Multi-Camera Monitoring of Human Activities at Critical Transportation Infrastructure Sites

Final Report

Prepared by:

Evan Ribnick
Ajay J. Joshi
Nikolaos P. Papanikolopoulos

Artificial Intelligence, Robotics and Vision Laboratory
Department of Computer Science and Engineering
University of Minnesota

CTS 08-08

Technical Report Documentation Page

1. Report No.	2.	3. Recipients Accession No.	
CTS 08-08			
4. Title and Subtitle Multi-Camera Monitoring of Human Activities at Critical Transportation Infrastructure Sites		5. Report Date June 2008 6.	
7. Author(s) Evan Ribnick, Ajay J. Joshi, Nikolaos P. Papanikolopoulos		8. Performing Organization Report No.	
9. Performing Organization Name and Address Artificial Intelligence, Robotics and Vision Laboratory Department of Computer Science and Engineering University of Minnesota 200 Union Street SE Minneapolis, Minnesota 55455		10. Project/Task/Work Unit No. CTS Project # 2006018 11. Contract (C) or Grant (G) No.	
12. Sponsoring Organization Name and Address Center for Transportation Studies University of Minnesota 511 Washington Avenue SE, Suite 200 Minneapolis, Minnesota 55455		13. Type of Report and Period Covered Final Report 14. Sponsoring Agency Code	
15. Supplementary Notes http://www.cts.umn.edu/pdf/CTS-08-08.pdf			

16. Abstract (Limit: 200 words)

The goal of this work is to provide a system which can aid in monitoring crowded urban environments, which often contain tight groups of people. In this report, we consider the problem of counting the number of people in the scene and also tracking them reliably. We propose a novel method for detecting and estimating the count of people in groups, dense or otherwise, as well as tracking them. Using prior knowledge obtained from the scene and accurate camera calibration, the system learns the parameters required for estimation. This information can then be used to estimate the count of people in the scene, in real time. Groups are tracked in the same manner as individuals, using Kalman filtering techniques. Favorable results are shown for groups of various sizes moving in an unconstrained fashion.

17. Document Analysis/Descriptors		18. Availability Statement	
People tracking and counting, crowd monitoring, computer vision, tracking		No restrictions. Document available from: National Technical Information Services, Springfield, Virginia 22161	
19. Security Class (this report) Unclassified	20. Security Class (this page) Unclassified	21. No. of Pages 42	22. Price

Multi-Camera Monitoring of Human Activities at Critical Transportation Infrastructure Sites

Final Report

Prepared by:

Evan Ribnick
Ajay J. Joshi
Nikolaos P. Papanikolopoulos

Artificial Intelligence, Robotics and Vision Laboratory
Department of Computer Science and Engineering
University of Minnesota

June 2008

Published by:

Intelligent Transportation Systems Institute
Center for Transportation Studies
University of Minnesota
200 Transportation and Safety Building
511 Washington Ave SE
Minneapolis, MN 55455

Table of Contents

List of Tables

List of Figures

Executive Summary

The goal of this work is to provide a system, which can aid in monitoring crowded urban environments, which often contain tight groups of people. In this paper, we consider the problem of counting the number of people in the scene and also tracking them reliably. We propose a novel method for detecting and estimating the count of people in groups, dense or otherwise, as well as tracking them. Using prior knowledge obtained from the scene and accurate camera calibration, the system learns the parameters required for estimation. This information can then be used to estimate the count of people in the scene, in real time. Groups are tracked in the same manner as individuals, using Kalman filtering techniques. Favorable results are shown for groups of various sizes moving in an unconstrained fashion.

CHAPTER 1

INTRODUCTION

Using computer vision techniques to monitor humans in urban environments is a difficult problem. In many human and group activity monitoring applications, it is important to get a sense of scale in the scene and estimate the number of people present. However, most of the cues that aid in detection and tracking of individuals like shape, texture, and appearance break down in the case of crowds, especially when low-resolution surveillance cameras are used. As such, it can be helpful to use an approach that treats a group of people as a single entity instead of processing each person individually. In this paper, we address the important problem of estimating the number of people in a group and being able to track it reliably.

The ability to accurately estimate the number or density of people in a scene has useful applications in many areas. In traffic control, automatic pedestrian and crowd monitoring techniques can be used to increase safety and improve timing of traffic. For example, at traffic intersections, intelligent walk-signal systems could be designed based on the proposed method. The system could be automated, such that the signal decides when to change based on the number of pedestrians waiting to cross. This would be more efficient than most systems currently in use, in which people have to press a button in order to cross intersections.

Other important applications have to do with estimating the number of people walking through a crowded area. For example, knowing the size and density of a group outside a school or public event can help authorities identify unsafe situations and regulate traffic appropriately. This capability can also be useful in planning an environment. For example, a sidewalk or corridor could be designed based on typical pedestrian traffic patterns in the area. Additionally, retailers could use information about traffic in their stores to improve efficiency and design spaces appropriately.

Although there is a lot of work in the literature on tracking people in crowded scenes, there has been limited research on the specific problem of counting the number of people in a scene, and most of the work that has been done is restricted to counting individuals. Systems based on detection and tracking of individuals suffer in areas with high crowd density and occlusion. As such, these systems are of limited use in urban environments, which often contain small or large groups of people which are dense, and in which many people are severely occluded.

We propose a method that can be used to accurately estimate the number of people in a scene without constraining ourselves to detection of individuals. The approach here is group-based, where the count of people in a dense moving group is estimated as a whole. In this paper, both, a heuristic-based and a shape-based method are presented for estimating group populations. Both methods are implemented in real-time. We then track each group as a single entity using a tracker based on an extended Kalman filter.

The methods presented here are most effective for scenes in which people travel as individuals or in social groups. Examples of places where this might be applicable include shopping malls, college campuses, and public transportation areas such as train

stations and airports. This system is not intended for use in environments with immense crowds of people that cannot be segmented into groups, such as mob scenes, political rallies, or the like.

The report is organized as follows. In Chapter 2, we discuss some of the previous work in the field of crowd monitoring and counting. We provide a brief overview of the entire system in Chapter 3. Chapter 4 explains the projective geometry techniques used in the group population estimators. In Chapter 5, we introduce the tracking mechanism and its extensions. We present the two counting methods in Chapters 6 and 7. An alternative bounds-based formulation is presented in Chapter 8. Chapter 9 describes how the system handles groups merging and splitting. We then present and analyze the results obtained from experiments in Chapter 10, and some concluding remarks and future direction are given in Chapter 11.

