Articulated Human Body Parts Detection Based on Cluster Background Subtraction and Foreground Matching

Harish Bhaskar[a], Lyudmila Mihaylova[b], Simon Maskell[c]

[a] *Dept. of Computer Engineering, Khalifa University, U.A.E.*
[b] *School of Computing and Communications, Lancaster University, U.K.*
[c] *Qinetiq, Malvern, U.K.*

Abstract

Detecting people or other articulated objects and localizing their body parts is a challenging computer vision problem as their movement is unpredictable under circumstances of partial and full occlusions. In this paper, a framework for human body parts tracking in video sequences using a self-adaptive *cluster* background subtraction (CBS) scheme is proposed based on a Gaussian mixture model (GMM) and foreground matching with rectangular pictorial structures. The efficiency of the designed human body parts tracking framework is illustrated over various real-world video sequences.

Keywords: human target tracking, background subtraction, optimisation, genetic algorithm, pictorial structures

1. Introduction and Related Work

Detection and tracking are critical tasks of any modern day visual tracking system. The core of any detection algorithm requires finding a clear distinction between the foreground and background regions. The literature on detection models has been focused on two fundamental categories of background subtraction and foreground learning models. The effectiveness of a reliable detection system relies on efficiently combining these background and foreground models in such a way that the system is accurate, robust and invariant to the presence of clutter and camera motion. In order to motivate the proposed model, a brief review of some of the important methods are presented in the subsections below.

Preprint submitted to NEUROCOMPUTING, Special issue on "Behaviours in Video", November 30, 2011

1.1. Background Subtraction

The simplest mechanism for accomplishing detection is through building a representation of the scene background and comparing the new frame with this representation. This procedure is known as *background subtraction* (BS) [1, 2, 3, 4, 5]. The various BS approaches differ in the ways of modeling the background and their learning. In the last several years, a number of different BS techniques have been proposed in the literature. These techniques include: the basic BS technique, extended basic BS using *average* or *median* of pixels from previous frames [6], pixel level Gaussian mixture models (GMMs) models [7, 8], *kernel density estimators* (KDEs) [9, 10] and *mean-shift estimators* [11].

Though these BS techniques satisfies the requirements of detection in some applications, they are limited in different ways. These modeling schemes are restricted especially in explicitly handling dynamic changes: gradual or sudden (e.g., moving clouds), camera motion (oscillations), background changes during the detection process, including tree branches, sea waves, etc., and changes in the background geometry such as parked cars [4]. These techniques also require specifying appropriate thresholds and are highly sensitive to the presence of clutter, movement of the camera etc.

Whilst building representations of the background of a scene helps the process of target detection, a learning model representing the foreground regions, e.g., the moving objects, is essential for efficient tracking. In the next subsection, a review of related approaches for foreground learning is presented.

1.2. Foreground Learning Models

Foreground modelling is concerned with forming 'blobs' around objects of interest. An important class of methods is concerned with splitting complex objects into separate simpler parts (rectangles for instance), modelling each part and then connecting each part into a whole structure to form the object. These methods called pictorial structures became quite popular and have shown efficiency in many image and video processing applications.

Pictorial Structure Methods

Many foreground learning models are motivated by the pictorial structure representation introduced in [12]. The procedure of pictorial structure matching generally consists of two main phases of *feature extraction* and *matching.*

In the first stage, discrete primitives, or features are detected. Features such as colour in combination with the object size can be used [13]. In the second stage, models are matched against those features using search techniques, e.g., genetic algorithms (GA) [14]. According to the pictorial structure model proposed in [12], an object is modeled by a collection of parts arranged in a deformable configuration. This representation allows encoding different parts of the body with the local properties of the object whilst, the deformable configuration characterises spring-like movements. The second phase of matching these pictorial structures to an image is typically done through minimising cost functions such as the Mahanalobis distance and energy for every part.

The pictorial structure representation is an appealing framework for foreground modeling in view its simplicity and generality, but its application to automatic detection has been limited due to the following reasons [15]:

- the model and its parameters are specific to different objects, and it is often hard to give general guidelines how to choose them;
- the resulting energy minimisation problem is highly complex and requires efficient techniques for real-time applicability;
- the matching process can contain a solution space with many outcomes resulting in ambiguities.

Other methods using pictorial structures are [16, 12] and [17, 18, 19]. Results achieved using pictorial structures even without subtracting the background are achieved in [20, 21].

Appearance-Based and Template Matching Methods

The inherent disadvantages of the feature-based techniques lead to the development of *appearance-based methods* (e.g., [22] and [23]) and *template matching methods* [24]. Such approaches treat objects in images as entities to be recognised, rather than having more abstract models based on features or other primitives. The general idea behind this class of techniques is to use a template and compare it with new images to determine whether or not the target is present, generally by explicitly considering possible transformations of the template.

Part-based Methods
Different *part-based methods* have been proposed. These techniques generally combine the appearance of the individual parts, spatial and geometrical constraints while matching. However, most of these part-based methods make binary decisions about potential part locations [25, 26, 27, 28].

In [29] the problem of finding people in images using *coarse part-based two-dimensional models* is considered. A two-stage process for pictorial people detection is proposed. In the first stage, binary decisions about the possible locations for individual parts are made and subsequently search for groups of parts that match the overall model. In the second level, a sequential importance sampling (particle filtering) mechanism is used to generate increasingly larger configurations of parts. The method from [29] provides efficient computation of the exact (discrete) posterior distribution for the object configuration and then sampling from that posterior PDF is therefore superior to other methodologies.

Deformable Matching Techniques
Silhouette-based deformable matching techniques have been proposed that match binary images obtained from BS to single parts [30, 31]. Models of pictorial structures have recently been used for tracking people by matching models when the cost function relies on the Chamfer distance function at each frame [18]. A great number of works on highly articulated object tracking such as people employ predominantly motion information [32, 33] and only performs incremental updates in the object configuration. These approaches perform an match initially the model to the image, and then tracking commences from that initial condition. Pictorial structures can be used to solve this track initialisation problem, or as demonstrated in [18] can be a tracking method on their own.

Some top-down random approaches are also proposed, based on Markov Chain Monte Carlo methods [34] and particle filtering [35].

Matching Cost Functions
Choosing reliable and robust cost functions for matching problems is often difficult and the reason is that this choice depends on the application area, on the features employed and other statistical properties. Some of the commonly used cost functions are Mahalanobis distance [12], Euclidean distance [36], Chamfer distance [37, 18] and energy minimisation using generalised

distance transform [16]. In [38], a comparison of different cost functions for stereo matching problems is presented. From their analysis, the authors have concluded that the cost function relies on the chosen technique for matching and that none of the cost functions could efficiently handle abrupt dynamic changes. In a similar study, the authors of [39] compare six different cost functions based on evaluations of gray value matching of 2D images in radiotherapy. This study also confirms that the application of a particular cost function largely depends on the target application and the features employed.

2. Our Approach: Combining CBS-GMM with Foreground Learning

The main idea behind our approach is to enhance the accuracy of the BS technique by learning the appearance models for the moving objects (the so-called foreground). The main contributions of the work are:

- To propose a novel cluster background subtraction model combined with foreground matching for automatic detection of human body parts
- To extend the cluster background subtraction framework proposed using dipping threshold feedback technique for adaptive parameter estimation.
- To model the evolutionary algorithm based body parts matching technique with a posterior based cost function for accurate measurement of the matching error between the template and the target body parts.
- To compare the effectiveness of the proposed model against state-of-the-art baseline systems
- To test the robustness of the detection framework on moving camera sequences

A block diagram illustrating the novel CBS-GMM method and the evolutionary pictorial matching scheme is presented in Figure 1.

The proposed automatic detection framework is composed of two phases. In the first phase, the image is clustered into regions according to colour feature and the CBS-GMM is performed. In the second stage, a pictorial structure rectangular human body model is matched to the background subtracted outputs of phase one using an evolutionary search strategy. The developed

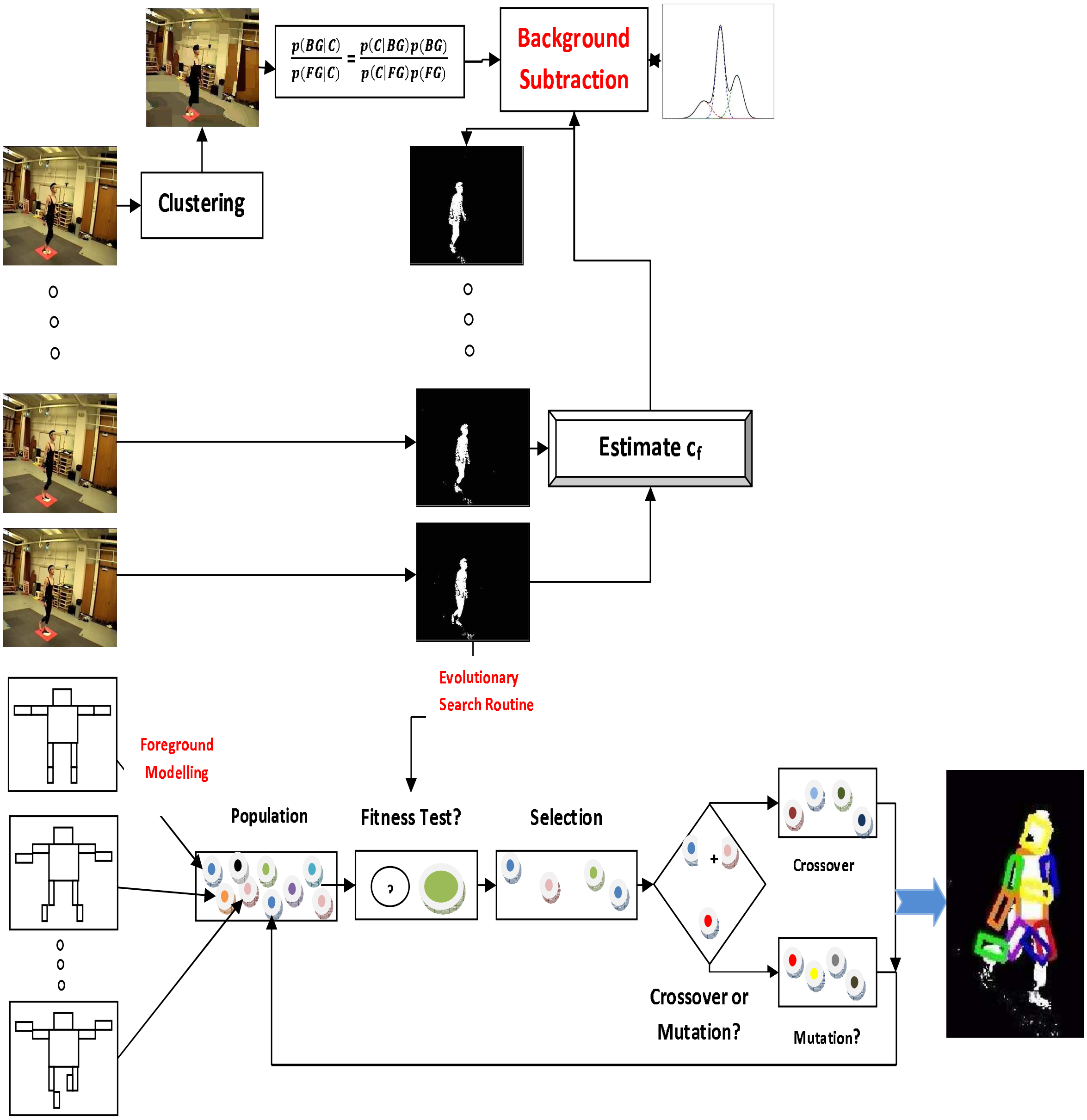


Figure 1: Block diagram of the body-part detection framework

CBS involves clustering image frames from the video sequence and further applies an adaptive Gaussian mixture model for reliable CBS. The CBS-GMM technique automatically adapts the parameters of the GMM including the detection threshold from the outputs of the previous processed frames (as indicated by the block estimate c_f) in contrast to [7] where some parameters of the pixel GMM are chosen heuristically.

