New Insights of Background Estimation and Region Localization

Author: Htet Htet Lin
Affiliation: Software Department Computer University (BanMaw), Myanmar
*Corresponding Author: htethtet.linnnnn@gmail.com

Abstract - Subtraction of background in a crowded scene is a crucial and challenging task of monitoring the surveillance systems. Because of the similarity between the foreground object and the background, it is known that the background detection and moving foreground objects is difficult. Most of the previous works emphasize this field but they cannot distinguish the foreground from background due to the challenges of gradual or sudden illumination changes, high-frequencies background objects of motion changes, background geometry changes and noise. After getting the foreground objects, segmentation is need to localize the objects region. Image segmentation is a useful tool in many areas, such as object recognition, image processing, medical image analysis, 3D reconstruction, etc. In order to provide a reliable foreground image, a carefully estimated background model is needed. To tackle the issues of illumination changes and motion changes, this paper establishes an effective new insight of background subtraction and segmentation that accurately detect and segment the foreground people. The scene background is investigates by a new insight, namely Mean Subtraction Background Estimation (MS), which identifies and modifies the pixels extracted from the difference of the background and the current frame. Unlike other works, the first frame is calculated by MS instead of taking the first frame as an initial background. Then, this paper make the foreground segmentation in the noisy scene by foreground detection and then localize these detected areas by analyzing various segmentation methods. Calculation experiments on the challenging public crowd counting dataset achieve the best accuracy than state-of-the-art results. This indicates the effectiveness of the proposed work.

Keywords: Mean Subtraction Background Estimation (MS); Background Subtraction; Segmentation; Crowd Counting

I. INTRODUCTION

The accurate foreground detection in the heavy cluster scenes is an essential task in many applications such as visual surveillance, operational, traffic and safety monitoring processes (Metro/railway stations, Airports, Bus Stand, Shopping Mall, Museums, Casinos, Bank, Classroom, Parking, Park, Zoom, Street Monitoring, etc.). Since this various applications consist of many people, this cause the heavy situation of people. Due to these heavy condition, many key issues have still exist. These are different articulations, various appearances, different sizes, static/dynamic occlusion, illumination change and strong perspective.

Subtraction many people, especially in large dense crowd area is very essential and an active area of research. Most of the previous works had been accepted the use of background subtraction a technique in a visual observation system that represents some important issues, including the calculation of the accuracy of cost and background estimates. They observed the image in the method of subtracting the background and compared with the estimated value of the background images. That does not contain objects, can be accessed through the background Modeling algorithm. Background subtraction technology that represents a visual observation system.

According to the literature, there are numerous methods for the background subtracting. Most of them are developed in accordance with the assumptions that visual differences are displayed in the foreground and in the background. Consequently, the front part can be detected once a good
background model was obtained. But, there exist some situations, the difference in intensity between the foreground and the background is very small, and existing methods can’t easily detect the foreground. This fact motivates to learn a framework that can handle this problem especially under illumination condition. Given a video, the two main objectives of the proposed system are:

1) to find out the clarity foreground, which can assist to get the better performance in the next steps and

2) to obtain the accurate person foreground region.

The proposed system first calculates mean subtraction background estimation (MS) to extract accurate foreground detection on all the three channel. The calculation of MS can avoid the illumination problems. After that analyze various segmentation methods to localize the region of foreground people area. Among them, k-mean clustering algorithm, which will be redundant, get the best localize results. Unlike other works, this paper uses the making MS background instead of the normal organization. For accurate detection and localization, this paper evolves the new insight, effective terminology of background estimation and region localization system has been addressed. This paper also shows the distinct progress in background estimations, foreground detections and region localizations that has been achieved by proposed methods. This paper organizes as follows. Section II presents the earlier methods and technologies and their applicable work of the crowd background estimation and localization system has been discussed. Section III demonstrates the proposed approach to contribute in crowd sensing field. Section IV investigates and discusses the experimental performance results compared with state-of-art results. The rest section, Section V has been attempted the conclusion and concerns.

