Efficient Method for Detection and Tracking of People

S. Anantha jothi
PG student, Department of CSE
GCE Tirunelveli

K. Thulasimani
Assistant Professor, Department of CSE
GCE Tirunelveli

ABSTRACT-This paper aims to develop an effective method to estimate the human in a frame of video stream. The paper has been threefold, first the preprocessing steps are performed using background subtraction results in foreground extraction and need training images to determine the relationship between foreground pixels and human oriented feature, second an Expectation Maximization based method has been used to cluster individuals in a low resolution scene. The cluster model is used to represent each person. Third the number of people is used as a priori for locating individuals based on feature points. The KLT tracker is used to track the people. Then the methods for estimating the number of people and for locating individuals are connected. Finally, a model is constructed to test the proposed system. Evaluation results on a number of images and videos and comparisons with previous methods are given.

Keywords
Expectation and Maximization, Cluster, Kanade Lucas Tracker.

1. INTRODUCTION

Visual surveilance has been a very active research area. In the recent years there has been a considerable growth in the availability of technology for real time surveilance. People counting is a challenging problem in surveilance field too. Detection and tracking of humans is important for many applications, such as surveilance, human computer interaction, and driving assistance systems. The principle sources of difficulty in performing real time monitoring are change in appearance of the objects with viewpoint, illumination and clothing also it is hard to maintain the identities of objects during tracking when humans are close to each other. The technique of clustering model is used to prior human shape. KLT features on human contours, KLT feature should be set such that the points from head-shoulder can be easily detected.

Many experimental studies have showed the evidence for segmentation and human tracking[1] and locating people by head top candidate. MCMC based method to compute the MAP estimate which has use of highly general reversible jump/diffusion and a mean shift tracking is handled[2]. The system detects the components of a person in an frame (i.e) the head, the left and right arms and the legs, instead of the full body[4]. Shadow pixels have same hue as the background but have lower intensity so the face can be easily detected[5]. The original shape contexts used binary edge presence voting into log polar spaced bins, irrespective of edge orientation[6]. Bayesian combination of part detector solving random fields using dynamic graph cuts[7]. To achieve multiscale human head detection preprocessing step is used with histogram equalization[8]. Shape priors and segmentation though produce good results, it had some short comings. It focused on obtaining good segmentations and did not provide the pose of the object explicitly. So there is a need to focus the concentration on clustering technique.

In this paper, presenting a new clustering model, called Expectation and Maximization(EM), to overcome the above limitations of shape prior segmentation an EM algorithm is used. Here expectation step uses complete data log likelihood function. By the continuation of Maximization step the algorithm is guaranteed to coverage to local maxima of the likelihood function with each iteration increasing the log likelihood.

The rest of the paper is organized as follows: Section 2 gives a brief review of Gaussian Mixture Model. Section 3, gives a brief review of EM Cluster Model. Section 4 provides a brief overview about Algorithm and Section 5 introduce KLT tracking and feature selecting. Section 6 gives the experimental result provides feature extraction result and clustering based blob detection system. Finally, Section 7 presents the conclusions of this paper.

2. RELATED WORKS

A mixture model with high likelihood tends to have the following traits:

1. Component distributions have high “peaks” (data in one cluster are tight).
2. The mixture model “covers” the data well (dominant patterns in the data are captured by component distributions).

Main characteristics of model-based clustering:

1. Well-studied statistical inference techniques available;
2. Flexibility in choosing the component distribution;
3. Obtain a density estimation for each cluster;
4. A “soft” classification is available.

The most widely used clustering method of this kind is the one based on learning a mixture of Gaussians [9]. Actually considering clusters as Gaussian distributions centred on their barycentres, as that can see in this picture, where the grey circle represents the first variance of the distribution:

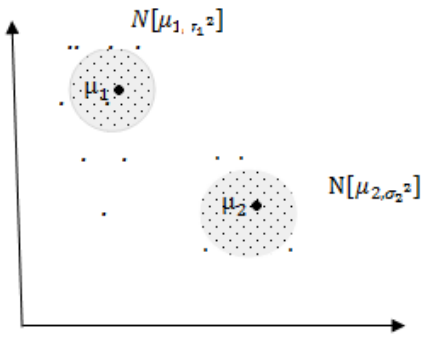

FIGURE 1: Mixture of Gaussians

3. EXPECTATION AND MAXIM IZATION (EM)

EM tries to find a set of basis vectors that can well approximate the dataset. Dataset and mask set can be used to learn the human-based representation. When applied for real world applications, the problem with Gaussian mixture model is that it fails to provide accuracy in human detection. To solve this problem, we use EM algorithm[3] over the Gaussian mixture model. The clustering information by constructing an ellipse of eh and ew it rounds up similar features according to data set.

To simplify the algorithm, only grayscale images are applied in our method. After getting the background image, a foreground image is obtained by subtracting the current image from the background image [10]. The foreground image is binarized based on a threshold to obtain the foreground pixel. The threshold in our evaluation is 25. When the intensity difference of a pixel between the current image and background image is larger than 25, the pixel is viewed as foreground pixel and then training set are needed to compare foreground blob.

