Continuous Adaptive Motion-based Person Understanding System (CAMPUS)

I. Gerg, J. Gibbs
Department of Electrical Engineering – Pennsylvania State University, University Park

Abstract - The authors present an algorithm for detecting and segmenting moving people in a campus scene. The people are observed by a stationary camera. A Gaussian mixture model is used to distinguish a moving foreground from the stationary, albeit dynamic, background. Further processing determines location and number of people in the scene. Two example campus scenes are processed to evaluate the efficacy of the proposed algorithm. These scenes are evaluated against an Expectation Maximization (EM) based algorithm and are shown to give comparable results in less computing time.

1. Introduction

Because of their rich informative value as compared to static images, motion based cues from a sequence of images have been used in many modern segmentation algorithms. In their algorithm the authors utilize the Adaptive Backgrounding for Motion Tracking algorithm using an online mixture model as first proposed by Stauffer and Grimson [1]. The algorithm is able to segment motion as observed from a stationary camera. The algorithm is sophisticated enough so as to discern foreground images atop a dynamic background environment. For instance, older background/foreground segmentation routines would not be able to account for a flashing light in the background of the scene due to its continual changing of pixel values. The online mixture model overcomes these problems and allows for a variety of background dynamics while still keeping the ability to accurately distinguish foreground images. The output of the online mixture model algorithm is pixel-level cues of bodies of "foreground objects". It is hoped that these pixels belong to people in our sequence of images, however it is shown that the online mixture model yields pixel level cues that are too noisy for discernment of upright people. Our algorithm aggregates these cues into a higher-level framework of objects. In this framework, we will distinguish between people and other objects not of interest.

Our higher-level processing consists of two steps that seek to identify people from the cues given by the online mixture model. The first performs connected-component analysis on the pixels eliminating noisy groups of pixels that belong to objects too small in size to be a person of interest. In the final step we propose a recursive (conditional split until idempotent) algorithm based upon the dimensions of an upright person. This technique allows for the discernment of closely spaced people who are considered to be one in simple connected component analysis. It is seen that shadows of moving individuals can cause problems for simplistic discernment methods like the aforementioned connected component analysis. As a comparative measure, we compare the higher level processing of our algorithm with an Expectation Maximization (EM) based algorithm seeded by the foreground as determined by the connected component analysis. Each connected "blob" was fit with a bounding box and the initial guess of foreground constituted the first of two regions (background and foreground) to be segmented by EM. It is seen that while in some cases the EM algorithm performs as well as the proposed algorithm, it fails to converge at all in others due to the assumptions made in the EM segmentation algorithm.

2. Algorithm Details

The authors chose a three-step process to segment upright moving people from the campus scenes. The first step determines the foreground of the scene via motion cues. This was done using the Gaussian Mixture Model proposed in [1]. To simplify the computational complexity of the mixture models, grayscale images were used instead of the three independent color channels as first proposed in the paper. After converting the original color images to grayscale, the algorithm determines the background and foreground pixels and outputs a binary representation of the foreground. The second step involved computing connected components from the binary video to determine significant regions of interest based on area. The third step involved a high level analysis of the interesting regions found in step two. The output of this step resembles binary motion blobs. A recursive technique was performed on the binary motion blobs of step two to distinguish upright, moving people from large object scene clutter.

For each of these steps there is a determination of the types of parameters to be used, each of which will be expounded upon in detail in the coming sections. The online mixture model requires three parameters: the number of Gaussian mixture models, the adaptation constant and the minimum percentage of background to be accounted for by the models. There are two remaining parameters used during the high level analysis to determine the upright, moving people in the scene: minimum area and height of people in the scene. In general, the authors found these two groups of parameters to remain very similar among different scenes of similar types of video (people moving in campus scenes). It is believed that in other scenarios they could be tweaked to provide adequate performance.

2.1 Initial Foreground Segmentation Via Motion Cues

The foreground was initially segmented using the Gaussian Mixture Model presented in [1]. The authors chose this method because of its demonstrated efficacy and computational efficacy. These two factors have led to the methods wide acceptance among researchers in the computer vision community.