II. RELATED WORK

Background subtraction method is one of the most commonly used methods for extracting foreground object from the video sequence. Many researcher used this method because of increasingly demands from a variety of domains such as human-computer interfaces, video surveillance monitoring and security, public safety and action recognition environment.

The simplest solution to modeling the background of the scene is to use statistical methods [1] and [2]. In the case, where a fixed object is retained for a long time, the limitation of the statistical method may incorrectly use the foreground object in the background. The Gaussian Mixture Model (GMM) is the most commonly used method for estimating backgrounds, originally proposed by Stauffer et al. [3]. In this method, each pixel value is estimated using a separate Gaussian mixture that is continuously studied by online approximation.

Several improved versions [4] to [9] were proposed as the major contributions to object detection methods using background subtraction methods. For example, Zivkovic [4] considers an improved adaptive GMM in which parameters and components of a hybrid model are continuously selected for each adaptive pixel. Lee et. al. [5] proposed to increase the speed of convergence without taking into account the stability of the model. There is another development of GMM. Elqursh et. al. [7] recently proposed to describe the tracking content in the low-dimensional space, and then synthesizes it through the GMM in each forthcoming frame. To exclude fluctuations in lighting and noise in intelligent video surveillance systems, Li et. al. [8] proposed a new solution based on GMM.

These method has three key elements: (1) explicit analysis of the spectrum reflection model, (2) online maximization algorithm and (3) two-stage foreground detection algorithm. After the probability regularization, a method based on the GMM Dirichlet method [9] is proposed to estimate the background distribution of pixels.

Although improvements in GMM improvements have been proposed in complex scenarios, they still have general limitations, such as fault detection and parameter estimation problems in high-speed mobility. To avoid the problem of finding the right form for the probabilistic model, the researchers focused on nonparametric methods of background modeling. The real-time algorithm in [10] quantizes the values of the background pixel in the codebook, which describes the base model of the compressed shape of many frames. Despite the high performance in real-time environments, the drawbacks of the codebook approach include long periods of time for creating models and using large memory for storing code words. Another highly used nonparametric method in subtracting the background is the kernel density estimate (KDE) [11]. This method uses a histogram to evaluate the probability density function to override the current value background pixels. Liu et. al. [12] proposed a hybrid model, integrated with KDE and GMM, to construct a function of the background probability density and moving target models.

Although KDE-based methods can provide fast response to high speed scenes, due to the first-in-first-out order, processing the companion events at different speeds of the method is limited. In recent years, background modeling techniques, known as background visual extraction (ViBE) by Barnich [13], have determined whether a pixel belongs to the background and unexpectedly compares its intensity to the neighborhood. Although ViBE can provide satisfactory detection results compared to existing methods, the problem
is problematic under harsh conditions, such as scenes with
darker backgrounds, shadows, and frequent background
changes. In the research framework of Gruenwedel et al.
[14], from the perspective of minimizing memory
requirements, a two-layer background model is proposed,
one with low adaptation rate with long-term background and
the other with high adaptation rate with short-term
background. The radial basis function (RBF) of the artificial
neural network is described in [15] as a multi-level
generation of inappropriate learning processes.

Although a number of progresses have been done in
background subtraction methods, it still remains challenges
such as illumination, inter and intra class variations, etc.
which are making the counting process to be immensely
difficult.

To tackle the issue illumination and to develop inaccurate
foreground localization error, this paper proposes a best
accurate background subtraction and localization framework
by segmenting foreground region with gradient k-mean after
detecting with the mean subtraction estimation background.
The paper also highlights the analysis to find out the best
strategies with some prominent existing methods.

III. PROPOSED APPROACH

The pseudo code form representation of the proposed
approach is depicted in Fig. 1.

Algorithm: Overall process of the proposed work for people counting
Input: Images from the dataset: input
Output: Images after classification results: res
If input ≠ 0 then
   Error
Else
   A = calculateMeanBy(input); //Calculate mean shape background
   B = foregroundDetection(A)
   C = k-mean(B); // foreground segmentation
   Res = performanceCalculation(C);
End

Fig. 1. Pseudo code form representation of the proposed work

A. Background Estimation

The foreground detection is an important influence on the
detection accuracy and is a difficult task that cannot be
carried out accurately in many scenes. To detect the
foreground as serve as the candidate detections more
accurately, the clarity background is very important.