3.1 Goal:

The goal of E-Step is to obtain the assignment probability, which associate the feature points with each cluster

$$\hat{p}(j|i) = \hat{p}_j h(i|j) / \sum \hat{p}_j h(i|j) \qquad (3.1.1)$$

The goal of M-step is to maximize the likelihood with respect to the cluster model parameters.

$$\hat{p}_j = \frac{1}{\hat{n}_j} \sum_{i=j}^{n} \hat{p}\left(\frac{j}{i}\right) s_i \qquad (3.1.2)$$

4. OVERVIEW OF ALGORITHMIC IMPLEMENTATION

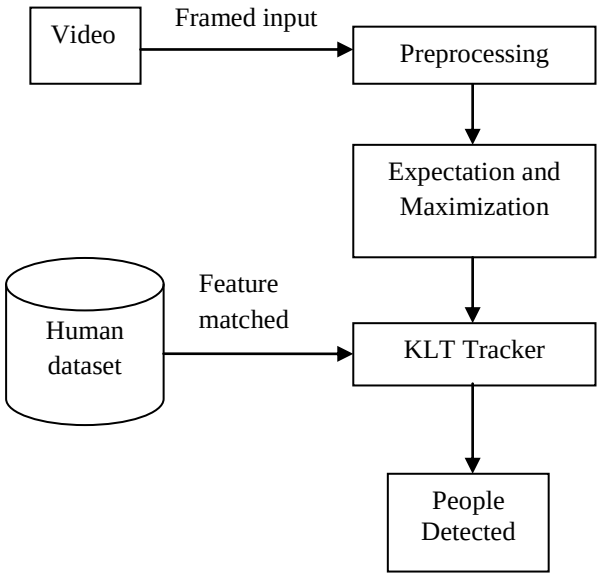

FIGURE 2: Schematic Representation of human detection

4.1 Preprocessing

This is the initial step where frames from the video input database are taken and the features of individual frame are extracted. Firstly frame denoising is performed to remove the noise in each frame using gaussian filter. Feature extraction processes use this enhanced frame for extracting the features namely shape context and KLT edge features. In this work, we consider a clustering technique and KLT feature for human detection. This block gives a feature matrix, X of size mxn where 'm' is the total number of features of an test image and 'n' is the total number of images in the trained dataset. The feature matrix, X is one of the main inputs to the EM algorithm and KLT tracker for processing the input frame.

4.2 Processing of foreground pixel

In this module, generate the foreground extraction a robust adaptive background estimation method based on the Gaussian mixture model is employed in this project[9]. To simply the algorithm only grayscale images are applied. After getting the background image a foreground image is obtained by subtracting the current image from the background image. To determine the relationship between foreground pixel and number of people some manually annotated training images from a similar scene are needed.

4.3 Expectation and maximization

The EM clustering results may contain some redundant ellipses. The feature points falling in these redundant ellipses are also included in other ellipses. It is reasonable to remove the ellipses without sufficient evidence from the feature points. In our test, the candidate ellipses are checked one by one and the redundant ellipses removed[1]. A very simple occlusion analysis is performed in this step. Humans not occluded by others should have more than three feature

points, while two feature points are acceptable for those who are occluded.

The EM algorithm has two main parts,

1. The first case occurs when the data has missing values due to limitations or problems with the observation process.
2. The second case occurs when the likelihood function can be obtained and simplified by assuming that there are additional but missing parameters.

The EM algorithm will find the proportion of belongs to each normal distribution along with other unknown parameters for the means and variance.

The density for the mixture of two Gaussian populations is,

$$f_W = p * \frac{1}{\sigma_1} * \varphi\left(\frac{w-\mu_1}{\sigma_1}\right) + (1-p) * \frac{1}{\sigma_2} * \varphi\left(\frac{w-\mu_2}{\sigma_2}\right) \quad (4.3.1)$$

P: Proportion of observation from normal distribution.
μ_1: Mean from normal distribution with shortest waiting time.

σ_1: Variance from normal distribution with shortest waiting time.

μ_2: Mean from normal distribution with longest waiting time.

σ_2: Variance from normal distribution with longest waiting time.

TABLE 1: The initial value of θ, decide to have the following values

$P^{(0)}$	$\mu_1^{(0)}$	$\mu_2^{(0)}$	$\sigma_1^{(0)}$	$\sigma_1^{(0)}$
0.4	40	90	4	4

4.4 Kanade Lucas Tracking

KLT is a popular corner detector and show good performance for tracking. In KLT points due to human shape, head shoulder part offer crucial KLT feature on human contour[10]. KLT feature point has 2 parameters, the number of features to be detected, minimum distance between 2 feature centres.

$$I(u) = I(x,y)$$
$$J(u) = J(x,y)$$

Let's consider two gray-scaled sequential images, I and J. For an interest point or feature u, where u = [x y]T
The image has width n_x and height n_y.
The goal of this method is to find the location,

$$v = u + d$$

On the second image such as I(u) and I(v) are similar. The vector d is the image velocity at the point u.
The next step is to choose a neighbourhood where we can analyze the similarity between u and v.