The proposed CBS-GMM scheme is much faster and more accurate than other BS techniques operating at pixel level which makes it suitable for real-time applications. This is because, in the pixel-based methods, pixels belonging to any particular region of an image have higher variability. This can lead to wrong classification of certain pixels within the same region, whereas the CBS reduces the pixel variability. Foreground learning is achieved by body-part matching using two different cost functions for the pictorial structure approach. The efficiency of the designed techniques is illustrated over various real-world video sequences static and moving cameras. The robustness of the proposed approach is proven as it reduces the clutter and achieves high level of precision.

2.1. Update of the GMM Parameters at Cluster Level

The problem of *cluster* BS (CBS) 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 $[Im]_k = [\boldsymbol{c}_k^1, \ldots, \boldsymbol{c}_k^q]$ represents the whole image; $p(bG|\boldsymbol{c}_k^i)$ is the PDF of the background, subtracted based on colour feature (though this can be generalised to other features such as texture, edges or combination) 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 CBS 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 [9, 7] take this decision at pixel level in contrast 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: $T = N/fps$, e.g., as a ratio between the total number of frames of the video sequence, N and the frame rate, fps, frames per second. At time instant k there is $\Re_k = \{\boldsymbol{c}_k^i, ..., \boldsymbol{c}_{k-T}^i\}$.

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 [40], 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, saturation, value* (HSV) colour model due to its inherent ability to cope with illumination changes. A detailed algorithm describing the clustering procedure can be found in [41].The results obtained from the clustering method suggested in [41] can vary in accordance with the chosen features.

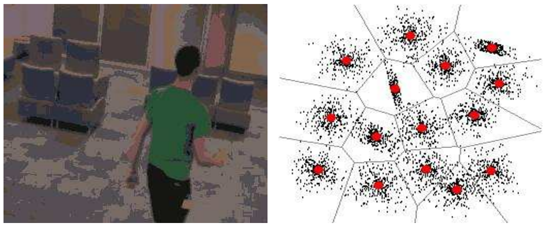


Figure 2: a) An image with clusters b) Clusters with their centres

A GMM, containing M components, is then used to represent the density distribution

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

both of the background and foreground where $\widetilde{\boldsymbol{\mu}}_{1,k}, ..., \widetilde{\boldsymbol{\mu}}_{M,k}$ and $\widetilde{\sigma}^2_{1,k}, ..., \widetilde{\sigma}^2_{M,k}$ are the estimates of the mean vectors and of the variances that describe the Gaussian components; $\boldsymbol{I}$ is the identity matrix. The estimated mixing weights $\widetilde{\pi}_m$ sum up to one. Given the new cluster $\boldsymbol{c}^i_k$ at time instant k, the update equations for the cluster parameters can be calculated as follows:

$$\widetilde{\pi}_{m,k+1} = \widetilde{\pi}_{m,k} + \frac{1}{T_k}(o_{m,k} - \widetilde{\pi}_{m,k}), \tag{4}$$

$$\widetilde{\boldsymbol{\mu}}_{m,k+1} = \widetilde{\boldsymbol{\mu}}_{m,k} + o_{m,k}\left(\frac{1}{T_k\widetilde{\pi}_{m,k}}\right)\boldsymbol{\delta}_{m,k}, \tag{5}$$

$$\widetilde{\sigma}^2_{m,k+1} = \widetilde{\sigma}^2_{m,k} + o_{m,k}(\frac{1}{T_k\widetilde{\pi}_{m,k}})(\boldsymbol{\delta}'_{m,k}\boldsymbol{\delta}_{m,k} - \sigma^2_{m,k}), \tag{6}$$

where

$$o_{m,k} = \begin{cases} 1, & \text{if the cluster centre is close to the mean} \\ & \text{of the particular GMM component;} \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

$\boldsymbol{\delta}_{m,k} = \boldsymbol{c}^i_k - \widetilde{\boldsymbol{\mu}}_{m,k}$, $'$ 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 if the Mahalanobis distance between the 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,

$$\widetilde{p}(\boldsymbol{c}^i_k|\Re_k, bG) \sim \sum_{m=1}^{B} \widetilde{\pi}_{m,k}\mathcal{N}(\widetilde{\boldsymbol{\mu}}_k, \widetilde{\sigma}^2_m\boldsymbol{I}), \tag{8}$$

$$B = \underset{b}{\text{argmin}}(\sum_{m=1}^{b} \widetilde{\pi}_{m,k} > (1 - c_f)), \tag{9}$$

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.

2.2. Dipping Detection Threshold

Most practical detection systems generally operate with the assumption that the proportionality between the pixels belonging to the foreground to the pixels from the background is assumed constant [7]. Therefore a fixed threshold is used as the detection threshold. An alternative to this is to employ the ratio of prior probabilities of the background and foreground regions based on the information from the training set. This allows a two way communication and perhaps a more appropriate "feedback" mechanism that enables the detection threshold to be decreased near where a target is expected to be and elevated where it is unexpected. This ratio defining the percentage of foreground and background pixels, i.e., the posterior probabilities can be updated from the training set as follows:

$$c_f = \frac{\widetilde{p}(\boldsymbol{c}_k^i|\Re_k, fG)}{\widetilde{p}(\boldsymbol{c}_k^i|\Re_k, bG)}. \tag{10}$$

The cluster background subtraction presents a number of advantages as against the pixel based method. In particular, the CBS-GMM algorithm can explicitly handle 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). The model can also reduce clutter and can operate faster and accurately than the pixel based methods.

The CBS-GMM is able to clearly segment the moving human targets from the background. However, in order to localise the human body parts, a model based on pictorial structures is constructed and matched using an efficient search mechanism. In the following section the pictorial structure model and the evolutionary matching algorithm is described.

2.3. Pictorial Structure

The considered pictorial structure model for an object is given by a collection of parts with connections between certain pairs of parts. More specifically, for the human body, the parts can correspond to the head, torso, arms and legs. The number of parts required to model well the object, depends on the application and the level of accuracy required. For example, whilst a 10 parts (head, torso, 4 arm parts and 4 leg parts) human body model can provide more accurate results, the time complexity of matching the 10 parts body model with each current frame is quite high. Throughout this paper, it is assumed that the human silhouette is represented by a model of maximum of 10 rectangular regions of specific dimensions unless specified otherwise. The simplest form of representing these parts as a structure is using a form of an undirected graph $G = (\mathbf{V}; \mathbf{E})$, where $\mathbf{V} = (v_1, v_2, ..., v_{10})$ is the set of 10 parts of the body and there is an edge between the connected parts in the graph; $\mathbf{E}$ is the set comprising the connecting edges. In each time instance a human body in any scene can be expressed in terms of a configuration $\mathbf{L}$ of different parts containing spatio-orientation parameters $l_i, \quad 1 \leq i \leq 10$. Therefore, a 10 body part model will contain a configuration vector $\mathbf{L} = [l_1, l_2, ..., l_{10}]$. The foreground modeling problem deals with identifying an optimal configuration vector $\mathbf{L}$ that accurately matches the pictorial structure template to human objects in a real-time video sequence.

The problem of matching a pictorial structure to an image can be defined in terms of minimisation of a cost function. The cost of a particular configuration depends both on how well the parts match to the image data and on how well the different parts relate to one another. According to [12], an optimal configuration that accurately matches the parts of the human body model to the image data is defined as

$$\mathbf{L}^* = argmin_{\mathbf{L}} \left(\sum_{i=1}^{10} m_i(l_i) + \sum f_{ij}(l_i, l_j) \right), \tag{11}$$

where $m_i(l_i)$ measures the *degree of mismatch* of any part $1 \leq i \leq 10$ at configuration l_i and f_{ij} is a *function of deformation* relating one part i to another part j. In [12, 16, 15], $f_{ij}(l_i, l_j)$ is defined as the weighted Mahalanobis distance between the transformed locations.

In this paper, an evolutionary strategy of searching parts of a human body on already background subtracted images is proposed. This procedure allows simultaneous learning of model parameters while achieving efficient cost minimisation. A detailed description of the procedure is illustrated in Section 2.4.

2.4. Searching for Parts

Matching of the body parts to the image is performed using a stochastic evolutionary algorithm [14]. The matching process is designed to determine the location and orientation of the body parts with specific constraints. These constraints, for example, can involve finding hands of people around the region of the upper torso. The algorithm is instantiated with a configuration vector $\mathbf{L}$ of the location and orientation parameters of the different body parts that include the pixel displacement value in x and y directions of the **torso**, together with the four parameters of transformation $(a_{11}, a_{12}, a_{21}, a_{22})$ of every body parts $\wp$ of the human target τ encoded as the chromosome $(T_x, T_y, a_{11}^{\wp,\tau}, a_{12}^{\wp,\tau}, a_{21}^{\wp,\tau}, a_{22}^{\wp,\tau})$.

Note: A large population of possible solutions $\mathbf{S} = [L_1, L_2, L_3, ...]$ if generated for every target τ within the bounds of the image dimensions and in the neighbourhood of the segmented regions from the previous step. An iterative procedure is then applied to obtain the optimal match of the pictorial structure of each target to the image. The stages of the iterative process include:

- *Preprocessing stage:* In the pre-processing stage, for all configurations in the solution space, the spatial location of every other part of the human body are estimated in relation to the spatial position (x_t, y_t) of the torso that is obtained from the configuration vector. These pivot points are the points that would attach the different human body parts to the torso or the relevant neighbouring part (such as lower arm and upper arm). The outcome of the preprocessing phase generates for every body part, the corners' positional coordinates of the rectangular regions surrounding each part. The region b_{target} corresponding to the body parts is then found from the target background subtracted image.