In many previous work, they took the first frame as the
original background from the input video sequence. But this
prior knowledge can't handle the nature of PET 2009 dataset.
The challenges of this dataset include severe occlusion,
clutter and similar appearance of people. This is shown in
Fig. 2. Testing results (assuming the first frame as the
original background like the previous work) are shown in Fig.
3.

Fig. 2. Example images of PET 2009 dataset

A well-estimated background model and an accurate
binarization algorithm is needed in order to provide a reliable
foreground image for extracting the target features. Due to
the nature of PET 2009 dataset, there have many challenges.
In here, the problem requirement is find out if the
challenging PET dataset is used. This fact motivate us to
develop the proposed framework. Unlike other, the
background is not clear and need to adapt.

In this study, given a video, the proposed system first
calculate and enhances the original input image. To detect
foreground as serve as candidate detections more accurately,
the clarity background is very important. In contrast to
previous work, they either run an object detector or perform
background subtraction with various approaches. They took
the first frame of the video. In this paper, the proposed system calculates mean shape on the whole frames to apply foreground detection. Like the other is that the system need to set threshold in the foreground detecting process. The proposed system can get the foreground detections with reasonable recall.

The mean subtraction background estimation is calculated on the whole frames to apply foreground detection by the following equation:

$$M = \left\lfloor \frac{\sum_{i} I_o(i)}{3} \right\rfloor$$ (1)

where $I_o$ is the image object of every frame $i$, $M$ is the mean subtraction background estimation, and the value 3 in eq (1) represents for the three color channels. Like other, the foreground image is then binarized based on a threshold to obtain the foreground pixels due to the following equation:

$$|\sum_{i} I_o(i) - M| > T$$ (2)

where $I_o$ is the image object of every frame $i$, $M$ is the mean subtraction background estimation, and $T$ is the threshold.

The notice fact is that there will often be significant overlap between the foreground and detections, which is complicated for the optimization process. This is hard to achieve convergence in practice. In reality, foreground detection is difficult because of the occlusion, illumination variations, shadows, etc. So, the proposed system use the mean shape estimation background to avoid the noise illumination. As a discussion, we set the threshold value as 60. Most of the people located regions will be cover up with the white pixels after putting the threshold value over 60.

This fact highlight the strong clarity of our proposed contribution point of background subtraction. After that, the proposed system use the morphological operation that need to estimate foreground. Most of the people located regions will be cover up with the white pixels after finishing the dilation process. As a discussion, we set the various values of the radius of the structure element. The insight is that the value 2 is the best for dilation process. The more value is the fewer regions that contain the people.

To vality of the proposed background subtraction method, this paper tests the results of the various previous methods compared with the proposed system. This is shown in Fig. 5.

![Fig. 5. Foreground detection results of the proposed system compared with the previous work](image)

### B. Region Localization

This paper extract the foreground area and segment these foreground localization by k-mean clustering. The proposed system adaptively sets the threshold to segment the foreground people regions. The system also estimates the number of person in each of the foreground segmented areas. These detected people gradients are clustered to get the whole area of people located. K-means is an essential technique in vector quantization, formerly used in signal processing. It is calculated as the following equation:

$$\arg \min_{\mu} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 = \arg \min_{\mu} \sum_{i=1}^{k} |s_i| Var_i$$ (3)

where $x$ is set of observation, $\mu_i$ is the mean of point in observation set $S_i$. This clustering use intends to find clusters of comparable white pixel, while these represent the clutter of people that have different shapes and size. We use the value 2 for the clutter centroid number. This is the best to clutter all the people location. Then, the region proposal of the original image is segmented.

Fig. 6 shows the result images of the various segmentation methods. According to the analysis of segmentation results, the Kmean segmentation is the best for all frames of the
proposed system. Hence, we got the new insight that the k-mean segmentation is the best to choice.