CHAPTER 2

RELATED WORK

There have been various approaches to the problem of counting people in crowded environments. Early work involved locating people by looking for heads in the vertical histograms of the blobs, where the number of peaks was assumed to correspond to the number of heads. Davies et al. [1], in 1995, proposed a system which provides an estimate based on the number of foreground pixels and edge pixels. It uses Fourier transform techniques to identify motion of the crowd. It is restricted to motion only in the vertical direction in the image. Crowd density estimation is done using elementary techniques such as estimating the area of the image segments which correspond to moving crowds, or using the perimeter of the region occupied by the crowd. Although the system works well for scenes with few people, it is incapable of accurately determining density in case of occlusions and illumination changes as mentioned by the authors. Since occlusions occur frequently in cases with large groups, the system also fails when there is a large number of people.

Pfinder [2], which used a statistical model for color and shape to segment a person, tracked heads and hands, and identified gestures, mainly dealt with individuals. W4 [3] was another system for detecting and tracking individuals based on shape models. However, detection of individuals becomes less feasible in the case of crowds, where people are typically severely occluded. Hydra [4], an extension of W4, proposed a method based on silhouettes to discern people moving as groups. It used heads to count people in groups. However, heads may not be a very reliable cue when they occupy only a few pixels in the image, and hence require a reasonably close view of the scene.

Texture was used as a cue by Marana et al. [5] in order to estimate crowd density using a Kohonen neural network and Haralick's gray level dependence matrix. One drawback is that the method requires the background to consist of low frequency variations only. Since texture is estimated using frequency variations in the image, high frequency changes in the background adversely affect this method. This limits applicability to controlled, indoor scenes only. Also, since the method uses two neural networks, extensive training is required for good performance. The authors mention training using 151 images in the paper. They were able to provide an estimate of the density but were not able to estimate the exact count of people.

Zhao and Nevatia [6] proposed a Bayesian model-based segmentation algorithm using shape models which segmented each individual from a scene. It was able to count individuals in groups of people. This method was based on Markov Chain Monte Carlo sampling and was prohibitively slow for large crowds. They later proposed a method for tracking people [7], which used 3D ellipsoid models to track individuals, which, the authors claim, performs better than the original method.

The major drawback of most of the methods mentioned above is that they assume there is a distinct visual separation between individuals, so that the motion-segmented image contains enough visual information to separate individuals moving as a group. However, this is not always true in dense groups, when people are severely occluded and visually inseparable. See, for example, Figure 1.

In [8], Marana et. al describe a technique for estimating the density of crowds using Minkowski fractal dimension as a characterization of image texture. Kettnaker et. al [9] outline a people counting method using multiple cameras. They use visual appearance matching and mutual content constraints together in order to count people. Using multiple cameras has the obvious advantage of alleviating problems due to occlusions. However, registration needs a significant amount of work and prior calibration. In [10], an MRF-based approach is used first to do foreground extraction, and then a density estimate of the crowd is obtained using calibration information on the extracted foreground. This method suffers when there is occlusion as well, and the authors propose using the time-median statistic for better average results. Kong et al. [11] present a learning method for the estimation of crowd densities. Edge orientations and blob size histograms in the image are used as features for learning. Normalization of the features is done by taking into account viewing geometry in the scene which makes the method viewpoint invariant. Actual training is done using linear filtering and neural networks. Occlusion and overlap are conditions under which the method suffers.

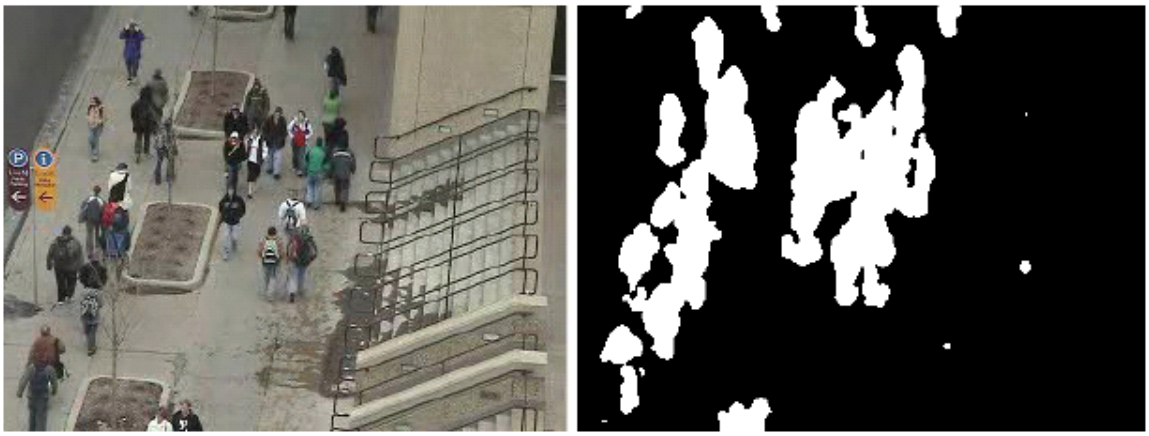


Fig. 1. A crowded scene and the corresponding motion-segmented blob image.

Alternatively, in [12], the authors use a clustering technique to cluster feature points tracked across frames in order to estimate the number of moving objects in a crowded scene. Because of the feature tracking mechanism, this method seems to handle occlusion and crowded scenes better. However, the method is still based on segmenting individuals rather than treating a group as a single entity. The technique can be used for counting any kind of moving objects and is not limited to humans, but it is assumed that the scene is homogenous (for example, it cannot contain both humans and vehicles at the same time).

In this work, we employ a different approach using geometric projections, and deal with the entire area occupied by a group as a whole, rather than trying to detect individuals separately. This reduces the effects of certain problems like occlusions and overlap which are faced by conventional techniques.

CHAPTER 3

SYSTEM OVERVIEW

In this paper we present two different methods for estimating the number of people in a group. Figure 2 shows an overall flowchart of the system. To start with, it is assumed that foreground regions have already been segmented. In our implementation this is accomplished in real-time using the adaptive mixture of Gaussians method proposed by Atev et al. [13]. The projected area on the ground in world coordinates is then estimated for each foreground blob using the method described in the next section. Each foreground region is labeled as an individual, a group, or neither (eg., a vehicle). These classifications are made based on the projected sizes and velocities of the regions, as will be described later in more detail. A tracker is then automatically initialized for each foreground blob that is identified as either an individual or a group.

There are two tracking modes used here. The first mode is used to track foreground regions labeled as individuals. This is done using extended Kalman filter (EKF) techniques as proposed by Masoud and Papanikolopoulos [14]. The second mode, which extends the EKF tracker to count the number of people in a group, is activated when a foreground blob is labeled as a group. In addition to tracking the group from one frame to another, the estimated count at each frame is maintained, as will be described later in more detail. For foreground regions identified as groups, the count of people in the group is estimated at each frame using either the heuristic-based or the shape-based method. In both methods, there is an initial training phase, in which a sample of individuals is analyzed in order to estimate the average projected area in the world occupied by one person. This information is one of the cues used to estimate the number of people in a group based on the area or shape of the group's projected blob, which relies on accurate camera calibration [19].