This segmentation method works by modeling each pixel by a mixture of Gaussian distributions. Each Gaussian is assigned a weight based on its likelihood to be a good model for the background. This is accomplished by weighting each of the Gaussians by how frequently they are observed in the past. This weight is an average with an exponentially decaying window. Specifically, each mixture's weight is updated in the following manner:

$$\omega_{k,t} = (1-\alpha)\omega_{k,t-1} + \alpha(M_{k,t}), \qquad (1)$$

Where $\omega_{k,t}$ is the weight for the k^{th} mixture at time t and α is the "learning coefficient" that controls how long previously seen mixture components remain in "memory". For the specific results shown in the following sections, α was set to 0.001 while k, the number of Gaussian mixtures was set to 4.

When a pixel is observed that does not belong to any of the existing mixture models it is most likely a "foreground" pixel. However, if it remains constant for long enough it can be incorporated into the background. Therefore if a new pixel is not within a certain amount of standard deviations of the existing mixtures, a new mixture is created replacing the least likely prior mixtures with arbitrarily high mean and variance (while the new mean and variance are arguably more tuning parameters for this method we view them as not as significant as the other parameter selections of this approach).

As the algorithm proceeds and each pixel has been assigned to an appropriate Gaussian in the mixture model, the background pixels are determined by ordering the Gaussians by weight and accounting for the minimum amount of data representing the background (T),

$$B = \arg\min_b \left(\sum_{k=1}^{b} \omega_k > T \right). \qquad (2)$$

The minimum portion of the data that should be accounted for by background, T, is set to be 0.5 in scenes that are depicted in the following sections.

Typically the Gaussian mixture model requires an initial representation of the background. For the scenes analyzed, the algorithm provides adequate results without the need training data. In this way, the authors do not calculate an initial probabilistic representation of the image background. In fact, 'training' the Gaussians with ten averaged frames of the entire scene does not provide significantly better results. The difference in segmentation between the methods is nearly imperceptible.

2.2 Connected Component Analysis

Eight-way connected component analysis was performed on each binary image resulting from the output of the Gaussian mixture model. Binary blobs smaller than an area threshold were removed from the image. For the following segmented images the minimum area that would constitute a person was set to be 550 pixels.

2.3 High Level Analysis

A recursive technique was performed on the binary motion blobs of step two to distinguish upright, moving people from large object scene clutter. Mainly, an 'upright, moving person' was discriminated by height in addition to area. The height threshold of a person was set to be 30 pixels.

In many cases, the people in the scene were joined by their shadow to create one large connected component blob. To remove the shadow portion of the connected blob and to discriminate between the two people, height was used. This was done by summing down the binary columns of each connect blob matrix to form a vector. The assumption of upright people is now exercise as we remove columns in the blob which have values less than our height threshold. The shadows had a low height and the upright person had a large height. A threshold was

used to distinguish the two. When two people were attached by one connected blob through one of their shadows, it is necessary to determine where break the two apart. Because of this, a recursive approach was used by which each blob was broken into fundamental parts and then each of those parts were analyzed until a base case was reached. The base case blob met the area and height requirements.

The recursive approach worked by first performing connected component analysis on a morphological dilation of the initial Gaussian mixture segmentation. Then, each blob was processed using the threshold technique described above. This process was then repeated for the resulting blobs until a steady state was reached for a given frame. The blobs remaining become the final segmentation.

3. Evaluation

The authors choose a standard set of campus video sequences encoded in MPEG format. The scenes feature students walking, sitting, and bicycle riding. Given an initial motion cue, we wish to segment an individual for the remainder of the scene regardless of their motion or lack of motion. We wish to distinguish between human motion and motion of other objects such as vegetation movement (wind) and other background distracters.

As an additional metric, we seek to determine the number of individuals present in a scene. This leads to the desires of a real-time tracking system in which the numbers of individuals present in the scene is a fundamental requirement. We determine the number of individuals in the scene by determining the number of total objects in the scene after the recursive algorithm has broken apart connected human objects.