Fig. 6. Result images of the various segmentation methods

IV. EXPERIMENTAL RESULT AND DISCUSSION

A. Dataset

To calculate the framework, we evaluate the experiments on the challenging crowd counting dataset, PET 2009[13]. It used as the crowded counting domain dataset. This dataset consisted of several sections to test the different surveillance scenes (indoor and outdoor scenes). This video sequences have good crowded properties and challenging variations such as both walking and running people, illumination changes and diverse crowd densities and sizes. This sequence also has the ground truth annotations [11].

B. Experiments

In this experiment, the proposed system uses Recall, Precision and F1 measure to compare with the state-of-the-art methods in the performance of foreground detection. This paper calculate the precision, recall and F1 measure by the following equations:

\[
Precision = \frac{TP}{TP + FP} \tag{4}
\]

\[
Recall = \frac{TP}{TP + FN} \tag{5}
\]

\[
F1 = \frac{2TP}{2TP + FP + FN} \tag{6}
\]

where TP is true positive, FP is false positive, TN is true negative and FN is false negative from the result image and ground truth. All metric values range from 0 to 1, with higher value pointing out the higher accuracy.

All of the experiments are performed on the laptop operating Windows 7 with a 3.1 GHz Intel Core i7 CPU and 4GB RAM. MATLAB R2016 was the software for simulation.

C. Results

The performances of the proposed system compares the previous different methods on challenging PETS dataset videos are reported in Table I. This table shows the quantitative metrics comparison between the proposed system and three state-of-the-arts. A highlighting fact is that the proposed system got the best value of precision, recall and f1 measure. We will evaluate this fact as a future work to assessment and clarity of our concept correctness.

Based on calculating operation on mean subtraction background estimation, foreground object pixel is extracted. K-mean clustering is also used to learn to localize the accurate foreground area and the related pixel. Finally, performance comes out by mapping the resulted out foreground object with the ground truth foreground object of the dataset. In this way, the system can compute the precision, recall and f1 measure performance. The evaluation metrics are calculated thanks to the ground truth with the binary classification. The proposed method performs quite well on the most sequences. Fig. 7 also shows some visualization of the foreground detection results, compared the proposed system with the three start-of-the-arts as an analysis.

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed System</td>
<td>0.9506</td>
<td>0.9860</td>
<td>0.9680</td>
</tr>
<tr>
<td>NIC [11]</td>
<td>0.947</td>
<td>0.953</td>
<td>0.950</td>
</tr>
<tr>
<td>I-GMM [12]</td>
<td>0.981</td>
<td>0.644</td>
<td>0.778</td>
</tr>
<tr>
<td>ViBE [13]</td>
<td>0.913</td>
<td>0.956</td>
<td>0.934</td>
</tr>
</tbody>
</table>
V. CONCLUSION

A robust system for background estimation with region localization is intended to solve the occlusion, illumination and various orientation. First input image is made preprocessing to avoid the noise and illumination. Then, clear background is taken out to get the discriminate foreground detection. The localized region containing people is extracted by foreground segmentation methods. Finally, the experimental result is calculated. Experimental result are significantly outperformed than the state-of-art methods.

As a discussion, we tested the proposed system only on background estimation and localization framework. In the frame difference method, the result come out with many black pixels i.e. missing the foreground object. In the HSV color background subtraction method, the foreground objects are not clear. In the adaptive background learning method, the foreground results contain many noise that make the decreased outcome. In the three frame difference method, they got the best performance than the other previous system. This paper motivate the way to get the clarity background and also get the high performance than the three frame difference method. The proposed mean subtraction background estimation got the highest result among them. The training time also took a very few time so the proposed system gets a balance tradeoff between processing time and results. As a future work, we will evaluate the proposed system on the field of people security system (fall down detection, thief-alarm detection, abnormal detection system, etc.) due to the finding of the foreground clarity. According to the proof of the analysis result, the proposed background subtraction will use to develop the people detection and counting system in near future.

VI. DECLARATION

Author have disclosed no conflicts of interest and the work was self-funded.

VII. ACKNOWLEDGEMENT

The author wishes to thank their parents and relatives whose inspirations make their work hard and who are always willing to give the authors all moral support.

REFERENCES


T. Huynh-The et al., “Background subtraction with neighbor-based intensity correction algorithm”, in Research Gate, 2015.