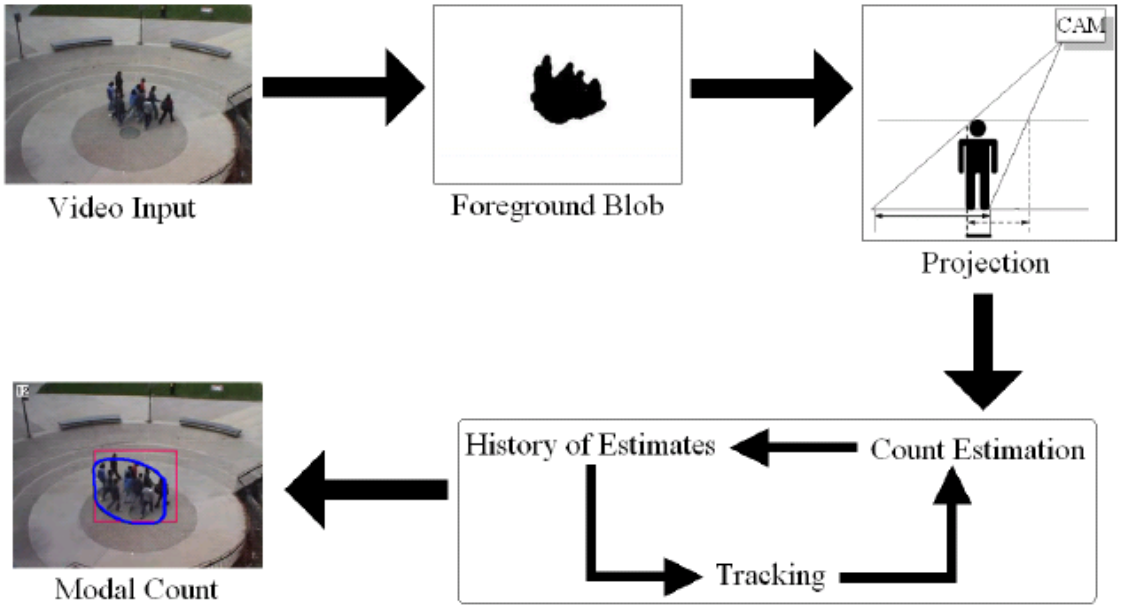


Fig. 2. A flowchart of the entire system.

These count estimation methods are robust to the motion, orientation, and spatial distribution of the group. Partial occlusions do not pose a problem as we do not try to

count the people in a group individually. Instead, they are treated as one whole group. This estimate is done every frame and a history of estimates for each blob is maintained throughout its lifetime, so that the most occurred value (modal estimate) is chosen at any given frame. This modal count estimate adds robustness to the system, since as groups move they often change shape, split, and merge temporarily due to the spatial constraints of the environment, all of which temporarily affect the instantaneous count estimates. Groups which merge or split for more than a fixed amount of time are reinitialized as new groups.

Both the heuristic-based and the shape-based method assume that there is, on average a certain distance maintained between people in the group. Note that this is not an absolute inter-person distance that people are forced to maintain, but rather a statistical average. Much research has been done in the Sociology community regarding interpersonal distances for people as a function of their interaction. See, for example [15], [16], and [17]. All of these papers maintain the claim that people prefer to maintain a certain, average distance between each other (which varies with gender, culture, etc.) when they interact.

Although there has not, to the best of our knowledge, been research which deals specifically with the case of people walking together in a group, it is a logical extension of the aforementioned work to assume that people in this case still prefer to maintain a certain separation between each other as a result of the same motivations. In the case of people walking in a group, however, interpersonal distance will exhibit transient fluctuations as groups temporarily deform to accommodate obstacles and other spatial constraints. Our method deals with this by using the modal count estimate as described above. Furthermore, in [18], Daamen et. al discuss the amount of longitudinal area that is required, on average, by a person walking at a certain speed. It is found that the area needed increases as the walking speed increases, due in part to the additional space required to take a longer step at a higher velocity. This is another factor affecting interpersonal distances in moving groups, and further emphasizes the fact that generalizations can be made about the average interpersonal distance in a group of walking people.

CHAPTER 4

FOREGROUND REGION PROJECTION AND AREA ESTIMATION

The area that a foreground region occupies on the ground plane in the world is one of the cues used to distinguish between individuals and groups. Additionally, for regions that are labeled as groups, the projected area is used to estimate the number of people in the group. This section describes the projective geometry used to project a foreground blob into world coordinates and approximate its area. It is assumed that we have a calibrated camera available.

All measurements of the area are done in world or scene coordinates. As such, each foreground blob is transformed into world coordinates through projection using the camera calibration information. It is assumed that people are moving on the ground plane in the real world. In this context, this amounts to projecting the blobs (corresponding to people) onto the ground plane as shown in the Figure 3. But we can see clearly from the figure that if we only used this projected area, then as objects move farther away from the camera, they will project to larger areas. One solution to this, as shown Figure 4, is to project the blob onto the ground plane as well as the head plane and take the intersected area in world coordinates. This eliminates the variation of area with the distance from the camera. This has the additional advantage that objects which are shorter than a human being (e.g., dogs, bicycles, and the like) will yield an intersected area of zero and therefore will not be counted as humans. The objective here is to have the head plane low enough so that all the individuals in the scene are detected. Having a head plane height too high will result in zero intersected area for shorter people. Very small head plane heights result in detecting shorter objects, which we want to eliminate using this projection technique. We found that a head plane height of 160 cm achieves this balance well. We therefore use this head plane height for the intersected area computation throughout the paper.

This helps make the method robust to false detections from other objects commonly found in urban environments. This also reduces the effects of shadows on the ground plane in cases where the shadow of a person or group appears shorter in the image than the height of the person or group. From here on, when we refer to area, we mean this intersected area in the world.