- *Cost Estimation:* In order to evaluate the fitness function on each chromosome:

  - The pictorial structure model corresponding to target τ is centred at the spatial location of the torso as extracted from the chromosome using the translation parameters (T_x, T_y). Using this location of the torso, all the pivot points corresponding to the different body parts of the template are estimated.
  - To each body part $\wp$ of target τ, the Affine Transform is applied using the affine parameters $(a_{11}^{\wp,\tau}, a_{12}^{\wp,\tau}, a_{21}^{\wp,\tau}, a_{22}^{\wp,\tau})$ using,

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_{11}^{\wp,\tau} & a_{12}^{\wp,\tau} \\ a_{21}^{\wp,\tau} & a_{22}^{\wp,\tau} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{12}$$

For each of these configurations, a cost function is further employed to determine the deviation of the target body part and the template body parts.

- *i*) The cost function is the mean absolute difference (MAD) between the target template b_{target} and the body part template $b_{template}$ and is computed in the following way

  $$m_i = |b_{template}\wp - b_{target}\wp|, \tag{13}$$

  where $\wp$ refers to the different parts and $b_{template}\wp$ are rectangular regions of specified dimensions. The absolute difference between the template and the target is averaged across all pixels. It is important to note that the number of targets and template regions correspond to the number of considered body parts. The vector cost function comprises the errors for every body part represented as $\boldsymbol{e} = [m_1, m_2, ...]$.

- *ii*) As an alternative to the absolute difference error, the posterior based error function illustrated in [42] is used. According to this function, it is assumed that the vector $\boldsymbol{D}$ of pixels of the image, is a nonlinear function of the configuration parameters, but a linear function of the template, $\boldsymbol{G}$, plus a zero-mean Gaussian noise.

  $$\boldsymbol{D} = \boldsymbol{AG} + \boldsymbol{Noise}. \tag{14}$$

  From this model, by assuming a uniform prior on $\boldsymbol{A}$ and a Jeffrey's prior on the variance of the Gaussian noise, the following error function is derived.

  $$\boldsymbol{e} = \frac{1}{|\boldsymbol{G}'\boldsymbol{G}|} \times (\boldsymbol{D}'\boldsymbol{D} - \boldsymbol{D}'\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{G})^{-1}\boldsymbol{G}'\boldsymbol{D})^{-[\frac{MN}{2}+1]} \tag{15}$$

  where $\boldsymbol{D}$ is a vector of pixels values of the template of size $M \times N$ and $|.|$ means the absolute value.

- *iii*) Although the above criterion would be efficient in matching the template to the target, it does not guarantee that the boundaries of the pictorial structure template overlaps the target. The error returned by the above functions would result in a minimised value even in the case when the template is scaled smaller to lie totally within the boundaries of the target. Therefore, in addition to measuring the deviation of the target region from the template of each body from the pictorial structure as above, the Hausdorff distance between the edge pixels of the body part template $e_{template}\wp$ and that of the target template $e_{target}\wp$ is also computed. The edges of the template and target are generated on the binary image of the segmented target using a simple canny operator. The Hausdorff distance between the two sets of edge pixels is defined as:

$$\begin{split} &H(e_{template}\wp, e_{target}\wp) = \\ &max(h(e_{template}\wp, e_{target}\wp), h(e_{target}\wp, e_{template}\wp)), \end{split} \tag{16}$$

where

$$h(e_{template}\wp, e_{target}\wp) = max_{a\in e_{template\wp}} min_{b\in e_{target}\wp}||a-b||.$$

  - *iv*) Finally, an additional constraint if enforced on the pairwise component f_{ij} of the objective function (11) between neighbouring body parts i and j. The computation of the pairwise condition automatically introduces a penalty on two body parts i and j as a function of their extent of overlap with each other measured as O(.). At this point it is critical to note that overlaps between body parts can also be valid if in the case that the body parts are undergoing partial or total occlusion. However, it is often very hard to autonomously distinguish between truly occluded body parts and a misfit of the search algorithm. In order to help distinguish between the two, a cascading approach to fitting different number of body parts for each target in the image is proposed. This cascading approach will be explained in detail in subsection 2.5.

- *Population Regeneration:* The estimation of the cost for every configuration in S through (13) allows a better prediction of the best configuration suited for the human body parts. However, obtaining a perfect match of the body model configuration, i.e, $\boldsymbol{e} = 0$ is rather difficult unless the solution space is assumed to be very large. In order to sequentially find a global solution, an evolutionary process can be used. In this work, two forms of population regeneration are introduced, one based on combining the goodness of low cost configurations and the other based on local neighbourhood.

  - Matching of the two configurations is performed by comparing the costs and by generating two new configurations with one containing the location and orientation parameters of parts that returned the lowest cost and the other containing the remainder of the parameters. The top performing candidates of the population are selected into this procedure. A hybrid member of the population is generated from the two candidate members by comparing the error scores. The output member of the population will characterise how well the two candidates match to each other.
  - The second form of population regeneration involves discarding configurations that have high values of cost and creating a new set of configurations within a specified region around configurations with low cost.

- *Termination:* The procedure is tested under three main stopping criteria including the *Zero Cost*, *Maximum Iterations* and *Number of Stall Iterations*. The zero cost condition is satisfied when a particular configuration vector returns an absolute match with zero error. The maximum generations condition is set when the number of iterations exceeds a predefined threshold. Stall generations refers to the consecutive iterations where the cost values remains unchanged. If the number of the stall iteration exceeds a predefined threshold, then the procedure terminates. If the event of the procedure is not satisfying any of the above conditions, the iteration continues.

2.5. Cascaded Implementation

As mentioned earlier, one of the main aspects of our objective function is to have the ability to distinguish between a true occlusion of body parts against a mismatch of body parts by the search algorithm. This is mainly due to the penalty factor introduced as a measure the pairwise constraint based on an overlap function of the two considered body parts. The proposed GA based search algorithm is extended to work in an cascaded manner, first localizing a small but well optimised subset of body parts such as the head and torso which are then used to initialize the subsequently matching containing a increased set of other body parts. At the first level of the cascade, matching is performed on the segmented binary image of the target obtained from the background subtraction algorithm. However, in subsequently higher levels, the matching of other body parts is performed on the gray-scale version of the original image. The number of levels of the cascade is automatically deduced as a function of the overlap between body parts. That is, the search mechanism would terminate when it is able to adequately describe the segmented target with the most optimum number of body parts without overlapping regions. For example, when matching a human target, a minimum 2-parts cascaded model is used at the lowest level of the cascade and incremented subsequently to a maximum 10-body parts model for robust and accurate matching. The effect of the cascaded implementation is studied in detail in the following section 3.

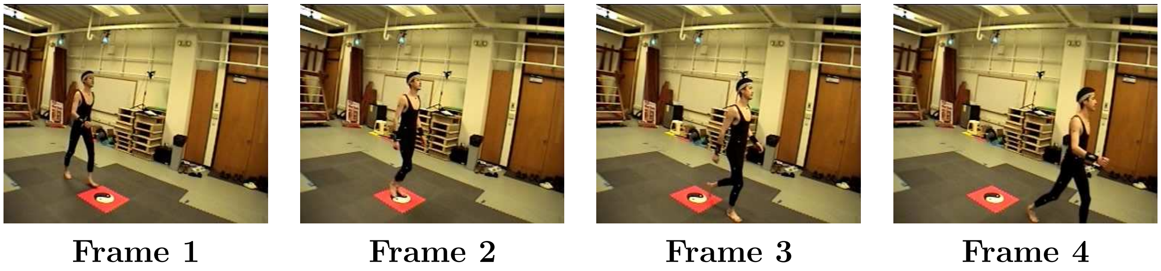
3. Results and Analysis

3.1. Qualitative Analysis of Background Subtraction

3.1.1. Results on Static Camera Sequences

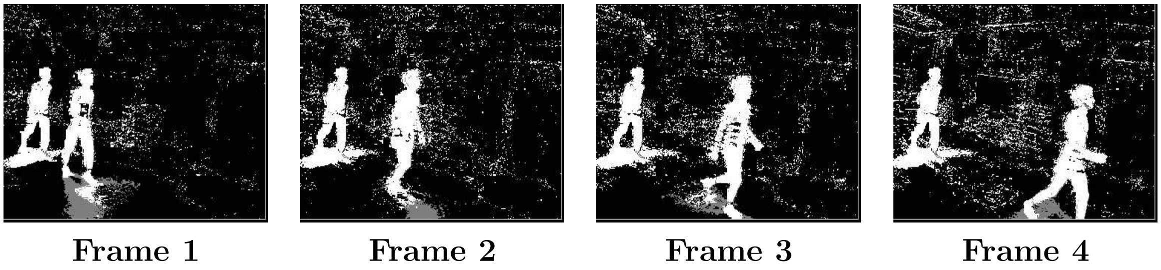
Experiments were conducted on a number of synthetic and real data sets from the Carnegie Mellon University [43] and CAVIAR data sets [44]. Frames from the two original video sequences used in the tests are shown in Figure 3 and Figure 7. A comparison of the outcome of the baseline algorithm (the pixel-based GMM) [7, 10] and the results of the proposed CBS technique is also presented on both sequences. Figures 4, 5, 6, 8, 9 and 10 illustrate the BS results of the pixel-level technique compared against the proposed cluster-level technique.

In Figure 11, recall-precision curves comparing the dipping and fixed thresholds for detection are presented. The graph is plotted for a 153 frames video sequence from the CMU data set [43]. It can clearly be observed that the dipping threshold mechanism (10) produces a much higher recall to precision ratio in comparison with varying percentages of the foreground. The dipping threshold mechanism is adaptive to different videos and thus is more flexible and efficient as against the fixed threshold methods.



