Computing the Pedestrians based on static & dynamic motions using Correlation based optical flow algorithm

[1]V.G.Janani, [2]R.NagaPriyanka, [3]G.Selvi, [4]S.MadhuShalini,
[1] Assistant Professor, ECE, [2,3,4] UG student, ECE,
Velammal college of Engineering & Technology, Madurai, India,

Abstract

People counting is an important problem in video surveillance as well as human detection in a crowded video is also difficult. This project aims to develop an effective method for estimating the people count and locate each individual even in a low resolution image with complicated scenes. The restriction on the work done so far includes: people must be moving, the background must be simple, and image resolution must be high. The problem of occlusion and foreground dependence is greater when it comes to detect & count individuals in a complicated scene. This project provides a method to overcome the above mentioned problems. We have worked in our datasets namely visor and peds to get reliable results. Initially we have performed preprocessing techniques such as background subtraction Barnard and thompson optical flow algorithm is used to segment and differentiate the people who are in motion as well as static in crowded and complex scenario. Moreover we have extended our work to count the number of people. We show promising reliable results and evaluations on some challenging data.

Keywords— barnard and thompson optical flow, segmentation, labelling

I. INTRODUCTION

People counting and human detection are two important problems in visual surveillance. It is useful for shopping mall managers to have information on the number of people in a mall each day. In addition, to ensure the safety of people and facilities, video surveillance has become more and more important. There has been considerable work in detecting and counting human in recent years. This has been a topic of considerable research in the recent past and robust methods for tracking isolated or small number of humans having only transient occlusion exist. The goal of our work is to develop a general framework to detect and count humans in crowd scenario without occlusion.

The ability to accurately detect and segment humans in video sequences represents an essential component in a wide range of application domains such as dynamic scene Analysis, human-computer interface design, driver assistance systems, and the development of intelligent environments. Nevertheless, the problem of human detection has numerous challenges associated with it. Effective solutions must be able to account not only for the nearly 250 degrees of freedom of which the human body is capable[1], but also the variability introduced by various factors such as different clothing styles, and the presence of occluding accessories such as backpacks and briefcases. Furthermore, a significant percentage of scenes, such as urban environments, contain substantial amounts of clutter and occlusion. Despite these challenges, detecting humans within video sequences has constituted an active area of research for a number of years, resulting in the proposal of numerous approaches. Nevertheless, only a small subset of the existing methods has been demonstrated to be effective in the presence of considerable overlaps and partial occlusion in video sequence.

Fig1: Crowded scenario

Various methods which estimate the number of people in input images have been previously proposed. They can be divided into three approaches:

Trajectory clustering approach

In this approach people are counted by tracking and identifying visual features over time. The feature trajectories which exhibit coherent motion are clustered and a number of clusters gives an estimation of the number of pedestrians.

Feature-based regression approach

This approach estimates a number of pedestrians by a regression on features extracted from an input image, e.g. neural networks.

Individual pedestrian detection

In this scheme, the pro- posed algorithm estimates the number of pedestrians who were detected in input images. Crowd counting has been an important component in video surveillance systems. For example, different levels of attention could be based on crowd of different density [2]. There are different approaches for counting the number of people from a crowded scene. One of the approaches is to directly detect humans from an image or videos. There has been a vast amount of work done in the area of human detection [3], and it usually performs well in scenes that are not very crowded. The other approach is indirect method which is based on a regression function to map the features from the region to the person count [4, 5, 6,7and 8]. The two approaches have certain advantages depending on the scene considered. Detection based approach provides the exact /location of the person, but might require sufficient resolution for reliable detection and would be challenging to detect multiple people in crowded/occluded situation. Whereas regression based approaches could be trained to work on low image resolution and crowded situations, but cannot provide the exact location of person, and tends to be scene specific.

Our work concentrates on to segment as well as count in ambiguity scenario. In this paper we investigate detection based approach for crowd counting task on PEDS and Visor databases .An example from PETS database and visor databases are shown in Figure 2. Given a crowded situation, Itis Detection based approach also has the advantage that it can be applied on a single image rather than a video.