The residual function is applied to a window size of ,

$$(2w_x + 1) \times (2w_y + 1)$$

This window is sometimes referred as integration window, and it has typical values between 2 and 7. To choose a good value we should consider two important aspects of any algorithm: accuracy and robustness. If we choose a small a value, we increase the accuracy, since it will consider only the nearest neighbours for the calculations, but it decreases the robustness, since the big motions will be ignored. The opposite problem occurs if we choose a big value for the integration window. In an idealistic situation, we should have $d_x = w_x$ and $d_y = w_y$, to cover all the possible motions.

5. KANADE LUCAS TRACKING (KLT)

To track features there are essentially two important steps. The first one is to decide which features to track, and the second one is the tracking in itself. One of the most suited methods is the Lucas-Kanade feature tracking algorithm.

The algorithm works as it follows:

1. The optical flow is computed at the deepest pyramid level Lm, using the classical Lucas Kanade optical flow algorithm.
2. The result is propagated to the upper level Lm-1 in a form of an initial guess for the pixel displacement.
3. The optical flow is computed for the pyramid level Lm-1.
4. The same procedure until reach the highest pyramidal level.

5.1 Kanade Lucas Feature Selecting:

The Eigen value G has to be invertible, or in another words, the minimum eigenvalue of G must be larger than a threshold. This characteristic is fundamental to decide which pixels are good to track. To select the features, the following steps are applied,

1. For every pixel in the Frame I, compute G and its minimum eigenvalue Cm.
2. Find the maximum value of Cm and call it Cmax.
3. Keep the pixels that have a Cm larger than a percentage of Cmax (10% or 5%).
4. For all these pixels, keep those ones that have a Cm bigger than the entire Cm in their 3x3 neighbourhood.
5. Finally, create a subset of these pixels so that the minimum distance between these pixels is larger than a given threshold.

6. EXPERIMENTAL RESULTS

In this section, we demonstrate some experimental results to show the performance of the proposed system are based on

video input of .avi format with width 608 and height 256, which can produce 24fps totally 264 frames for a video.

TABLE 2: Values of the feature extraction for an input video comparing with human mask in the dataset

Test feature	Feature Extraction	Feature Matching
1580	31.10	37.90
1136	17.34	25.17
758	9.56	17.40
527	5.39	13.23
326	3.15	9.96

The Maximum vote score is calculated from the matching feature of features in query frame and features in the trained dataset. The probability for maximum vote score is calculated for human detection. The estimation of human is in accordance with the human data shape set and foreground shape context set which is already trained. The processes of resizing the frame have to be done for better estimation and feature detection.

7. CONCLUSION AND FUTURE WORK:

The proposed system for People counting and human detection maintains good performance and tolerable for different situation including people in different direction and movement in different scenario. The system consists of a Tracking algorithm followed by an approximate search method. The proposed Expectation and Maximization algorithm (EM steps) extracts robust, discriminant, and compact estimation from videos in a fast and reliable fashion. It consumes very low search time.

This paper provides a simple and efficient method for foreground clustering, and human tracking, which performs clustering based on the computation of the Expectation and maximization, KLT tracker between the features of number of query frame and the data in the database. This is a one to one mapping where the accuracy of the retrieval is fixed always and cannot be improved. The system performance can be enhanced by incorporating Neural networks. This training process can be repeated until an improved accuracy as desired is reached thus providing better detection and tracking performance.

8. REFERENCES

[1] Ya Li Hou, People Counting and Human Detection in a Challenging Situation. IEEE Transactions on System, Man, and Cybernetics Vol. 41, No.1, 24:33, January 2011.

[2] Tao Zhao, Tracking Multiple Humans in Complex Situations. IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 26, No. 9, 1208:1220, September 2004.

[3] Tao Zhao, Segmentation and Tracking of Multiple Humans in Crowded Environments. IEEE Transaction on Pattern Analysis And Machine Intelligence, Vol. 30, No. 7, 1198:1211, July 2008.

[4] Anuj Mohan, Constantine Papageorgiou and Tomaso Poggio. IEEE Transaction on Pattern Analysis And Machine Intelligence, Vol. 23, No. 4, 349:361, April 2001.

[5] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in Proc. Eur. Conf. Comput. Vis., pp. 69–82, 2004.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 886–893, 2005.

[7] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," Int. J. Comput. Vis., vol. 75, no. 2, pp. 247–266, Nov. 2007.

[8] S. F. Lin, J. Y. Chen, and H. X. Chao, "Estimation of number of people in crowded scenes using perspective transformation," IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 31, no. 6, pp. 645–654, Nov. 2001.

[9] G. J. Brostow and R. Cipolla, "Unsupervised Bayesian detection of independent motion in crowds," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 594–601, 2006.

[10] V. Rabaud and S. Belongie, "Counting crowded moving objects," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 705–711, 2006.