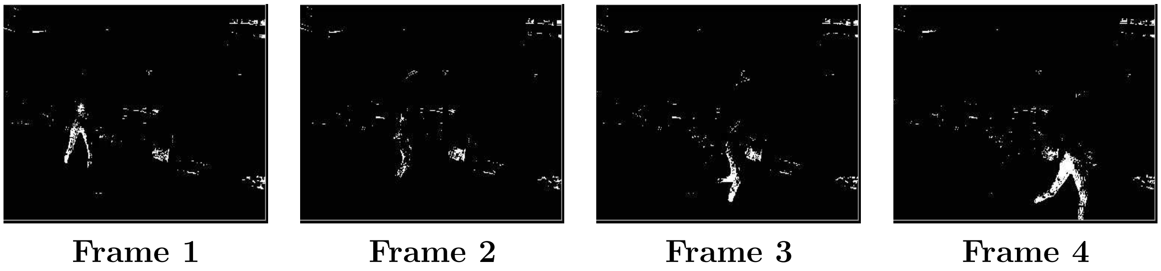
Frame 1 **Frame 2** **Frame 3** **Frame 4**

Figure 3: Original sequence from the CMU data set



Frame 1 **Frame 2** **Frame 3** **Frame 4**

Figure 4: Results from the pixel-based GMM [7] on the CMU data set



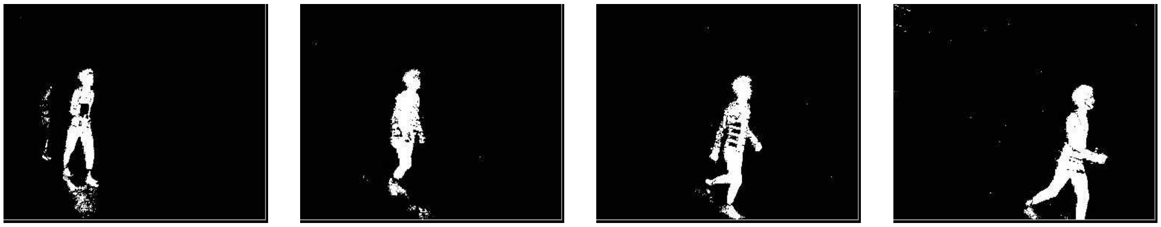
Frame 1 **Frame 2** **Frame 3** **Frame 4**

Figure 5: Results from nonparametric kernel background subtraction [9, 10] on the CMU data set

3.1.2. Results on Moving Camera Sequences

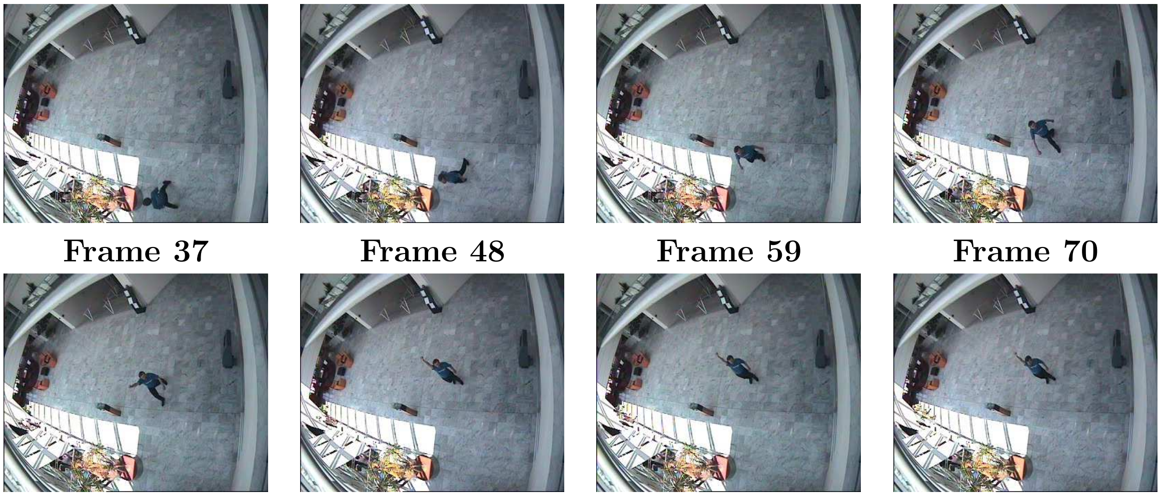
The proposed CBS technique was investigated on surveillance camera video sequences (Figure 12 captured using a hand held camera). Figure 15 presents results from the proposed CBS, Figure 13 gives results with the pixel-level GMM BS technique [7], and Figure 14 show results with the pixel-level GMM BS technique developed in [10].

Some of the major implications that can be derived from the results are: *i*) the pixel-level BS mechanism produces noisy/ cluttered BS as against the



Frame 1 **Frame 2** **Frame 3** **Frame 4**

Figure 6: Results from the proposed CBS with adaptive parameters on the CMU data set



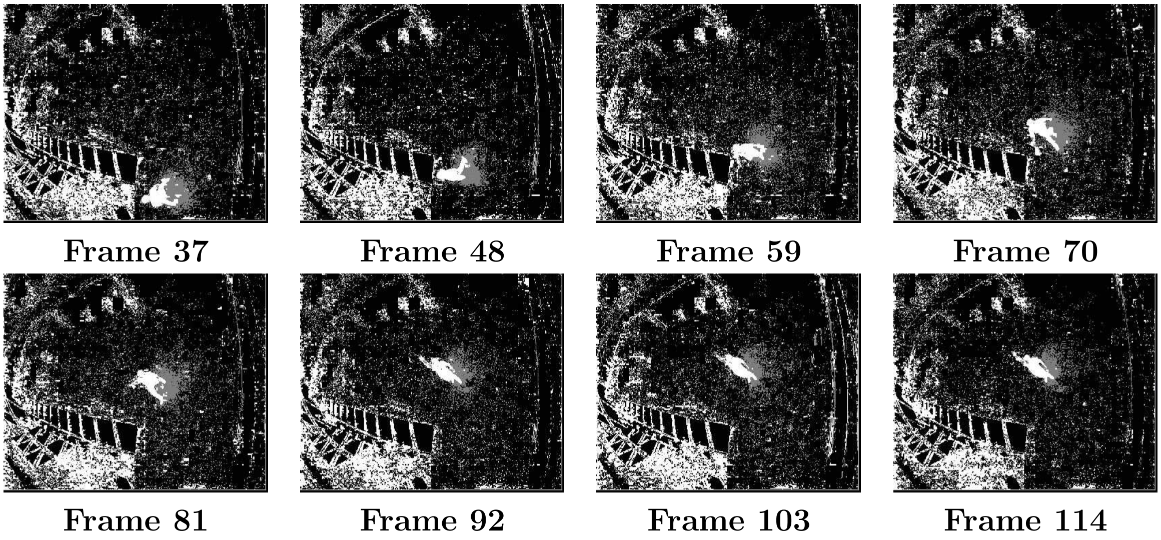
Frame 37 **Frame 48** **Frame 59** **Frame 70**

Frame 81 **Frame 92** **Frame 103** **Frame 114**

Figure 7: Original sequence from the CAVIAR data set

cluster-level techniques; *ii*) the output of the CBS scheme clearly distinguishes the foreground from the background regions thus permitting further processing of the output which is especially useful for tracking purposes; *iii*) false detection in the proposed technique leads to over segmented foreground regions and is attributed to the variation in dipping threshold; *iv*) the self-adaptive procedure used by CBS helps coping with camera motions whilst the pixel-level methods cannot handle moving cameras thereby not distinguishing regions of foreground and background.

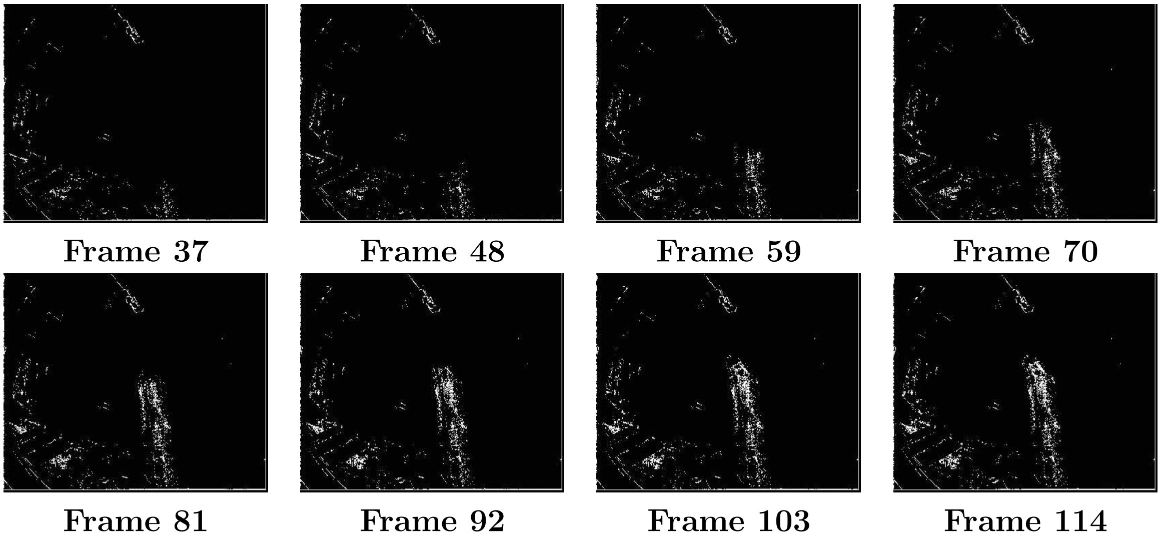
The video sequence from [18] consists of 109 frames captured at a frame rate of 15 frames per second. From the results shown in Figure 16, it is clear that the proposed CBS-$S\alpha S$ technique is capable of suppressing clutter simultaneously handle camera displacements with small movements in



Frame 37 Frame 48 Frame 59 Frame 70

Frame 81 Frame 92 Frame 103 Frame 114

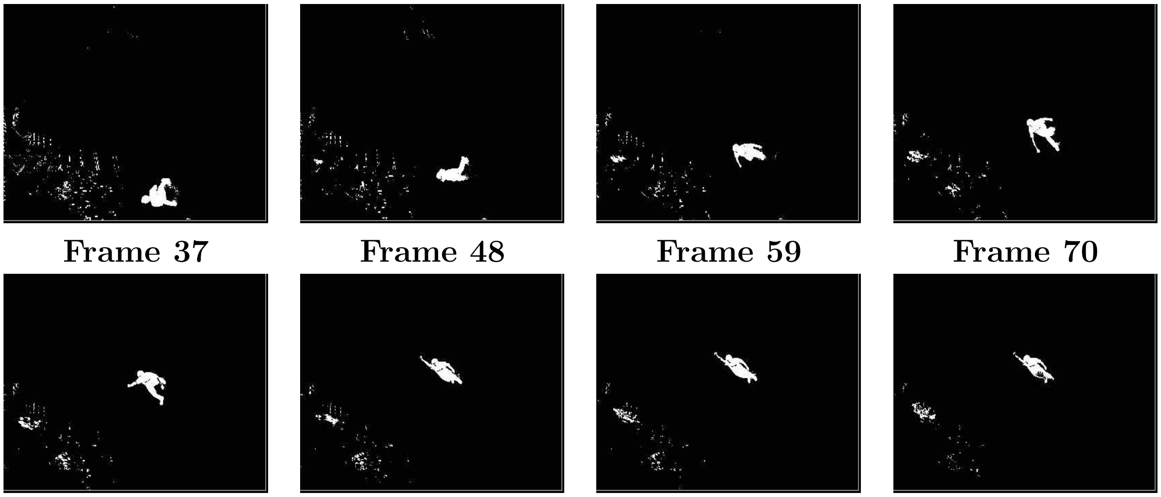
Figure 8: Results from the pixel-based GMM [7] on CAVIAR data set