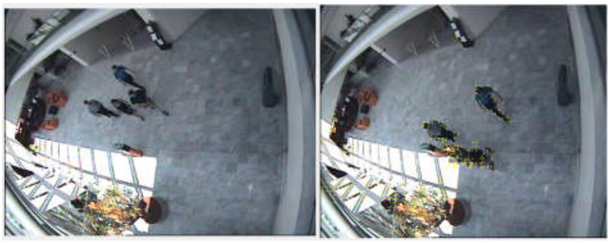


Fig2(a).Input **fig2(b).Motion vector**

Fig (2) Peds dataset

II. RELATED WORK

People counting and human detection have become hot topics in recent years. Currently, a great deal of research has been carried out in these areas. Moreover the most prevalent class of methods present in the literature is the detector-style method, in which detectors are trained to search for humans within a video sequence over a range of scales. Approaches such as AdaBoost have been used with some degree of success to learn body part detectors such as the face [10], hands, arms, legs, and torso [5] [8].While this class of approaches is attractive, detection of parts is itself a challenging task. This is particularly difficult in the class of scenes in which we are interested, which consist of crowded scenes containing significant occlusion amongst many parts. Rittscher et al [11] tried to reduce the requirements for an accurate foreground contour by sampling only the informative feature points from the contour. The sampled points were labelled as top, bottom, left, and right, which was based on their local contour information. A variant of expectation (E)–maximization (M) algorithm has been used to find the best grouping of the points within rectangles. In the E step, feature points are assigned to the rectangle candidates with a probability based on the distance to the corresponding top, bottom, left, or right borders of the rectangles. In the M step, rectangle sizes and locations are updated based on their associations with the feature points. Points with a low assignment probability have a low influence on the rectangles. A considerable amount of work has also focused on shape-based detection. Zhao et al [12] use a neural network that is trained on human silhouettes to verify whether the extracted silhouettes correspond to a human subject. However, a potential disadvantage of the approach resides in the fact that they rely on depth data to extract the silhouettes. Others, such as Davis et al [14] have also attempted to make use of shape-based cues by comparing edges to a series of learned models. Wu et al [13] have proposed learning human shape models and representing them via a Boltzmann distribution in a Markov Field. Although a number of these methods have proved to be successful in detecting humans in still images, most of them assume isolated human subjects with a minimal presence of clutter and occlusion. The work of Dong et al. [15] is also based on the insight that the foreground contour is a strong indication about the number and positions of human beings in a crowd. Since the number of people inside a foreground blob will never suddenly change, ambiguities inside the crowd region are mitigated by considering a set of consecutive frames. However, this strategy does not solve the problem inherently; for example, when the crowd does not show significant movement, ambiguities inside a dense crowd will cause a significant drop in the performance of the method. Another fatal problem of the method is that a great number of labelled training samples are required. To include almost all the possible occlusion situations, a small number of people will need a huge number of training samples. The requirements will seriously limit the application of the method in a large crowd. We are interested in methods that enable effective use of shape-based cues in the presence of heavy occlusion of the kind present in most urban environments.

There are different approaches that have been proposed for estimating the people count from an image and videos. Most of the approaches use background subtraction to segment humans from the background. In [8, 9], the authors mapped some feature statistics extracted from the blob to the count of people. In [5], a feature response belonging to the object is given a weight such that the sum of feature weights extracted from a full object sums to one. Thus even if partial object is visible their method will give partial count to the object. In [3, 4] background subtraction is used to obtain foreground pedestrian pixels which are mapped to the number of people in the scene. The issue is as the crowd increases the estimation will deviate from true density due to occlusion and it is not easy to figure out just by foreground pixels if it belongs to the same or different person. In [10], maximally stable extremely regions (MSER) are used as a region extraction prototype for crowd feature description. In [11], corner points based on Kanade-Lucas- Tomasi are extracted which are clustered to detect human like structures after removing any background features by using foreground mask In [12], top of head region is detected after background subtraction to segment roughly the human by an ellipsoid. To reduce the dependence on an accurate foreground con- tour, which may be easily corrupted by noise, Rittscher et al. [13] extracted some additional feature points from the con- tour. A variant of Expectation Maximization (EM) is used to group these features into some human-sized rectangles. The authors have shown that their method work well even under low resolutions. However, if the background model depends on pixel value average over time, as used in their paper, obtaining good foreground might be difficult when the people in the scene are stationary. Head detection using skeleton graph for people counting is proposed in [14]. The skeleton graph is extracted from the foreground mask obtained using background subtraction. In [15], a SVM classifier is trained to detect contours of head region, and then a perspective transform technique is used to estimate the crowd size more accurately. Our work is more closely related to method in [15] and follows a detection based approach to count people from an image. In summary, the methods described previously have three serious limitations: First, they usually rely on an informative foreground contour, which cannot be easily obtained in most situations. Second, with little information inside the foreground area, the methods would find it difficult to handle the high ambiguity at the center of the dense crowds, which limits their applications in dense scenarios. Third, the methods assume that the region shown in the foreground is from human beings. No specific measure has been taken to deal with nonhuman objects in the foreground. In addition, except for the background subtraction, the methods in [12]– [15] do not consider the motion features explicitly. Although the temporal information has been used in [15] to handle the ambiguities inside the crowd region, it has not solved the problem inherently. If no significant information can be extracted about the number of people on the foreground contour for the entire sequence, people in the dense region will never be counted.