As a measure of comparison we compare our higher-level processing with that of the Expectation Maximization (EM) algorithm as presented in [Forsyth, Ponce]. The EM algorithm is performed on objects enclosed by blobs that meet our height/area requirements. It is performed in the smallest rectangular bounding box about the blob. Two regions are used in the segmentation with the assumption of a background region with a normal distribution and a different foreground region with a separate normal distribution. The initial mean and variance of each is computed from the first guess as to where the person is in the scene. This guess corresponds to the binary mask of foreground pixels from the Gaussian mixture model. The EM algorithm then iterates until convergence. In the images of Table 4, it can be seen that in some instances it provides very good results. But in others, it is seen that the EM completely fails. This is because the gray values are far from the assumption in the EM algorithm of two very different homogenous regions constituting the foreground and background. Additional filtering could be done to try to mitigate these problems and force the EM algorithm to converge to better results (vertical smoothing of the image would result in more homogenous regions). Nevertheless, the straightforward EM segmentation as proposed in [2] yields results that when they converge are comparable to those of the CAMPUS method, however it is ineffective in converging in nearly half the frames for all of the people in the scene that the CAMPUS algorithm detects.

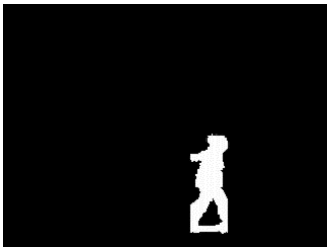
An important metric in the evaluation of the CAMPUS algorithm the computational efficiency. For 320x240 images the throughput of the CAMPUS algorithm is 125 frames/minute as coded in MATLAB on a 1Ghz Pentium Celeron with 512MB RAM running Windows XP. In comparison the EM based segmentation provides 115 frame/minute throughput on the same hardware.

4. CAMPUS Results

Two campus scenes were analyzed using the CAMPUS algorithm. These scenes involve students walking in and out of the stationary camera's view.

The first scene scene, oldmain, depicts students sitting in the grass and walking on a sidewalk. The algorithm successfully segments the upright, walking people and does not include those sitting in the grass. Table 1 displays a selected frame from the results.

(a)



(b)

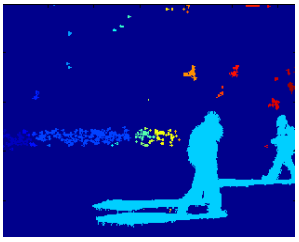
(c)

Table 1. Depiction of segmentation steps of oldmain video. (a) Initial segmentation via Gaussian mixture model algorithm. (b) High level segmentation via our

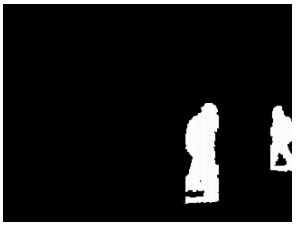
recursion technique. (c) An overlay of the segmentation onto the colored frame.

The recursive process CAMPUS uses to segment connected components further into moving people is depicted in Table 2. From the pictures, we can see 2.c is segmented correctly despite the two groups of people joined through connected components as shown in 2.b.

(a)



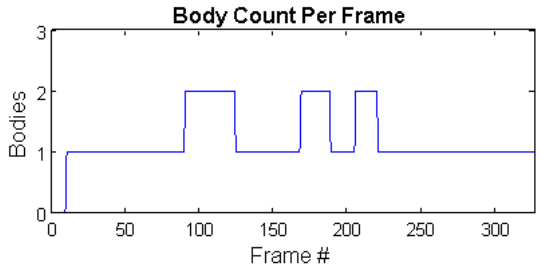
(b)



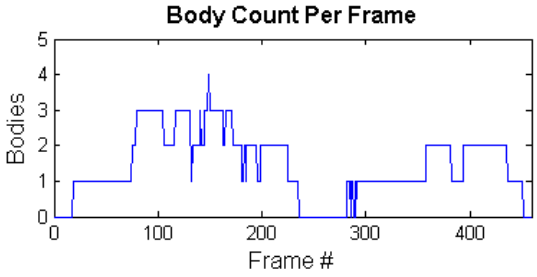
(c)

Table 2. Demonstration and solution of the connected components problem. (a) Initial segmentation via Gaussian mixture model. (b) Connected components of a. (c) The output of the recursive algorithm we developed to deal with overlapping motion blobs.