Frame 37 Frame 48 Frame 59 Frame 70

Frame 81 Frame 92 Frame 103 Frame 114

Figure 9: Results from nonparametric kernel BS [10] on CAVIAR data set



Frame 37 Frame 48 Frame 59 Frame 70

Frame 81 Frame 92 Frame 103 Frame 114

Figure 10: Results from the proposed CBS with adaptive parameters on CAVIAR data set

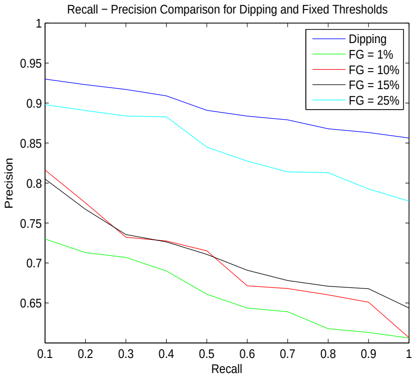
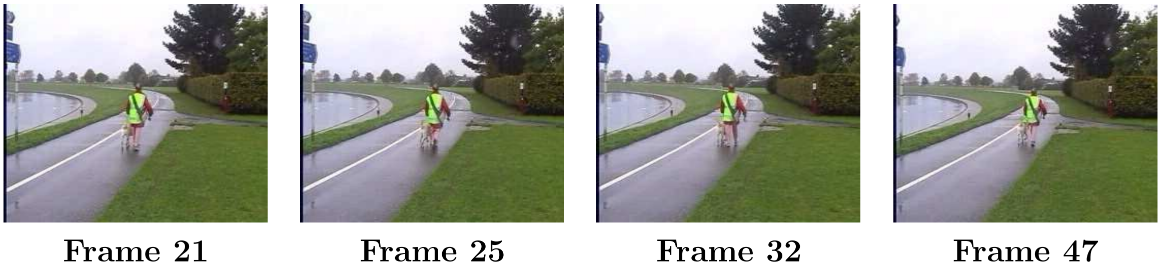


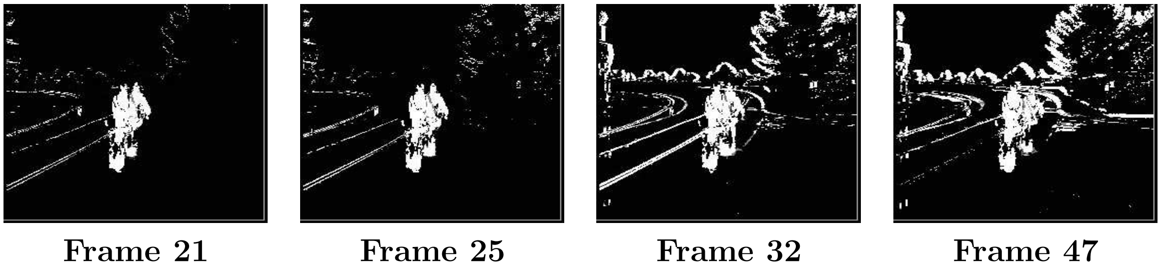
Figure 11: Recall-precision curves comparing fixed and dipping thresholds

the background. This sensitivity of the $S\alpha S$ densities relies on exploiting the heavier tails of the distribution to accommodate such variations in the background. In addition of this, the model also displays higher robustness to illumination changes. For more detailed review of the proposed CBS-$S\alpha S$ technique kindly refer to [45]. In Figure 16, the edge silhouette of the target is also isolated and presented. At level 1 of the proposed detection algorithm, these silhouettes are used for matching the target to the pictorial model.



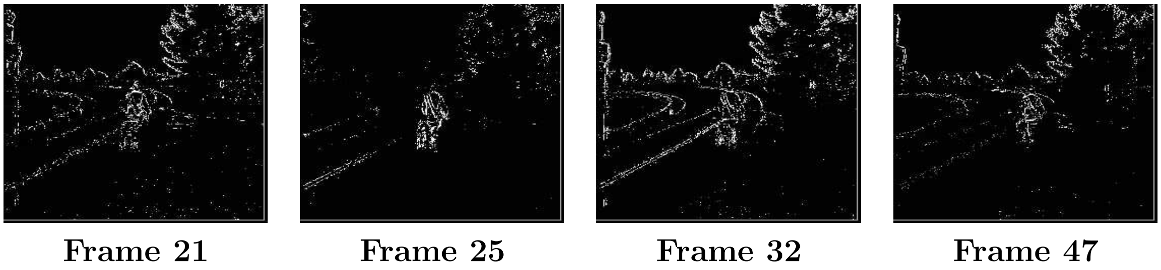
Frame 21 **Frame 25** **Frame 32** **Frame 47**

Figure 12: Original sequence from a moving camera



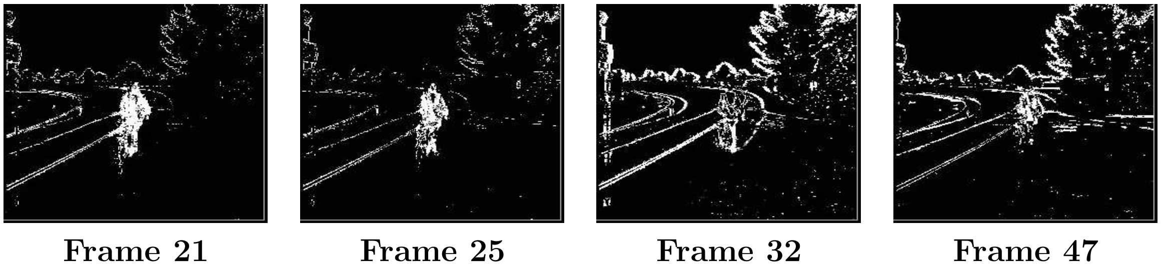
Frame 21 **Frame 25** **Frame 32** **Frame 47**

Figure 13: Results from the pixel-based GMM [7] on the moving camera sequence



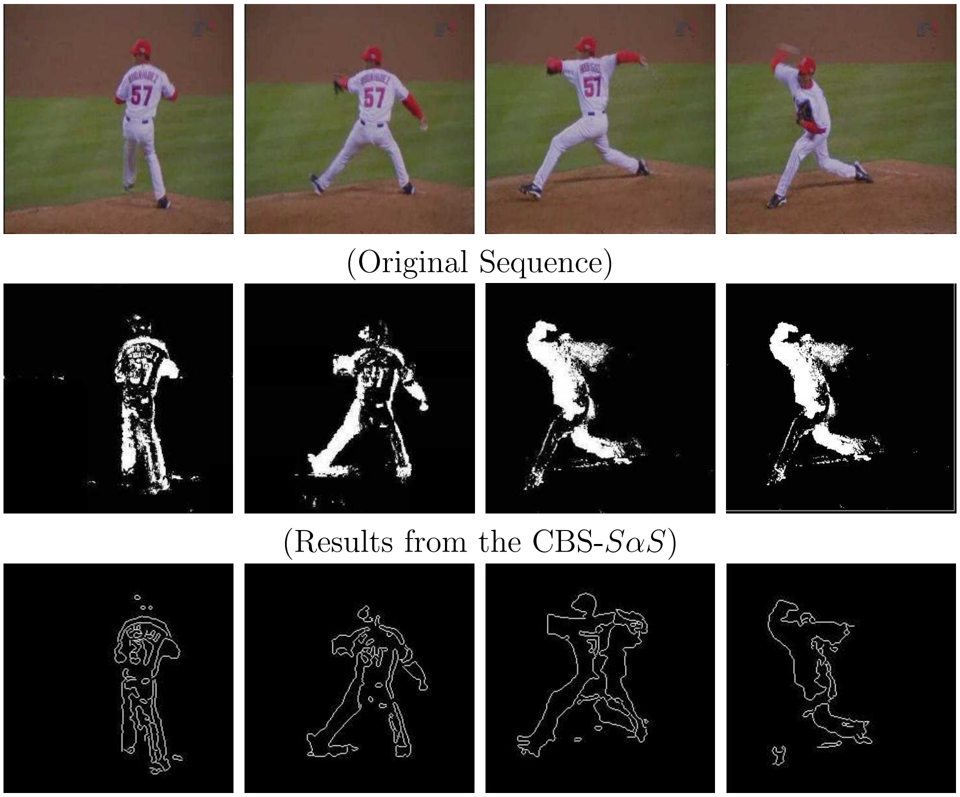
Frame 21 **Frame 25** **Frame 32** **Frame 47**

Figure 14: Results from nonparametric kernel BS [10] on the moving camera sequence



Frame 21 **Frame 25** **Frame 32** **Frame 47**

Figure 15: Results from the proposed cluster-based GMM with adaptive parameters on the moving camera sequence



(Original Sequence)

(Results from the CBS-$S\alpha S$)

(Extracted target edges)

Figure 16: Results from the proposed CBS-$S\alpha S$ model on the moving camera sequence along with extracted target silhouettes

3.2. Quantitative Analysis of Background Subtraction

The techniques are also evaluated using the quantitative measures recall and precision of the objects compared with the hand-labeled ground truth (TP: true positive (correct)) whilst accounting for the missed detections (FN: false negative) and the false detections (FP: false positives) [46]. Recall and precision measures quantify how well an algorithm matches the ground truth [47]

$$Recall = \frac{TP}{TP+FN}, \quad Precision = \frac{TP}{TP+FP}. \tag{17}$$

Figure 17 shows that the proposed algorithm has higher level of precision for the same values of recall. The precision values are directly related with the number of correctly classified foreground pixels [46], and are inversely

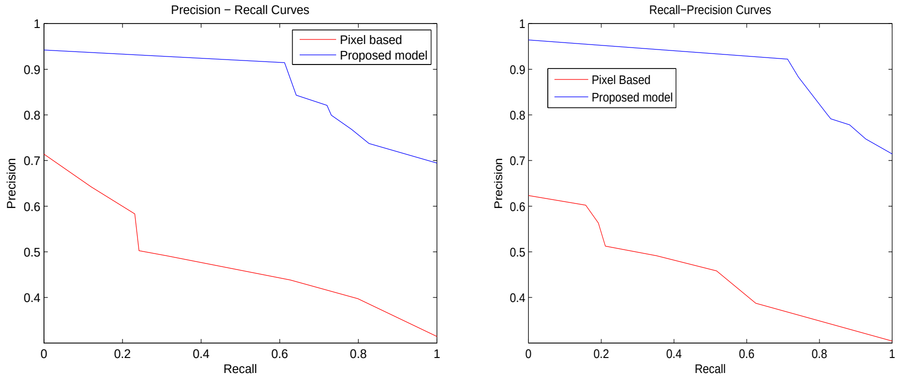


Figure 17: Recall-precision curves for the proposed model on CMU and CAVIAR data sets

proportional to the misclassified foreground pixels. It is evident that compared with the pixel-based GMM [7] the proposed CBS technique maximises the proportion of correctly classified pixels and minimises the misclassification.