Our method works on dense scenario in which the people segmented and counted without occlusions and shadow illuminations. We have used barnard and thompson optical flow algorithm utilized for segmentation and labelling technique for counting.

Fig(3).Crowded scenario

III. DATASETS

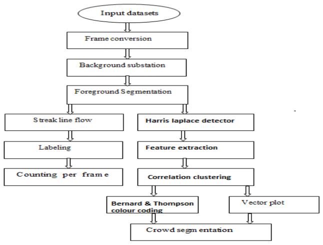
Our works performed on datasets such as peds and visor. Initially we have done frame conversion from these datasets.

A. Tabular column 1: Dataset details

Dataset	Format	No of frame	Size
Peds	Mpeg	1036	5.43MB
Visor	Avi	960	4.52MB

We have been working on different dataset formats which is very opt for segmentation and counting. From these datasets we can easily worked out and get reliable results.

IV. FLOW CHART



A. Pre-processing

Pre-processing is the first step in background subtraction algorithm. The purpose of this step is to prepare the modified video frames by removing noise and unwanted objects in the frame in order to increase the amount of information gained from the frame and the sensitivity of the algorithm. Pre-processing is a process of collecting simple image processing tasks that change the raw input video into a format that can be processed by subsequent steps. Pre processing of the video is necessary to improve the detection of moving objects.

In our case, a camera is used to capture a short video of human moving to and fro within a specified indoor area using avi format. The video is then processed using Matlab program. The video which are true-color image RGB are first converted to gray-scale intensity image by eliminating the hue and saturation information while retaining the luminance.

Fig(4).Preprocessing

B. Background Modeling

Background modeling and subtraction is a core component in motion analysis. The idea behind such module is to create a probabilistic representation of the static scene that is compared with the current input to perform subtraction. Background modeling is an important stage of any background subtraction algorithm. There are many research work done on developing a background model that is robust against environmental changes in the background, however sensitive enough to identify all moving object of interest. There are a few background techniques used and it can be classified into two broad categories namely the recursive and non-recursive.

Non recursive technique uses a sliding window approach for background estimation. It stores a buffer of a certain number of video frames and estimates the background image based on the temporal variation of each pixel within the buffer. This technique does not depend on the history beyond those frames stored in the buffer. The storage requirement can be significant if a large buffer is needed to cope with slow moving traffic. some example of some of the non recursive techniques are frame differencing, median filter, linear predictive filter and Non parametric model.

Recursive techniques do not maintain a buffer for background estimation, they recursively update a single background model based on each input frame. As a result, input frames from distant past could have an effect on the current background model. As compared with non recursive techniques, recursive techniques require less storage but any error in the background model can linger for a much longer period of time. Most schemes include exponential weighting to discount the past and incorporate positive decision feedback to use background pixels for updating. Recursive techniques consist of Approximated median filter, Kalman filter and Mixture of Gaussians (MOG).

In this paper, one of the simplest non-recursive techniques which is the frame differencing technique is used. This method is chosen as it is very quick to adapt to changes in lighting or camera motion. Frame differencing is an important step for counting human which consist of making a pixel by pixel absolute difference between the two consecutive frames. The result of the new image will show the difference between the two frames which represents a motion detector. If there is no motion detected, the image will be a full black one. A threshold value need to be set to eliminate those cases whereby the result of the frame differencing of the frames captured having some noisy pixels or some pixels which are very close to the background image. This threshold is very important as it acts as filter to eliminate out any unwanted information and to determine the existence of the motion.

Fig(5).Background modeling

C. Foreground Detection

Foreground Detection compares the input video frame with the background model and identifies candidate foreground pixels from the input frame. The technique uses a single image as the background model. The most commonly used approach for foreground detection is to check whether the input pixel is significantly different from the corresponding background estimate

It can be seen that this foreground detection stage is actually the detection of the foreground using the threshold and background model as described in the earlier section whereby applying the threshold is used. The difference between the background model and the 101th frame will deduce a foreground with the supposed human detected only if there is a motion.

Fig(6).Foreground detection

D. Barnard & Thompson barnard and thompson optical flow algorithm

The work of Barnard and Thompson [3] provides a good example of a more sophisticated correlationb based method. Firstly, although this method is feature-based, it does use correlation to determine similarity between points. It then uses a probabilistic relaxation labeling algorithm to help make the final determination for the flow vector.

