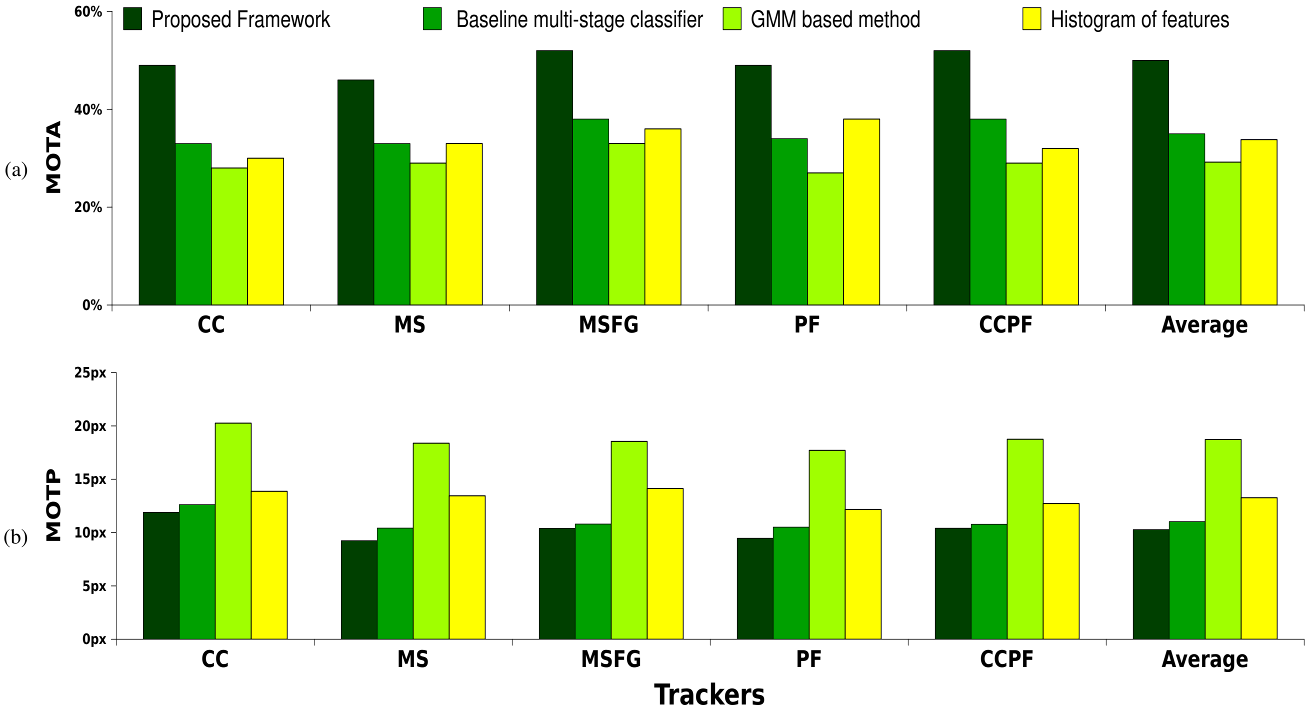

**Figure 5.** Effect of foreground detection methods on: **(a)** multiple object tracking accuracy (MOTA), where taller bars indicate better accuracy; **(b)** multiple object tracking precision (MOTP), where shorter bars indicate better precision (lower distance). Results are grouped by tracking algorithm: blob matching (CC), mean shift trackers (MS and MSFG), particle filter (PF) and hybrid tracking (CCPF).

## 5. Main Findings

In this paper we have proposed a foreground segmentation framework which effectively segments foreground objects in cluttered environments. The MRF-based model initialisation strategy allows the training sequence to contain foreground objects. We have shown that good background model initialisation results in considerably improved foreground detection, which leads to better tracking.

We noticed (via subjective observations) that all evaluated algorithms perform reasonably well when foreground objects are always in motion (i.e., where the background is visible for a longer duration when compared to the foreground). However, accurate estimation by methods solely relying on temporal statistics to initialise their background model becomes problematic if the above condition is not satisfied. This is the main area where the proposed framework is able to detect foreground objects accurately.