Finally, a comparison between the time complexity of the pixel-based and the proposed cluster-based methods is made. Time is measured as the CPU runtime of different video sequences on an Intel i7 2.0-2.9GHz processor with Matlab 2010(b) version. The computational time is shown in the last two columns of Table 1. Column 2 indicates the number of frames analysed for time complexity of each of the video sequences. The time complexity of the proposed CBS mechanism is much lower in comparison with the pixel-level BS. The main reason for this reduced computational time is due to the fact that the CBS evaluates the parameters of a small number of clusters of the image as against all individual image pixels in the pixel-based methods.

In [45], a more detailed quantitative comparison of the CBS-$S\alpha S$ technique has been reported. In summary, it has been proven that the proposed CBS-$S\alpha S$ technique is capable to isolating targets by suppressing noise and clutter enabling the detection process to remain more robust and reliable.

3.3. Pictorial Structure Based Foreground Modelling

Both qualitative and quantitative evaluation are performed on the proposed pictorial structure based foreground modeling technique with GA search

Sequence	Frames	Pixel Based (s)	CBS (s)
CMU 1	202	141.3	112.6
CMU 2	257	178.5	136.4
CMU 3	153	139.1	89.8
CAVIAR 1	1055	512.2	312.1
CAVIAR 2	1135	678.9	439.6
Baseball	109	117.6	101.3

Table 1: Time complexity of pixel-based and CBS methods

method. The quality of matching is evaluated through visual inspection of results on the video sequences. The results of the qualitative evaluation is presented on some sample frames from both static camera sequence and on moving camera sequence. The matching process of the pictorial structure using the GA search for these experiments has been carried out on the background subtracted image for the level 1 of the cascade and on the gray scale of the original image for subsequent levels using the original objective function with combined posterior, edge error and overlap based pairwise constraint. For each image frame, the search technique is initialised with 500 object configurations around the neighbourhood of the center of mass of the segmented target and limit the maximum number of search iterations to 1000. The run time recorded for detecting these single human target parts on the video sequence was estimated to be around 68 seconds on an Intel Core i7 2.0-2.9GHz processor with Matlab 2010(b) version. Traditional approaches to human parts detection involve training and typically take several minutes. The time complexity of the proposed approach is low considering that it is an unoptimised code running in Matlab and can be adapted so that it is suitable for real-time applications.

The proposed automatic detection framework is compared with the baseline model proposed in [18] over the baseball sequence (moving camera with scale variations) shown in Figure 18. The model proposed in [18] relies on prior semantic knowledge and learns the appearance of objects to detect them. Here, no semantic knowledge is used in the CBS-GMM and foreground modeling technique.

Figure 19 illustrates the classification of body parts as obtained using

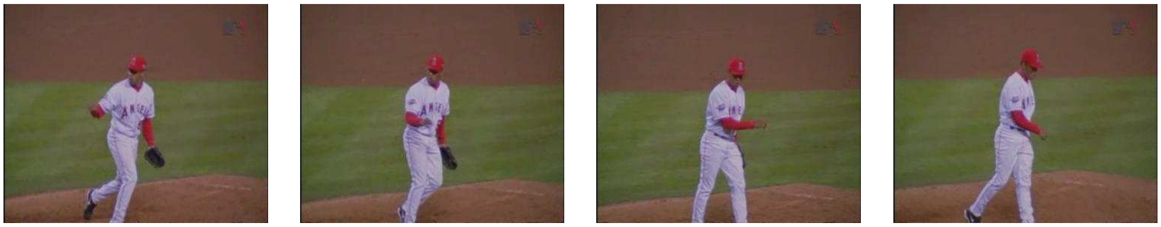


Figure 18: Original baseball sequence containing a total of 200 frames

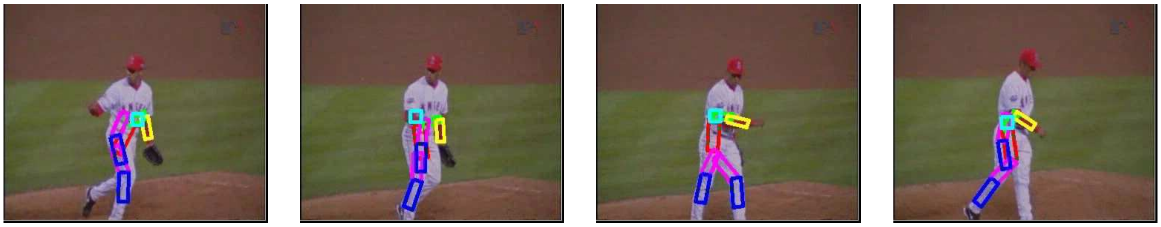


Figure 19: Results from the foreground modeling of human body parts using the algorithm proposed in [18].

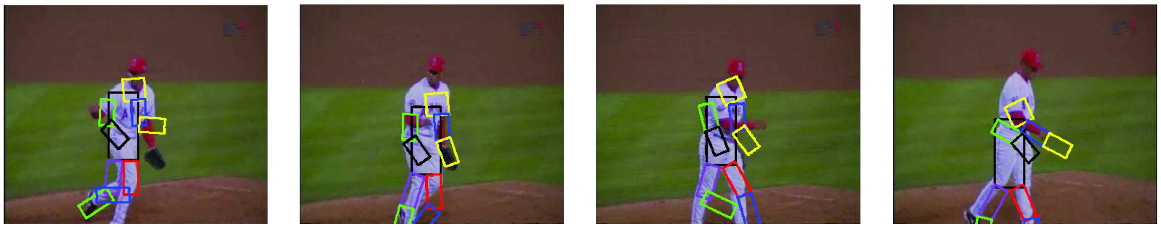


Figure 20: Results from the foreground modeling of human body parts using the proposed automatic detection framework without edge constraint and overlap penalty

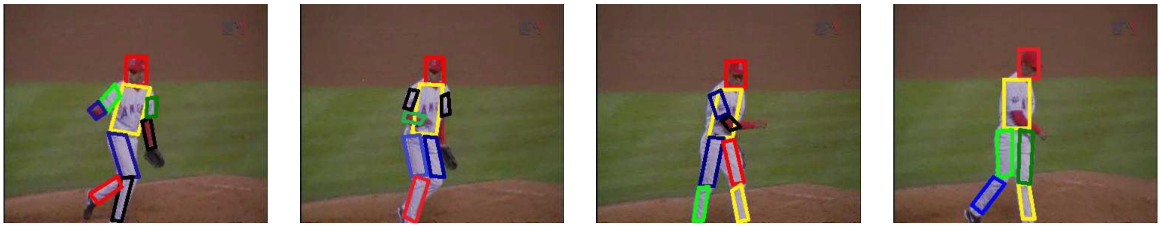


Figure 21: Results from the foreground modeling of human body parts using the proposed automatic detection framework with edge and overlap constraints

Video	Error (pixels)	Time (s)
CMU 1	8.2	61.23
CMU 2	6.4	63.64
CMU 3	7.1	69.52
CAVIAR 1	5.2	75.49
CAVIAR 2	4.7	71.97

Table 2: Tabular description of the chosen video clips and the error in localisation (pixels) with time complexity (msec)

the technique proposed in [18] and compare it with our technique. Foreground modeling is performed on the cluster background subtracted frames and thus showing the connection between the CBS and foreground learning. It is evident from Figure 20 that the proposed automatic detection framework based on CBS and evolutionary matching (without affine matching) produces comparable matching results even without the use of edge and overlap constraints. In Figure 21 the results of the proposed algorithm using the Hausdorff distance based edge error and the pairwise penalty using overlap constraints is presented. It can also be noticed that the number of body parts chosen to match to a particular frame is different from other frames. This is the result of the effect of the cascaded implementation of the search process. These results of detection are good for majority of the image frames. The only limitation of the system noticed so far is that the detection of body parts at the second level (for example, lower arms, leg parts below the knee, etc.) rely on the proper detection of their parent body parts. This dependence is inherited from the pictorial structure model.

The error between the locations of the matched body parts and their manually labelled ground-truth on 6 different video sequences is also measured and these results are tabulated in Table 2.

In addition, a performance curve is generated where the performance is measured as the percentage of correctly detected parts (PCP) against the L2 norm distance from the ground-truth.

The proposed algorithm performs comparably better than other techniques, some results from other techniques can be found in [48]. Although our results cannot be directly compared to the results of other methods in

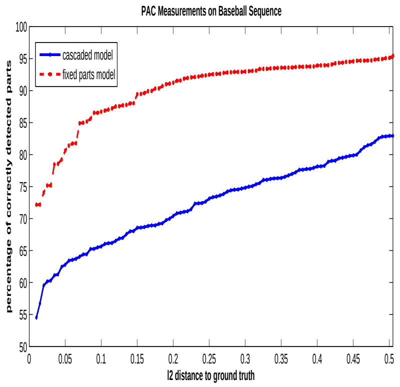


Figure 22: Percentage of correctly detected parts on the baseball sequence comparing the fixed and cascading models of the proposed framework

Video	Head	Torso	U.Arms	L.Arms	U.Legs	L.Legs
CMU 1	83.24	100	98.72	94.87	97.62	91.26
CMU 2	88.72	100	96.46	88.64	90.08	86.65
CMU 3	91.16	100	97.47	92.98	91.86	86.56
CAVIAR 1	75.62	98.14	87.62	81.46	88.92	80.45
CAVIAR 2	80.04	96.28	82.36	76.65	81.57	78.98
Baseball	89.90	100	98.64	90.96	96.37	95.23
Pet Walk	81.98	99.6	93.59	87.66	91.05	87.82

Table 3: Tabular description of the chosen video clips (short clips from original sequence) and the percent of correctly detected different body parts

[48] because our experiments are carried out on different sequences, atleast numerically the proposed strategy outperforms the other methods. It is clearly evident that the proposed cascaded implementation has a significant contribution to the success of the proposed framework in comparison to a fixed 10-parts body model matching. With the cascaded model it has been possible in some short sequences to nearly achieve 100% correctly detected body parts at around 0.35 L2 distance from the ground-truth. A detailed description of the proposed cascaded model performance on various sequences for all the localized body parts is listed in the Table 3.

Finally in this section, the overall performance of the proposed strategy is compared with the quantitative results of [18] over the baseball sequence (moving camera with scale variations) using the metric of percentage of frames that the respective body parts were correctly classified. The proposed system has localised the torso of the target in the baseline sequence 100% and the head in 89.90% of the frames in comparison to 98.4% torso detection without the head in the case of [18]. In addition, our model localizes the arms at 94.8% (averages between upper and lower arms) and legs at 95.8%. [18] presents 93.75% frames where arms have been correctly detected and 95.3% of frames where the legs have been correctly detected. Although that the percentage of frames comparing the models is nearly comparable for the detection of arms and legs, the failure rate in the proposed model is primarily due to low rates of the lower limbs (both arms and legs) which is due to inaccurate localisation of the corresponding upper limbs.