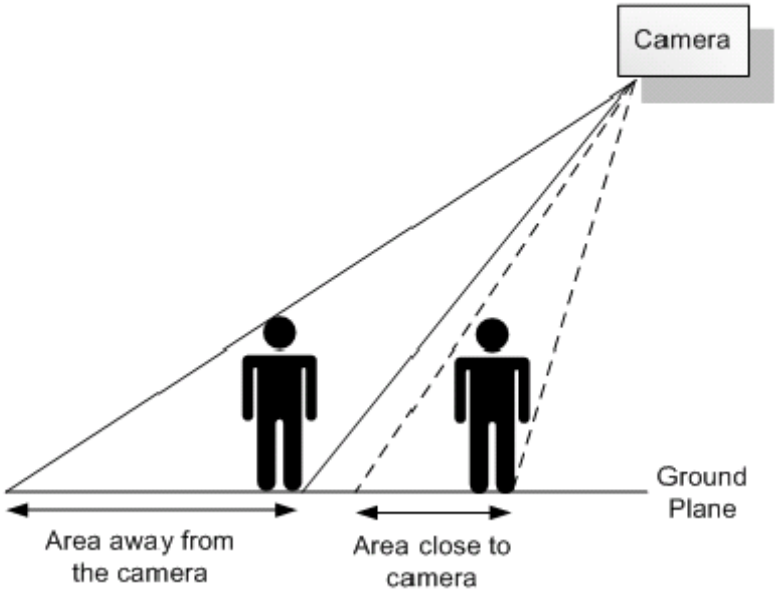


Fig. 3. Projection onto ground plane only. Objects farther from the camera have a bigger projection.

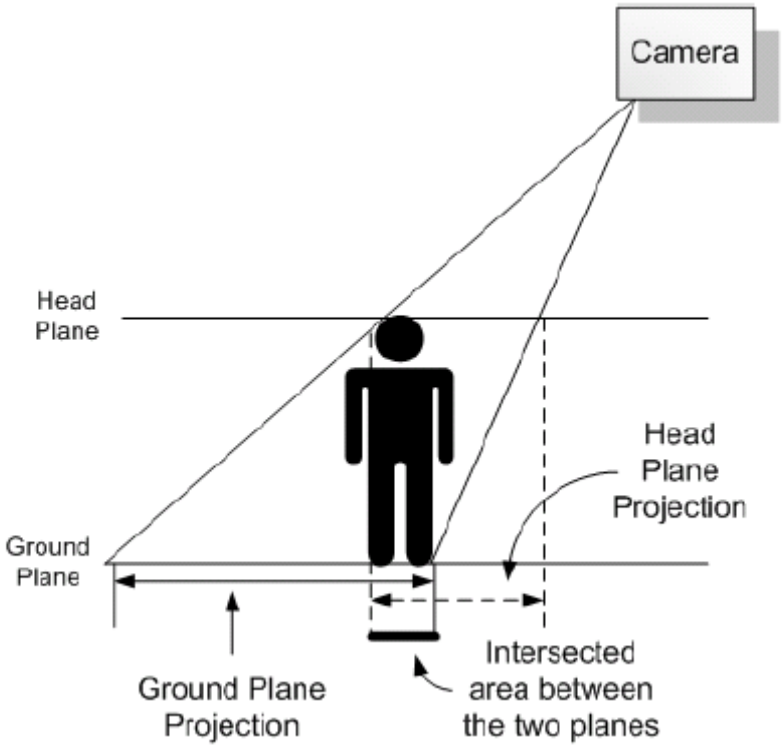


Fig. 4. Using intersected area between the ground plane projection and the head plane projection to eliminate the problem.

CHAPTER 5
TRACKING

The basic tracker used here is the extended Kalman filter (EKF) pedestrian tracker Proposed by Masoud et al in [14]. First of all, it is used to help classify foreground regions as either individuals, groups, or vehicles. For the case of groups of people, The basic tracker is extended to include group count estimate information, and this history of estimates is used to combat the effects of occlusion in the scene. This is described in detail in this section.

5.1 Foreground Region Classification

The system presented here aims to count the total number of people in the scene, including the sum of all individual and group count estimates. Before this can be done, each foreground region must first be classified as either an individual, a group, or a vehicle. This classification is done as follows:

- All blobs whose projected area (computed as described in Section 4) exceeds a threshold are classified either as a group or a large object (bus or car). This threshold is selected to be just under the area corresponding to two individuals in the real world.
- For these large blobs, we initialize the EKF tracker by observing their velocity for a small number of frames. If it stays above a velocity threshold, they are classified as vehicles and ignored. The remaining large blobs are classified as groups, and the counting algorithm is applied and a group tracker is initialized.
- For all blobs whose area is less than the area threshold, we first check if these are taller than they are wide, as individuals walking upright would be. For these blobs, we initialize an EKF tracker and track them for a minimum number of frames. If this can be done reliably, then the region is classified as an individual, and correspondingly is counted as one person in the total scene count. All other blobs are classified as noise and are discarded.

5.2 Group Tracker

Foreground blobs that were classified as individuals are tracked using the basic EKF mentioned above. For regions that were classified as groups, the basic EKF is extended to include information about the group count estimates at each frame. As the group is tracked, at each frame, the group count is estimated using one of the algorithms described in the following sections. Each of these instantaneous estimates is stored, along with an initial estimate age of zero. Every time a new instantaneous estimate is stored, the ages of all the previous estimates are updated. This history of count estimates is attached to the EKF tracker, and all of this together is referred to as a group tracker. It is not sufficient, however, just to use the instantaneous group count estimate at each frame. This instantaneous count may fluctuate rapidly due to temporary group deformations and occlusions. To make the overall measurement more robust, the history of estimates, which is part of the group tracker, can be used to make a smooth count estimate. As such, we use the mode (most frequently occurring value) of the history of estimates as a robust estimator of the group count. The number of occurrences of an estimate can be thought of

as its confidence score, and thus the mode of the history of estimates is the estimate in which the counting algorithm places the most confidence.

CHAPTER 6
HEURISTIC-BASED METHOD

In this method, the number of people in a group is estimated based only on the area they occupy on the ground plane. The area occupied by the group is estimated using the projection and intersection discussed in Section 4. This area is then divided by the average density of a group, which is learned during the training phase.

6.1 Training

For this method, we fix the height of the head plane at 160 cm above the ground, which, as discussed earlier, is an intentionally low estimate of the average adult height in order to ensure that the intersected area of objects of interest is nonzero. We then observe the average intersected area occupied by people moving in the scene. This is done for a large sample of people at different locations in order to reduce errors due to specific camera locations and other factors specific to a scene. We take the mean of these areas to obtain the average area occupied by a single human, K. This becomes the heuristic based on which groups are counted. As mentioned in Section 3, it is assumed that, on average, people maintain a specific distance between each other as they walk in groups. Since the method is based on training at different locations with different sets of people, the average distance between people is also accounted for in the heuristic K. Once the training phase is complete, we are ready to provide count estimates for groups in the scene.