A minor limitation exists, as there is a potential to misestimate the background in cases where an occluding foreground object is smooth (uniform intensity value), has intensity value similar to that of the background (i.e., low contrast between the foreground and the background) and the true background is characterised by strong edges. Under these conditions, the energy potential of the label containing the foreground object is smaller (i.e., smoother spectral response) than that of the label corresponding to the true background. This limitation will be addressed in future work.

Overall, the parameter settings for the proposed algorithm appear to be quite robust against a variety of sequences and the method does not require explicit post-processing of the foreground masks. Experiments conducted to evaluate the effect on tracking performance (using the CAVIAR dataset) show the proposed framework obtains considerably better results (both qualitatively and quantitatively) than approaches based on Gaussian mixture models (GMMs) [17] and feature histograms [10].

## Acknowledgements

NICTA is funded by the Australian Government as represented by the *Department of Broadband, Communications and the Digital Economy* as well as the Australian Research Council through the *ICT Centre of Excellence* program.

## References

[1] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the CLEAR MOT metrics. *EURASIP Journal on Image Video Processing*, 2008.

[2] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(2):192–236, 1974.

[3] J. Besag. On the statistical analysis of dirty pictures. *Journal of Royal Statistics Society*, 48(3):259–302, 1986.

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *Proc. Intl. Conf. Computer Vision (ICCV)*, volume 1, pages 377–384, 1999.

[5] G. Bradski and A. Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, 2008.

[6] A. Colombari, A. Fusiello, and V. Murino. Background Initialization in Cluttered Sequences. In *CVPRW*, pages 197–202, Washington DC, USA, 2006.

[7] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, pages 721–741, 1984.

[8] R. Gonzales and R. Woods. *Digital Image Processing*. Prentice Hall, 3rd edition, 2007.

[9] K. Kim, T. Chalidabhongse, D. Harwood, and L. Davis. Real-time foreground–background segmentation using codebook model. *Real-Time Imaging*, 11(3):172–185, 2005.

[10] L. Li, W. Huang, I. Gu, and Q. Tian. Foreground object detection from videos containing complex background. In *ACM Int. Conf. Multimedia*, pages 2–10, 2003.

[11] L. Maddalena and A. Petrosino. A Self-Organizing Approach to Background Subtraction for Visual Surveillance Applications. *IEEE Trans. Image Processing*, 17:1168–1177, 2008.

[12] V. Reddy, C. Sanderson, and B. Lovell. An efficient and robust sequential algorithm for background estimation in video surveillance. In *Proc. Int. Conf. Image Processing (ICIP)*, pages 1109–1112, 2009.

[13] V. Reddy, C. Sanderson, and B. Lovell. Improved foreground detection via block-based classifier cascade with probabilistic decision integration. *IEEE Transactions on Circuits and Systems for Video Technology*, (in press). http://dx.doi.org/10.1109/TCSVT.2012.2203199 .

[14] C. Sanderson. Armadillo: An open source C++ linear algebra library for fast prototyping and computationally intensive experiments. Technical report, NICTA, 2010.

[15] A. Sanin, C. Sanderson, and B. Lovell. Shadow detection: A survey and comparative evaluation of recent methods. *Pattern Recognition*, (in press). http://dx.doi.org/10.1016/j.patcog.2011.10.001 .

[16] Y. Sheikh and M. Shah. Bayesian Modeling of Dynamic Scenes for Object Detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(11):1778–1792, 2005.

[17] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 246–252, 1999.

[18] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *Proc. Int. Conf. Computer Vision (ICCV)*, volume 1, pages 255–261, 1999.

[19] H. Wang and D. Suter. A Novel Robust Statistical Method for Background Initialization and Visual Surveillance. *ACCV 2006, Lecture Notes in Computer Science*, 3851:328–337, 2006.