3.4. Effect of the Cascaded Implementation

One of the important considerations in understanding the capabilities of the cascaded model is to being by understanding the connection between the background subtraction process to the foreground matching stage. As it has been mentioned earlier, the output of the background subtracted procedure is the extracted silhouette of the target. At level 1 of the cascading model the matching of the pictorial structure happens on this background subtracted output thus allowing the prediction of the most useful subset of body parts as shown in Figure 23. At this stage no emphasis on the colour components of the body parts is taken. However, in the subsequently higher levels, both colour and edge constraints are simultaneously engaged to match the larger set of secondary body parts to the original image. In this way we are able to achieve much higher accuracy of body parts matching.

It is important to note that the cascaded model is not mandatory to the functioning of the proposed framework. It is quite possible to match with the proposed model a fixed number of body parts model to the target using the GA search scheme without having to begin with a subset of most salient body parts and consequently adding lesser salient body parts at higher levels. However, as the results in Figure 22 suggested, the use of the cascaded implementation significantly improves the performance of detection. In either case, the initialisation of the body parts pictorial model begins with matching the model to the background subtracted image and subsequently considering other features such as color and texture.

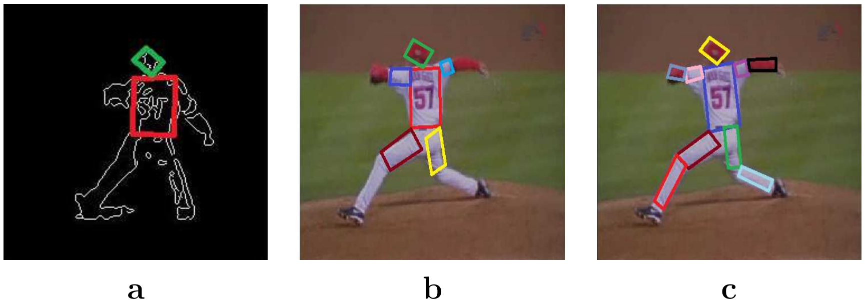


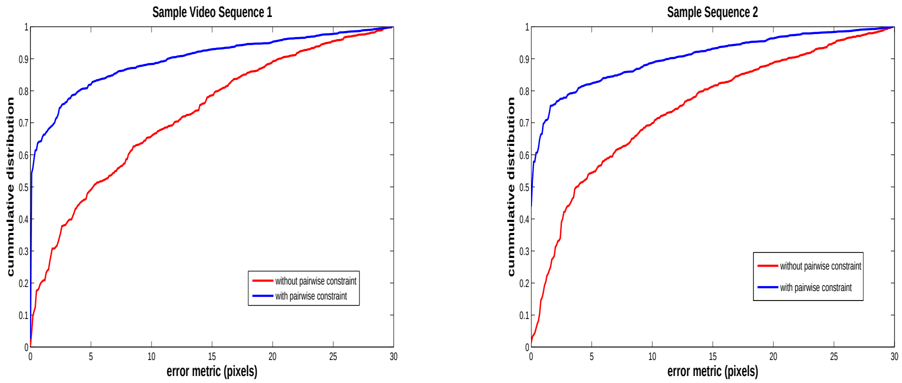
Figure 23: Cascaded implementation of the proposed framework and the link of background subtraction and foreground matching. a) illustrates the matching process at level 1 on the background subtracted silhouette, b) is the output at level 2 of the cascade when higher number of lesser salient body parts are added and finally, c) shows the matching of all body parts at level 3.

The usefulness of the proposed cascading implementation can be highlighted on several video sequences. A quantitative evaluation of this has partly been represented in our previous results as in Figure 25. In addition to that, some sample frames from the original sequence and their corresponding matched parts are also presented in Figure 21.

3.5. Effect of System Parameters on Performance

As mentioned earlier, detailed quantitative evaluation of the proposed method is conducted by measuring the effect of different system parameters on the performance. In our first experiments, the importance of initialisation of the solution space on the time complexity of the proposed search algorithm is tested. The solution space of the GA search mechanism can be initialised either a) randomly from within the dimensions of the image or b) in the neighbourhood of the center of mass of each segmented target from the background subtracted image or c) from the localisations of the previous frame (if used on a video sequence). The time complexity of the proposed matching framework when initialised with a random population was estimated to be 112.4s in comparison to 70.3s when the population was initialised around the neighbourhood of the target and 30.1s when the population was initialised from the estimates of the previous frame.

In our next experiment, the importance of the pairwise constraint on the accuracy of the target detection process is demonstrated. The error is measured as the mean difference between all the localised body parts and its manually



a **b**

Figure 24: Effect of overlap function based pairwise constraint

labeled counterpart against different number of frames from sample chosen video sequences containing 300 frames. The results presented in Figure 24a and Figure 24b are in the form of a cumulative error distribution. The results indicate that the use of the pairwise constraint (blue curve) shows a significant improvement (greater percentage of frames producing lesser error) as against the objective used without the pairwise constraint (red curve). The pairwise constraint not alone helps in improving the accuracy of the model but also ensures that only the optimal number of body parts are always chosen to represent the human target. In the case otherwise as shown by the red curve in Figure 25, the model always uses a fixed number of 10 body parts, some of which overlap with others or even sometimes not localised properly. To support this claim, the curve demonstrating the number of body parts obtained whilst the pairwise constraint was used on the baseball sequence containing the first 98 frames as indicated by the blue curve in Figure 25 is presented.

In the next group of experiments, the impact of the two error functions employed during matching of templates to targets: the MAD cost function (13) versus the probabilistic cost function (15) both without the use of the Hausdorff distance between the edges are compared. Again for these experiments, as in the previous case the error metric used is the difference of the localised target body parts and the manually labeled counter-part. The plots in Figure 26a shows the spatial error difference (in pixels) between the two

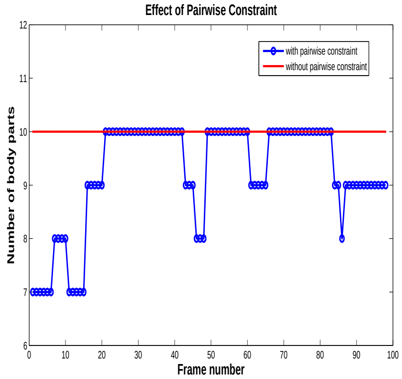
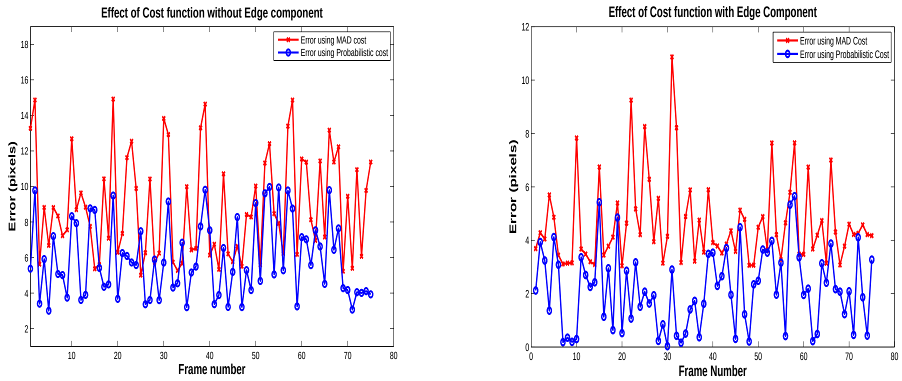


Figure 25: Effect of overlap function based pairwise constraint on the number of body parts chosen to represent the target

cost functions across a sample video sequence consisting of 75 frames. It is clear that the spatial error between the cost measures is significant. The same comparison between the two error functions is repeated but now with the use of the error between edges of the template and target. This is presented in Figure 26b. These results also conforms to our earlier estimates of the difference. However, in general the probabilistic cost function yields better results in terms of accuracy than the MAD function. It is also critical to note from these graphs, the high impact of the edge based criteria on the matching accuracy. It is obvious that the edge criteria not alone improves the qualitative accuracy of the matching process but also play an important role in advancing the quantitative performance of the model.

The evolutionary algorithm provides an efficient search mechanism for matching different parts of the human body and provides also *global* optimisation. This advantage of the evolutionary algorithm helps the model achieving efficient solutions in human body parts detection for movements with a high degree of freedom. To illustrate this claim further, an experiment comparing the evolutionary strategy to popular stochastic methods (in the form of a generic particle filter) is performed. In this experiment, 1 in every 10 frames of the baseball sequence are selected and the matching process is repeated using both the proposed GA based search technique and the generic particle filter (GPF). The GPF technique is initialised using the pictorial structure model of [16, 15]. The results presented in the Figure 27



a **b**

Figure 26: Difference error between the cost functions (13) and (15)

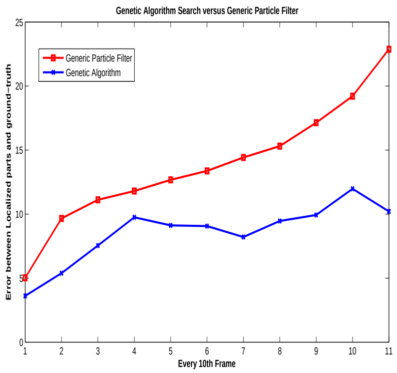


Figure 27: GA search compared to a Generic Particle Filter on the Pet Walk sequence

compares the error between the predicted locations of the body parts and their corresponding manually labelled ground truth data.

It is evident that the proposed GA search mechanism is capable of localizing the body parts of the target better than other stochastic methods. In addition, it has also been found that the error difference of matching the torso and the head regions of the target of both algorithms (GA and GPF) are nearly the same. This behaviour is fairly easy to explain as the motion of the head and torso regions are fairly predictable and less randomised than the regions of the arms and legs. The main reason for the superior performance of the GA search scheme is from the evolutionary nature of the GA search that allows the algorithm to unconditionally cope with randomized movements of the body parts at higher degrees of freedom. However, most stochastic models relies heavily on constant velocity movement to the different body parts. Although the GA search scheme has been found to be more accurate than other methods, it computational complexity of on an average 20% higher than the GPF methods.

Overall, the model has also been proven to be robust against the presence of clutter and occlusion without the use of additional heuristics. However, sometimes the pose estimation error is increased due to the nature of evolution of the population and the deformation model. Current research is focussed on refining the results with a different matching criterion and different evolutionary models.

4. Conclusions

In this paper, a technique for automatic people detection based on a *cluster* background subtraction using a GMM and an evolutionary algorithm with pictorial structure matching is proposed. First, each video frame is clustered in regions according to certain features such as colour. Next, the parameters of the GMM are calculated for each cluster centre. Operating at cluster level, the CBS technique is less dependent on variations of separate pixel intensities and noises compared with pixel level BS. As a result the CBS technique has shown to be more robust to intensity variations and is superior than pixel BS methods in terms of better accuracy, robustness to clutter and reduced computational complexity.