The CAMPUS algorithm also determines the number of people's bodies in a scene at each frame. This was used as an additional metric for detection performance. Table 3 depicts the results. As shown, the oldmain video segmented much better than the pattee video. This was due to confusion in the Gaussian mixture model because only grayscale frames were processed. This was done for performance considerations.



(a)



(b)

Table 3. The body counts per frame of the oldmain and pattee videos. (a) oldmain video results. (b) pattee video results. Notice the results of a are much cleaner than those of b due to background/foreground confusion during intial segmentation (Gaussian mixture)

The second scene, pattee, depicts students walking on two sidewalks. The algorithm has generally good performance and separates the people. However, because the initial foreground segmentation was done using grayscale video, some of the people are not cleanly segmented from the scene. This is an inherent problem with the Gaussian mixture model. If the moving object resembles the background behind it, it may be labeled as background also. Table 5 depicts this problem.

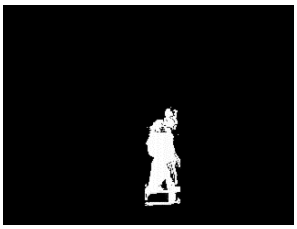
We also evaluated the performance and EM based method for person segmentation. This worked reasonably well as shown in Table 4.

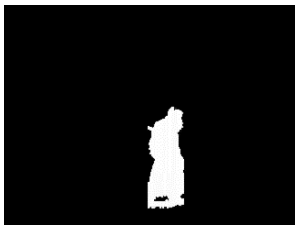

(a)


(b)

Table 4. Initial segmentation and final segmentation overlay of the pattee video. (a) Initial segmentation via Gaussian mixture model. (b) Final segmentation overlayed onto color video.

The results of the comparison of the EM algorithm against CAMPUS are shown in Table 5. Notice, 5.b depicts a much cleaner segmentation than that of 5.a. There were instances in the processed video where the EM algorithm failed entirely marking and entire person as background. This is depicted in 5.d as the girl in red is not identifies as a person.


(a)


(b)


(c)


(d)

Table 5. A comparison of the EM segmentation versus the CAMPUS algorithm. (a) EM segmentation. (b) CAMPUS segmentation. (c) Overlay of EM segmentation onto color video. (d) A missed detection using EM.

The algorithm fails when two people walk closely together. This is visible in the oldmain video. This is an inherent problem of the CAMPUS algorithm.

5. Conclusions

The CAMPUS algorithm provides good segmentation of upright moving people in a collegiate campus scene. It does so in a time frame that is near-real time considering none of the code used is optimized and is being implemented in MATLAB, a notoriously slow platform. Segmentation using the CAMPUS algorithm is comparable with that of EM when EM is initialized with the lower-level data generated by the Gaussian mixture model and the connected component labeling. CAMPUS performs faster than even a simplistic algorithm using EM to perform the finest level of segmentation. More robust versions of the EM algorithm that have more than two layers of segmentation and filter the image would require more processing time to overcome the problem of the algorithm converging so that all the points in the scene belong to the same region. CAMPUS is somewhat less effective than EM when it works in that the dilation/erosion morphological operators lead to segmentations which are less sharp than those of the EM algorithm. Furthermore people's feet tend to be cut off, a non-desirous result of the resolution of two different people connected by foreground effects by CAMPUS. Nevertheless, it is seen that CAMPUS provides good segmentation in a variety of scenes in a manner far more robust than the simple EM approach.

6. References

[1] **"Learning patterns of activity using real-time tracking,"** *Stauffer, C.; Grimson, W.E.L.;* Pattern Analysis and Machine Intelligence, IEEE Transactions on , Volume: 22 , Issue: 8 , Aug. 2000 Pages:747 – 757.

[2] **"Computer Vision -- A Modern Approach,"** *Forsyth, D.; Ponce, J.*; Prentice Hall of India. New Delhi, India. Pages: 354-372.

Video sequences and source code available at:
http://www.gergltd.com/users/isaac.gerg/cse586/project1