6.2 Count Estimation

This counting procedure is performed only on blobs that have been labeled as groups using the procedure described in Section 5. Once we know that we are tracking a group, counting is done as follows:

(1) Given the intersected area of the group being tracked, estimate the count for this current frame to be:

Count = Area/K.

(2) If the blob is already being tracked, update the estimate of the group tracker; otherwise initialize a group tracker with this as the initial estimate of the count for the group.

(3) Compute the total number of people in the scene by summing together the count estimates for all groups in the scene, along with the number of individuals present. This is done in real-time on a frame-by-frame basis.

6.3 Bounds on Count Based on Training Statistics

Since the value of the heuristic K is critical to the counting procedure, it is important to perform some statistical analysis in order to provide confidence intervals for the count estimate. To do this, a sample of 20 individuals is selected from the video sequence from various points in the scene and uniformly spread out in time. This is to avoid any bias due to location in the scene or illumination change with time. The area occupied by each of the people is estimated using the projection and intersection technique. Then, the mean μ_K and standard deviation a_K are computed for these samples. For a large enough sample,

this mean would be the average area occupied by a human. Once we have these statistics, assuming an underlying Gaussian distribution, we can provide confidence intervals for our estimates. Given the area of the group, the mean count estimate would be given as:
$C = \text{Area}/\mu_K$.

Therefore, the 95% confidence interval for the count estimate is the range:
$(\text{Area}/(\mu_K - 2a_K) , \text{Area}/(\mu_K + 2a_K))$.

6.4 Discussion

Even though the heuristic-based method provides an extremely simple and efficient solution to the problem of counting people in groups, there are some cases in which it fails.

(1) The estimated count for a group depends on its spatial density. The heuristic method assumes a relatively high spatial density corresponding to the average density observed in the training phase, and therefore performs poorly in the case of groups that are more spread out. If there are significant gaps, this technique will typically overestimate the count since the number of people occupying the same area is less than the average.

(2) This method does not handle changes in configuration or the dynamics of the group explicitly. As such, if the configuration of a group changes, the projected area it occupies may change, even though the number of people remains the same.

(3) This method also uses a fixed head plane, which is at a constant height of 160 cm. Although this conservative estimate has been shown to produce relatively stable results, in some cases the height of a group might differ significantly from this, which would affect the accuracy of the count.

To deal with these issues, we propose a more flexible, probabilistic approach that provides count estimates for a group based on its shape. This is presented in detail in the next section.

CHAPTER 7

SHAPE-BASED METHOD

In this method, the shape of a group's intersected area (intersection of ground and head plane projections) is used in order to estimate the number of people present. Groups are modeled as elliptical cylinders, and probability priors are used to incorporate prior information. In some cases, multiple elliptical cylinders are used to model a single group. An initial solution for the elliptical cylinder is computed based on some simple shape statistics of the intersected area. Then, a cost function is minimized in order to find the model that best approximates the shape of the group.

7.1 Shape Model

As before, this method assumes that there is an average interpersonal distance maintained in groups, which is backed by the Sociology and Transportation literature ([15] - [18]). Here, elliptical cylinders are used to approximate the shape of a group (Figure 5). An elliptical cylinder is a general model that can be configured to accurately fit many group geometries. For example, a group that has a circular geometry can be accurately modeled by a standard cylinder with a circular base. In some cases, such as on narrow sidewalks, groups are elongated and become single-file lines. This case can be modeled by an elliptical cylinder with an elongated base, as demonstrated in Section 10. We use the elliptical cylinder because it is a flexible model which encompasses the diversity of group configurations that are typically encountered in natural scenes. Furthermore, it is a simple model that can be specified by relatively few parameters, and has been shown to achieve good results for the application of estimating group counts. When a group that appears as a single motion segmented blob in the image becomes two disjoint polygons after the projection and intersection routine, each of the disjoint polygons is fitted with a separate elliptical cylinder. An example of this is shown in Figure 7, and will be discussed later.

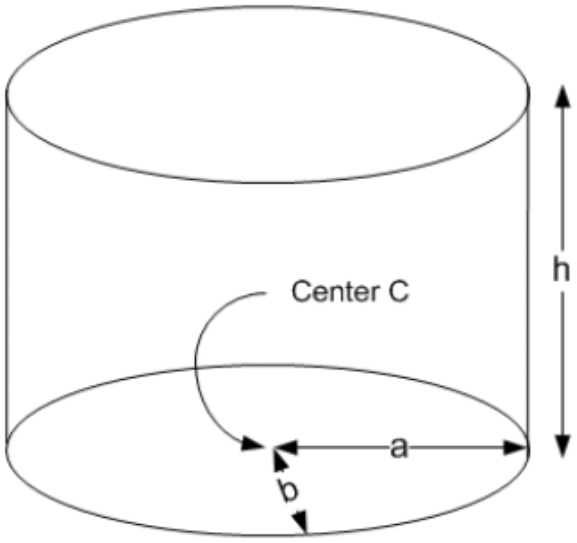


Fig. 5. An elliptical cylinder.

Mathematically, an elliptical cylinder S can be specified as S(A, h), where A = {x and y coordinates of the center, radii of major and minor axes, angle of orientation with respect to the x-axis} completely specifies the elliptical base, and h is the height of the elliptical cylinder corresponding to the human height.

7.2 Shape Priors

If we have prior information about a scene, and the way people move in it, we can make use of it in terms of certain prior probabilities. For example, if we know that there is a narrow pavement that is being monitored, it is likely that people will form long, narrow groups rather than more conventional elliptical groups. This prior information can then be incorporated into the shape model. The prior for the shape S is p(S) = p(r)p(h), where p(r) is the prior probability distribution of the ratio between the major and minor axes of the elliptical base, r = a/b, and p(h) is the distribution of the height of the elliptical cylinder, h, which corresponds to the average human height in the group.

We know that the distribution of the ratio of the axes controls the shape of the ellipse, which is the base of the elliptical cylinder. For most of the experiments presented here, we use a prior that is biased toward groups with circular geometries, which was found to work well in experiments with most scenes. Clearly, r = 1 for a cylinder with a circular base. This is modeled as a Gaussian distribution $N(\mu_r, s_r)$, where $\mu_r = 1$ and $s_r = 0.2$. The distribution of the height corresponds to the height of the elliptical cylinder and hence the average height of the group. We assume that the most likely average height is h = 160 cm which, as explained before, is a conservative estimate for objects of interest and was found to work well in experiments. Values which are far away from this are assumed to be less likely. The height is also modeled as a Gaussian $N(\mu_h, s_h)$, with μ_h = 160cm and $s_h = 20$. Note that even though the prior for group height is biased towards this mean, the group height is not fixed and is one of the parameters of the optimization. These models can be modified by the user according to the nature of the scene. If we do not have prior information at hand or do not want to make assumptions about the shape of groups, without loss of generality, we can weight all shapes with equal prior probabilities (i.e., uniform distributions).