Additionally, a foreground modeling scheme for learning the appearance of human body parts is developed and linked with the CBS. This technique combines a rectangular pictorial structure representation with evolutionary learning for matching the different body parts over the CBS frames. Efficient and quick body part matching can be accomplished with the proposed mechanism. Robust automatic human detection is demonstrated and results over real video sequences from static and moving video cameras.

Acknowledgements

We are thankful the anonymous reviewers for their suggestions helping us to improve this work. We also acknowledge the support of UK MOD Data and Information Fusion Defence Technology Centre under the Tracking Cluster project DIFDTC/CSIPC1/02 for supporting this research.

References

[1] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings of IEEE Computer Vision and Pattern Recognition Conference*, 1999, pp. 246–252.

[2] C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 7747–757, 2000.

[3] A. McIvor, "Background subtraction techniques," in *Proceedings of Image and Vision Computing*, Auckland, New Zealand, 2000.

[4] 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 Inteligence*, vol. 25, no. 10, pp. 1337–1342, 2005.

[5] G. Foresti, C. Micheloni, L. Snidaro, P. Remagnino, and T. Ellis, "Active video-based surveillance system: the low-level image and video processing techniques needed for implementation," *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 25– 37, March 2005.

[6] B. Lo and S. Velastin, "Automatic congestion detection system for underground platforms," in *Proc. of the 2001 International Symp. on Intelligent Multimedia, Video and Speech Processing*, 2001, pp. 158 – 161.

[7] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773–780, 2006.

[8] I. Pavlidis, V. Morellas, P. Tsiamyrtzis, and S. Harp, "Urban surveillance systems: from the laboratory to the commercial world," *Proceedings of the IEEE*, vol. 89, no. 10, pp. 1478 –1497, 2001.

[9] A. Elgammal, D. Harwood, and L. Davis, “Non-parametric model for background subtraction,” in *Proc. 6th Europ. Conf. on Computer Vision*, June/July 2000.

[10] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, “Background and foreground modeling using non-parametric kernel density estimation for visual surveillance,” *Proceedings of the IEEE*, vol. 90, pp. 1151–1163, July 2002.

[11] B. Han, D. Comaniciu, Y. Zhu, and L. Davis, “Incremental density approximation and kernel-based Bayesian filtering for object tracking,” in *Proc. IEEE Conf. CVPR*, 2004.

[12] P. Felzenswalb, “Learning models for object recognition,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.

[13] D. R. Magee, “Tracking multiple vehicles using foreground, background and motion models,” *Image and Vision Computing. Statistical Methods in Video Processing*, vol. 22, no. 2, pp. 143–155.

[14] M. Mitchell, *An Introduction to Genetic Algorithms (Complex Adaptive Systems)* . MIT Press, 1998.

[15] P. F. Felzenszwalb and D. P. Huttenlocher, “Pictorial structures for object recognition,” *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[16] P. Felzenswalb and D. Huttenlocher, “Efficient matching of pictorial structures,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2000.

[17] D. Forsyth, O. Arikan, L. Ikemoto, and D. Ramanan, *Computational Studies of Human Motion: Part 1, Tracking and Motion Synthesis. Foundations and Trends in Computer Graphics and Vision.* Hanover, Massachusetts. Now Publishers Inc., 2006.

[18] D. Ramanan, D. A. Forsyth, and A. Zisserman, “Tracking people by learning their appearance,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 65–81, 2007.

[19] M. Bergtholdt, J. Kappes, S. Schmidt, and C. Schnrr, “A study of parts-based object class detection using complete graphs,” *International Journal of Computer Vision*, vol. 87, pp. 93–117, 2010.

[20] M. Andriluka, S. Roth, and B. Schiele, “Pictorial structures revisited: People detection and articulated pose estimation,” in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1014 – 1021.

[21] M. Andriluka, S. Roth, and B. Schiele, “Monocular 3D pose estimation and tracking by detection,” in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 623–630.

[22] M. Turk and A. Pentland, “Eigenfaces for recognition,” *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–96, 1991.

[23] H. Murase and S. Nayar, “Visual learning and recognition of 3-D objects from appearance,” *International Journal of Computer Vision*, vol. 14, no. 1, pp. 5–24, 1995.

[24] D. Huttenlocher, G. Klanderman, and W. Rucklidge, “Comparing images using the Hausdorff distance,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[25] A. Pentland, “Recognition by parts,” in *IEEE International Conf. on Computer Vision*, 1987, pp. 612–620.

[26] S. Dickinson, I. Biederman, A. Pentland, J. Eklundh, R. Bergevin, and R. Munck-Fairwood, “The use of geons for generic 3-D object recognition,” in *In International Joint Conference on Artificial Intelligence*, 1993, pp. 1693–1699.

[27] E. Rivlin, S. Dickinson, and A. Rosenfeld, “Recognition by functional parts,” *Computer Vision and Image Understanding*, vol. 62, no. 2, pp. 164–176, 1995.

[28] M. Burl and P. Perona, “Recognition of planar object classes,” in *In IEEE Conf. on Computer Vision and Pattern Recognition*, 1996, pp. 223 – 230.

[29] S. Ioffe and D. Forsyth, “Probabilistic methods for finding people,” *International J. of Computer Vision*, vol. 43, no. 1, pp. 45–68, 2001.

[30] Y. Gdalyahu and D. Weinshall, “Flexible syntactic matching of curves and its application to automatic hierarchical classification of silhouettes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1312–1328, 1999.

[31] T. Sebastian, P. Klein, and B. Kimia, “Recognition of shapes by editing shock graphs,” in *In IEEE International Conf. on Computer Vision*, 2001, pp. 755–762.

[32] C. Bregler and J. Malik, “Tracking people with twists and exponential maps,” in *In IEEE Conference on Computer Vision and Pattern Recognition*, 1998, pp. 8–15.

[33] S. Ju, M. Black, and Y. Yacoob, “Cardboard people: A parameterized model of articulated motion,” in *In International Conference on Automatic Face and Gesture Recognition*, 1996, pp. 38–44.

[34] M. W. Lee and I. Cohen, “Proposal maps driven mcmc for estimating human body pose in static images,” in *Proceedings of IEEE Computer Vision and Pattern Recognition Conference*, 2004.

[35] P. Peursum, S. Venkatesh, and G. West, “A study on smoothing for particle-filtered 3D human body tracking,” *International Journal of Computer Vision*, vol. 87, pp. 53–74, 2010.

[36] F. DiMaio, J. W. Shavlik, and G. N. Phillips, “Pictorial structures for molecular modeling: Interpreting density maps,” in *NIPS*, 2004.

[37] M. Kumar, P. H. S.Torr, and A. Zisserman, “Extending pictorial structures for object recognition,” in *Proceedings of the British Machine Vision Conference*, 2004, pp. 789–798.

[38] H. Hirschmller and D. Scharstein, “Evaluation of cost functions for stereo matching,” in *CVPR.* IEEE Computer Society.

[39] N. Dekker, L. S. Ploeger, and M. van Herk, “Evaluation of cost functions for gray value matching of 2D images in radiotherapy,” in *Proc. of the*

4th International Conf. on Medical Image Computing and Computer-Assisted Intervention. London, UK: Springer-Verlag, 2001, pp. 1354–1357.

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

[41] A. M. Payne, H. Bhaskar, and L. Mihaylova, "Multi-resolution learning vector quantisation based automatic colour clustering," in *FUSION Conference*, Cologne, Germany, 2008.

[42] S. Maskell, "A Bayesian approach to fusing uncertain, imprecise and conflicting information," *Information Fusion*, 2007, available on line.

[43] R. Gross and J. Shi, "The CMU motion of body (MoBo) database (CMU-RI-TR-01-18)," Robotics Inst., Carnegie Mellon Univ. The data are available at: $http://mocap.cs.cmu.edu/$, Tech. Rep., 2001.

[44] "Caviar test case scenarios (http://homepages.inf.ed.ac.uk/rbf/caviardata1/)," http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/, 2005.

[45] H. Bhaskar, L. Mihaylova, and A. Achim, "Video foreground detection based on symmetric alpha-stable mixture models," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 8, pp. 1133 –1138, Aug. 2010.

[46] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd Intl. Conf. on Machine Learning*, 2006.

[47] S.-C. Cheung and C. Kamath, "Robust techniques for background subtraction in urban traffic video," *Video Communications and Image Processing*, vol. 5308, no. 1, pp. 881–892, 2004.

[48] B. Sapp, A. Toshev, and B. Taskar, "Cascaded models for articulated pose estimation," in *Proc. of the European Conference on Computer Vision (ECCV)*, 2010.

[ ] Dr. Harish Bhaskar (MIEEE) is a Assistant Professor at the Department of Computer Engineering, Khalifa University, Sharjah Campus, United Arab Emirates. Before moving to the UAE for an academic career, Dr. Bhaskar worked as a Researcher at Manchester and Lancaster Universities, U.K. Dr. Bhaskar has been actively associated with several European Research institutes and the Ministry of Defense UK. His research interests are in the field of computer vision, image processing, visual cryptography, artificial intelligence and robotics.

[ ] Dr. Lyudmila Mihaylova (SMIEEE) is a Reader in Advanced Signal Processing at the School of Computing and Communications, Lancaster University, United Kingdom. Her interests are in the area of nonlinear filtering, sequential Monte Carlo Methods, statistical signal processing and sensor data fusion. Her work involves the development of novel Bayesian techniques, e.g. for high dimensional problems (including for vehicular traffic flow estimation and for image processing), localisation and positioning in sensor networks. On these areas she publishes book chapters and numerous journal and conference papers. Dr. Mihaylova is the Editor-in-Chief of the Open Transportation Journal and an Associate Editor of the IEEE Transactions on Aerospace Systems and Elsevier Signal Processing Journal. She is a member of the International Society of Information Fusion (ISIF). She has given a number of invited tutorials including for the COST-NEARCTIS workshop and is involved in the organisation of international conferences/ workshops. Her research is funded by grants from the EPSRC, EU, MOD and industry.

[ ]

Dr. Simon Maskell had an IEE scholarship to Cambridge University Engineering Department, from where he graduated with Distinction. His PhD was then funded by a Royal Commission of 1851 Industrial Fellowship and ran concurrently with a UK MoD fellowship. One paper that emerged from this research has been cited more than 3000 times. Simon now leads projects at QinetiQ on developing state-of-the-art Bayesian algorithms for tracking, data fusion, intelligence processing and video processing. Simon regularly reviews conference and journal papers and has given courses and tutorials at international conferences. Simon has also recently written the Wiley Encyclopaedia of Computer Science definition of Tracking and been general chair of International Conference on Information Fusion 2010, the first time this conference has been to the UK.