Features are first chosen by applying a simple interest operator described by Moravec [22] to the entire image. Each feature pixel in the first image is defined as a node ai and then assigned a series of labels lj which correspond to potential matches in the second image. All potential pixels within a distance r of ai will be given such a label. The algorithm then determines a probability associated with each of the labels. Initially, the sum of the square difference (SSD) between a neighborhood surrounding the feature pixel and the candidate pixel are found. We represent these SSD values as si(l) for node ai and label l. A SSD values closer to 0 represents a more similar region, so we first invert these values to get an associated weight for each label that increases with the similarity.

$$w_i(l) = \frac{1}{1 + cs_i(l)}$$

where c is a positive constant and si(l) was the SSD value for node ai and label l. Therefore, an area which has a high similarity with the SSD measure, and therefore a low SSD value, will have a high weight value. These weights are then normalized to obtain the probabilities.

$$p_i(l) - \frac{w_i(l)}{\sum_{l'} w_{l'}(l')}$$

where for each label l for ai , we divide the weight by the sum of the weights for all the labels. This gives us the probability this label is correct. Once these initial probabilities are determined, an iterative relaxation algorithm is applied to update the probability of l based on the similarities of this displacement to neighboring displacements which are below some threshold Θ.

$$q_i^k(l) = \sum_{ajneara_j} \left[\sum_{l' s.t. \|l-l'\| \leq \theta} p_j^k(l') \right]$$

where the sum is over all the nodes aj which are within some distance of ai . For each of these close nodes, we take a further sum of all the labels l 0 which are within a threshold Θ of the considered label l. In this case, we mean the difference between the associated displacements for each label. The sum of all probabilities for nearby similar displacement labels are then used to update the current label's probability. Intuitively, we are simply performing a process similar to the smoothness constraint of Horn and Schunk. We are looking to maximize the similarity of the label's displacement within a neighborhood, providing a smooth flow field.

$$\hat{p}_i^{k-1}(l) - p_i^k(l)(\alpha + \beta q_i^k(l))$$

where α, β are parameters controlling how much influence the neighboring values should have. p k i (l) is the probability associated with label l at iteration k. q k i (l) is the previous measurement of the sum of similar label's probability in a surrounding neighborhood for iteration k. We use these values to compute a new probability for label l at iteration k + 1. Finally, the probabilities are normalized.

$$p_i^{k+1}(l) = \frac{\hat{p}_i^{k+1}(l)}{\sum_{l'} \hat{p}_i^{k+1}(l')}$$

This process is then repeated for a number of iterations or until stabilization occurs. The labels for each node are then checked and the label with the highest probability will be chosen, if it exceeds some threshold, as the barnard and thompson optical flow vector.

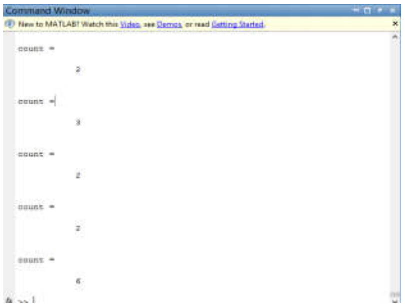
E. Counting

Our fundamental problem comes down to estimating the number of items in a dataset that satisfy a predicate or belong to a group. These counts can be used to answer aggregate queries or estimate selectivity. We explore two methods for counting: a label-based approach and account-based approach.

The label-based approach is based on traditional sampling theory. We sample tuples and ask the crowd to label the category assigned to each tuple (e.g., whether a photo is of a male or a female)until we achieve a desired confidence interval around the frequency of each category. The count-based approach displays a collection of items to a worker and asks them to approximate how many of the items fall into a particular category (e.g., the number of images with males displayed). Both approaches result in estimates of the true frequency of each category that, in the absence of faulty worker responses, converge on the true frequency. Because in practice all crowd sourced worker output has some uncertainty, all crowd-powered techniques result in approximate answers. We study the convergence rate and accuracy of various approaches in Section 6. We assume that categories are known ahead of time (e.g., we know the domain of the group by attributes). One interesting line of research lies in how we can determine all of the distinct categories covered by our estimation technique. Various projects explore how this can be done through crowd sourcing.