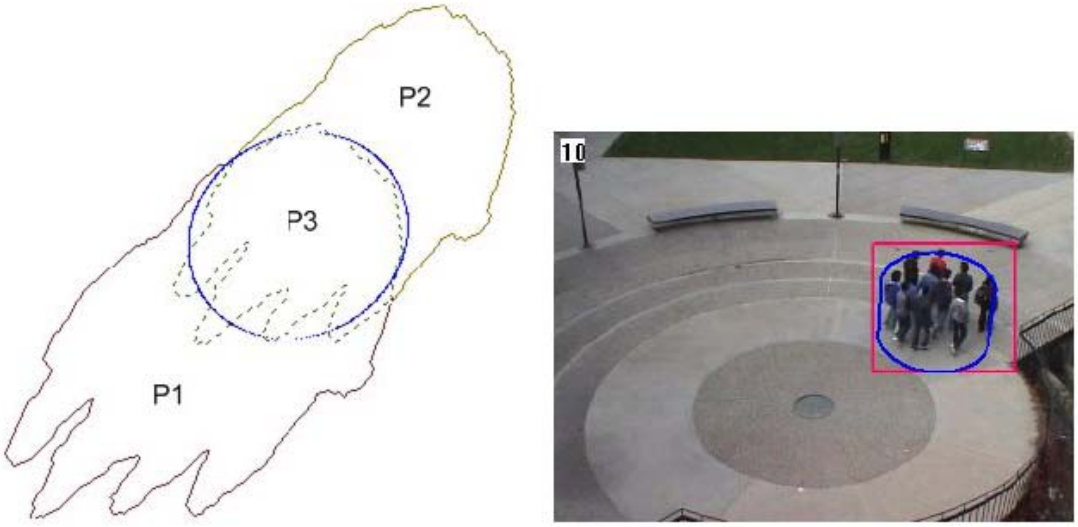
7.3 Finding the Optimal Shape

Now that we have a general model, we need to find the shape that best approximates the blob at hand. In other words, we want the elliptical cylinder in world coordinates whose projection into the image is the best fit for the group's blob in image space. The elliptical cylinder is completely specified by S(A, h). We want a cost function which measures the normalized overlap between the projection of this shape in image space, W(S), and the motion blob B. The cost function can, therefore, be defined as:

$$C(B,S) = 1 - \left(\frac{OverlapArea(B,W(S))}{maxArea(B,W(S))} p(S) \right), (1)$$

where W(S) is the world to image transform applied to the shape S, and p(S) is the shape prior. Now the problem can be framed as minimizing this cost function C as follows. The function C is a non-linear function of blob B and the elliptical cylinder S. Since B is given, this amounts to finding the values of the parameters A and h that minimize C.

There is no closed form solution to this problem. As such, it can be solved using an iterative non-linear minimizer such as the Conjugate Gradient method or LBFGS, which searches for a solution based on an initial estimate and the gradients of the cost function. If the initial solution is good (close to the optimal), the iteration converges quickly to the optimal value. The gradients for the cost function are computed using central differences. For all experiments reported here, LBFGS was used for optimization.



(a) (b)

Fig. 6. (a) The head and ground plane projections of the blob, and the base of the elliptical cylinder that approximates the group. (b)The projection of the elliptical cylinder into the image.

Initial Solution to the Minimization

In order for the optimization to to converge quickly, an initial solution which is relatively close to the minimum should be provided. To compute this, we start with the polygon P3 (in world coordinates) that is the result of the projection and intersection described in Section 4. Also, we assume initially that the average height of the group is 160 cm. Then, the elliptical cylinder that is used as an initial solution for the minimization is specified as follows:

(1) Find the first two eigenvectors of the contour of the polygon P3. Let these be v_1 and v_2. The major and minor axes of the ellipse are along these vectors, respectively. Also the ratio r = a/b is the ratio of the first two eigenvalues, l_1/l_2.

(2) The center of the ellipse (x, y) is computed from the moments of the contour of the polygon P3.

(3) The initial height of the cylinder is 160 cm, since that was the height used to obtain the projected polygon P3.

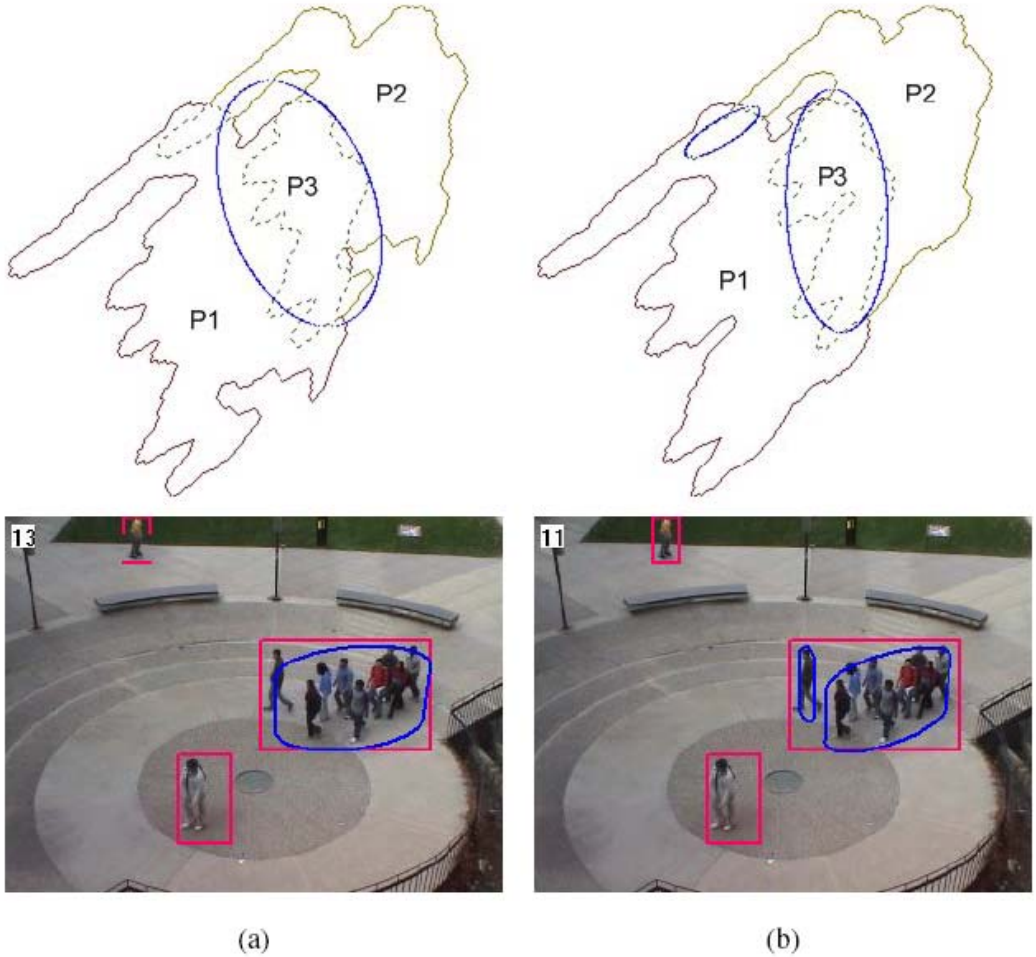
This elliptical cylinder is the initial solution for the minimization of the cost function C, which is used to find the optimal shape S with parameters A and h. An example of these polygons are shown in Figure 6(a). The polygon of the ground lane projection, P1, is towards the lower left with many peaks corresponding to heads. The blob projected on the head plane forms a smoother polygon, P2, towards the top right. The dotted polygon is the intersected polygon P3. The ellipse that best approximates this shape is shown in

blue, which is the base of the optimal elliptical cylinder. The elliptical cylinder is then projected back to the image using the reverse transform. The contour of this projection can be seen in blue in Figure 6(b). The area of this world ellipse (base of elliptical cylinder) will be used to estimate the count of the group.

Although this procedure typically provides a good approximation of the group, sometimes the projected and intersected polygon P3 is actually made up of multiple disjoint polygons. This can happen if the groups have unusual gaps and orientations. For example, consider the case depicted in Figure 7. The person at the front of the group is slightly detached from the rest of the group so in the projected world space, after intersection, it shows up as two disjoint polygons (Figure 7), even though they are part of the same blob in the image. If we used a single ellipse to approximate this, we would end up with an ellipse which tries to compensate for this but cannot come up with a good fit. In this case, we use multiple ellipses to approximate the two shapes in world coordinates. Hence, the same procedure described above for fitting a single ellipse is now applied separately for each of the disjoint polygons. The cost function to be minimized becomes:

$$C(B,S) = 1 - \left(\sum_i \frac{OverlapArea(B, W(S_i))}{maxArea(B, W(S_i))} p(S_i) \right), \quad (2)$$

where S_i is the set of elliptical cylinders. This yields a solution as shown in Fig. 7(b).



(a) (b)

Fig. 7. (a) Fitting one elliptical cylinder does not produce a good fit. (b) Fitting two elliptical cylinders produces a more accurate fit when projected back to the image.

Computing the Cost Function

This procedure is called multiple times in every iteration of the non-linear optimization. Given the approximation S, we need to compute the cost function C. The procedure for doing this for the case of one elliptical cylinder is explained below. When there are multiple cylinders, the procedure is extended as explained above.

(1) Project the ellipse from the ground plane (which is the base of the elliptical cylinder) to the image coordinates according to:

$C_I = H^T \cdot C_G \cdot H,$

Where C_i and C_g are the conic section matrix representations of the image and ground plane ellipses, respectively. H is the homography matrix which is a result of camera calibration. This will yield a conic in the image plane, which is the projection of the ground plane ellipse.

(2) Project the ellipse from the head plane (top of elliptical cylinder) to the image coordinates using a technique similar to the one described above. Here, the homography matrix is modified to account for projection from the head plane (which is parallel to the ground plane but a given height h above it). This will yield another conic in the image plane, which is the projection of the head plane ellipse.

(3) Now we have projected two conics from the ground and head planes, into the image plane. We compute the convex hull of these two projected conics in the image, which is the projection of the elliptical cylinder S. This is the polygon P4.

(4) The cost may now be computed as:

$$C = 1 - \frac{Area(P4 \cap B)}{maxArea(P4,B)} \cdot p(S),$$

where B is the polygon representing the blob.

7.4 Training for Spatial Density of a Group

From a reasonably large sample of individuals in the video sequence, we find the average area, Ah, of the polygon in world coordinates that is the intersection of the head and ground plane projections. For individuals, this intersected area has typically been found to be a circle. As mentioned earlier, there is a certain average gap assumed between individuals when walking in a group. This gap changes all the time, but we model it using an average value. We use a value of 30.48 cm (1 foot) as the gap between two individuals, which has been found to work well in experiments. This gap can be chosen to be a different value based on prior knowledge about the scene, since it is a tunable parameter. We therefore add 15.24 cm (0.5 feet) to the average projected radius of individuals found from the training, and recompute the average area A_h. This gives us a cylinder which approximates the average individual and the assumed space around him/her in a group.

7.5 Count Estimation

After the optimization, we have the elliptical cylinder S, which minimizes the cost function. Let the area of the ellipse that is the base of S be A. In the multi-ellipse case, it is the sum of the areas of each ellipse. Let us also assume that each human is a cylinder, whose base occupies a constant area A_h, as computed during training in the previous section.

Now the actual count estimation is given by:

Count $=A/A_h$.
This amounts to packing as many circles (humans) as possible into the ellipse (group). For each group, this count is computed using the optimal shape approximation and added to the group tracker's history of estimates.

CHAPTER 8

BOUNDS-BASED COUNT ESTIMATES

Given a motion-segmented image of a group of people (Figure 1), it is difficult to estimate the size of the group without making an assumption about its spatial density. In both the heuristic-based and shape-based methods discussed earlier, assumptions were made regarding this average interpersonal distance. Training data was used to determine the typical area occupied by an individual, and an average interpersonal spacing was assumed. While these methods perform well, their accuracy is bounded by the accuracy of the assumptions. One alternative is to provide upper and lower bounds on the number of people in a group. These can be computed as follows:

• Upper Bound: The upper bound can be obtained by assuming a tight packing or high spatial density. Given a blob, this is the maximum number of people that can occupy it. This is computed using either of the techniques described in Sections 6 and 7.

• Lower Bound: The lower bound is the minimum number of people that can occupy a given blob. This corresponds to a configuration of people which produces the maximum non-overlapping area in the world with the projected blob. This can be easily computed by back-projecting a single circular shape (corresponding to one person) centered at the ellipse in the world onto the image and using this area as the area for one person in that blob. The minimum count is then the ratio between the blob area and the area corresponding to one person.

CHAPTER 9

GROUP MERGING AND SPLITTING

When two groups of people come in close proximity of each other, they become visually indistinguishable. Often times, however, these group merges are temporary, and the two groups will split within a short time. This occurs when groups walk past each other in opposite directions, or when two groups traveling in the same direction are forced to compress temporarily due to spatial constraints or obstacles. Since group merging and splitting are commonplace, they are handled explicitly by our system, as explained below. Figure 8 shows an example of two groups walking past each other in close spatial proximity. In the motion segmented image, it is impossible to visually separate the two groups.