Fig(7).Counting

V. EXPERIMENTAL RESULT


Command Window
New to MATLAB! Watch this Video, see Demos, or read Getting Started.

```
count =

         2

count =

         3

count =

         2

count =

         2

count =

         6

>> 
```

Fig(8).Output for counting

Thus We have used barnard and thompson optical flow algorithm and worked on dense scenario in which the people are segmented and counted without occlusions and shadow illuminations. Pre-processing, background modeling, Foreground detection and counting were the steps we employed and arrived at the above result using matlab

VI. CONCLUSIONS

We have presented a framework for detecting and segmenting humans in real-world crowded scenes which integrate both local and global shape cues. Cluttered scenes containing many occlusions render the lone use of global shape representations as interactive. Instead we aggregate local shape evidence via a codebook of local shape distributions for humans in various postures. Additionally, we found that a set of learned global posture clusters aids the segmentation process. Our experiments indicate that local shape distribution represents a powerful cue which can be integrated into existing lines of research. We presented a method for counting crowd in images using head detection with interest points based on gradient orientation. Experiments on PETS and visor databases show the potential for such an approach in different conditions of people moving in the scene, and the same head detector was tested on both the databases. There was also no fine tuning of the background subtraction parameters to obtain the best results in case of PETS and no background subtraction was used in Turin database, which makes our approach more ideal to estimate the crowd from a single frame. Nevertheless, there is a lot of scope for improvement in our approach. In our future work we would like to reduce the false detections, and incorporate additional methods to reason out occlusions in a crowded scene.

ACKNOWLEDGEMENT

No words to express our pleasure and sincere thanks to our project guide and well-wisher Ms.janani for assisting us to finish our project

REFERENCES

[1] ZATSIORSKY. KINETICS OF HUMAN MOTION. HUMAN KINETICS, 2002.B. ZHAN, D. N. MONEKOSSO, P. REMAGNINO, S. A. VELASTIN, AND L.-Q. XU, "CROWD ANALYSIS: A SURVEY," MACHINE VISION AND APPLICATIONS, PP. 345–357, 2008

[2] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," PAMI, vol. 34, no. 4, pp. 743–761, 2012.

[3] Y. J., V. S., and A. Davies, "Image processing techniques for crowd density estimation using a reference image," in ACCV, vol. 3, 1995, pp. 6–10.

[4] R. Ma, L. L., H. W., and T. Q., "On pixel count based crowd density estimation for visual surveillance," in IEEE Conf. Cybernet. Intell. Syst., vol. 1, 2004.

[5] V. Lempitsky and A. Zisserman, "Learning to count objects in images," NIPS, 2010.

[6] D. Conte, P. Foggia, G. Percannella, F. Tufano, and M. Vento, "An effective method for counting people in video-surveillance applications." in VISAPP, 2011, pp. 67– 74. [8] D. Conte, P. Foggia, G. Percannella, and M. Vento, "A method based on the indirect approach for counting people in crowded scenes," in AVSS, 2010, pp. 111–118.

[7] T. Roberts, S. McKenna, and I. Ricketts. Human pose estimation using learnt probabilistic region similarities and partial configurations. ECCV, 4:291–303, 2004.

[8] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. CVPR, 1:511–518, 2001.

[9] J. Rittscher, P. H. Tu, and N. Krahnstoever, "Simultaneous estimation of segmentation and shape," in Proc. IEEE Conf. ComputER. Vis. Pattern Recognit., 2005, pp. 486–493.

[10] L. Zhao and C. Thorpe. Stereo-and neural network-based pedestrian detection. ITS, IEEE Transactions on, 1(3):148–154, 2000.

[11] Y. Wu and T. Yu. A Field Model for Human Detection and Tracking. CVPR, 28, 2006.

[12] L. Zhao and L. Davis. Closely Coupled Object Detection and Segmentation. ICCV, 1, 2005.

[13] L. Dong, V. Parameswaran, V. Ramesh, and I. Zoghlami, "Fast crowd segmentation using shape indexing," in Proc. IEEE Int. Conf. Comput. Vis., 2007, pp. 1–8.

[14] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003., vol. 2, pp. II–459–66 vol.2, June 2003.

[15] T. Zhao and R. Nevatia, "Tracking multiple humans in crowded environment," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004., vol. 2, pp. II–406–II–413 Vol.2, June-2 July 2004.

[16] V. Rabaud and S. Belongie, "Counting crowded moving objects," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, vol. 1, pp. 705–711, June 2006.

[17] D. Kong, D. Gray, and H. Tao, "Counting pedestrians in crowds using viewpoint invariant training," in 18th International Conference on Pattern Recognition, ICPR, pp. 1187– 1190, 2006.

[18] P. Kilambi, E. Ribnick, A. J. Joshi, O. Masoud, and N. Papanikolopoulos, "Estimating pedestrian counts in groups," Computer Vision and Image Understanding, vol. 110, no. 1, pp. 43 – 59, 2008.

***[19]** Z. Kim, "Real time object tracking based on dynamic feature grouping with background subtraction," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., Los Alamitos, CA, 2008, pp. 1–8.*