9.1 Group Merging

The main idea here is that when two entities merge, it is usually temporary, as explained above. As such, they are tracked as two separate objects, corresponding to the original objects, using only the predicted values of the Kalman filter. Therefore, when a merge occurs involving a group, the history of count estimates for the group is maintained without updating it. The modal estimate before the merge is used as the count for the group. The count for the entire merged group then becomes the sum of counts of all groups and individuals that were part of the merge. As soon as the merged entities split, the count estimation proceeds for each one as explained in the previous sections. If, however, the group remains merged for longer than a time threshold, they come to be considered as one group. In this case, a new group tracker is initialized on this merged group, and the count estimation process begins.

Fig. 8. Two groups of people walking past each other in opposite directions, and their corresponding motion-segmented image.

9.2 Group Splitting

As before, the main idea here is that when a group splits, it is often only temporary and due to obstacles, spatial constraints, or occlusions. As such, it does not make sense to discard the previous history of estimates and initialize new group trackers on each of the

split parts. Furthermore, the number of people in the scene has not changed. We therefore maintain the previous count estimate of the group before the split for a limited time. This also helps the system deal with static occlusions which are part of the background, such as signposts or signal poles. If the group remains split for longer than a time threshold, the split groups are then treated as separate entities and new trackers are initialized.

CHAPTER 10

ANALYSIS AND RESULTS

In this section we analyze the performance of the methods described above. In total, experiments were performed on three different scenes with 8 different positions of the camera. The camera height was varied from 8.23 meters to 27.43 meters, with various camera tilt angles. In addition to individuals, the scenes analyzed contain groups of people ranging in size from 2 to 11. Note that 11 is not the total number of people in the frame, but the maximum size of one group, and the total number of pedestrians is often larger. The videos used were shot on campus with some of the scenes being more crowded than others. All of the scenes were natural traffic. Both the heuristic- and shape-based approaches described above have been tested on these video sequences and the results are fairly consistent across them. All tests were implemented on a Pentium IV 3.0Ghz PC. Both of the proposed methods run in real-time for all the tested videos. The scenes are calibrated using Masoud's camera calibration technique [19]. A calibrated camera is essential to the operation of the count estimator.

10.1 Examples and Discussion

Here we present several representative results in order to demonstrate the performance of the approach in different situations. We discuss the strengths of the methods, and also point out places in which there were difficulties. The following results show frames with different densities of people in the scene, differing camera tilt angles, and different locations and illumination conditions. Figures 9 through 17 show natural outdoor settings with no user control on the scenes. The blue contour that can be seen in every frame is the projection onto the image plane of the elliptical cylinder in world coordinates that best approximates the blob representing the group. Red bounding boxes are shown for all motion segmented blobs, independent of their classification. These are sometimes affected by shadows. For each scene, the estimated count is displayed in the upper left corner of the frame. In some of the examples, a large 'G' or 'I' is superimposed over a blob to show that it has been classified as either a group or an individual, respectively. In other figures, these labels are not shown, since they severely occlude the image. Objects that are not classified as either individuals or groups do not have labels. In addition to the examples shown here, the reader is encouraged to refer to the snapshots presented in Section 7 while describing the algorithm.

Figure 9 shows two separate groups of people walking in opposite directions in the camera view. An advantage here is that the scene is relatively sparsely populated other than the groups themselves. The groups of people are dense, which causes significant occlusion within the group. Also, note the dark colored clothing of many of the people, which would have made it extremely challenging for any system which relied on segmenting individuals within the group. Our method deals with the entire group as a single entity, which alleviates the problem caused by color similarity and inability to visually separate people. The count is estimated using the shape-based method, and the result shown at the top left of the image is close to the actual number of people in the scene.

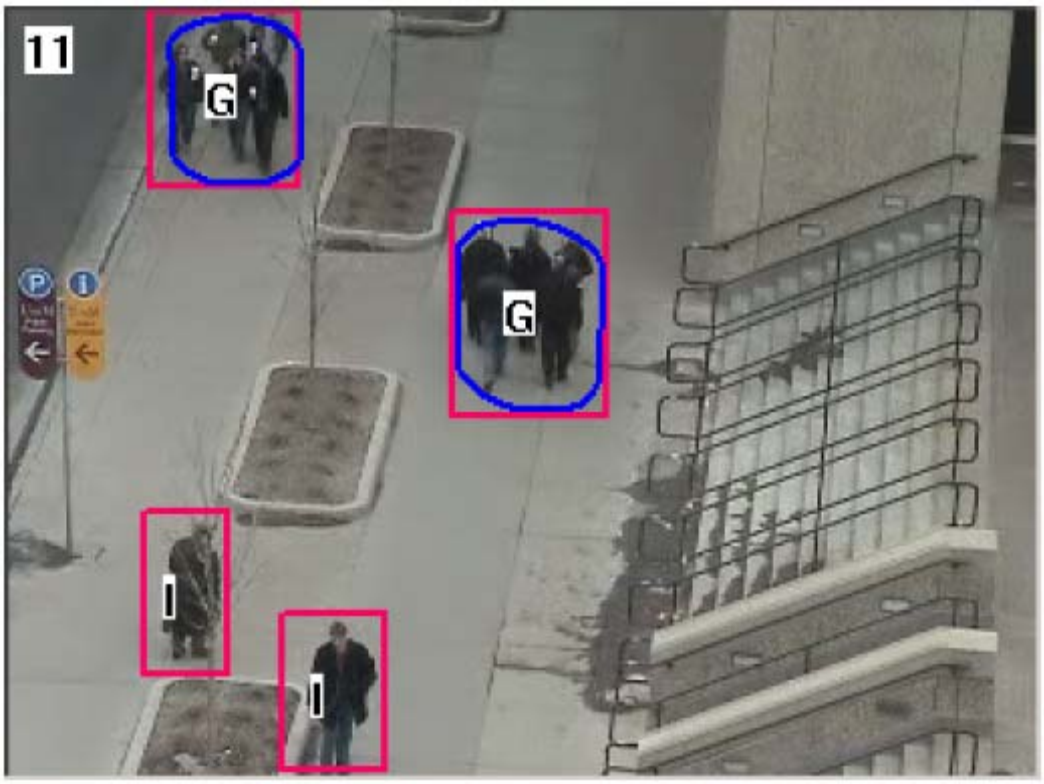


Fig. 9. Two separate groups walking in opposite directions. Note the high density of people within each group and the corresponding occlusions.

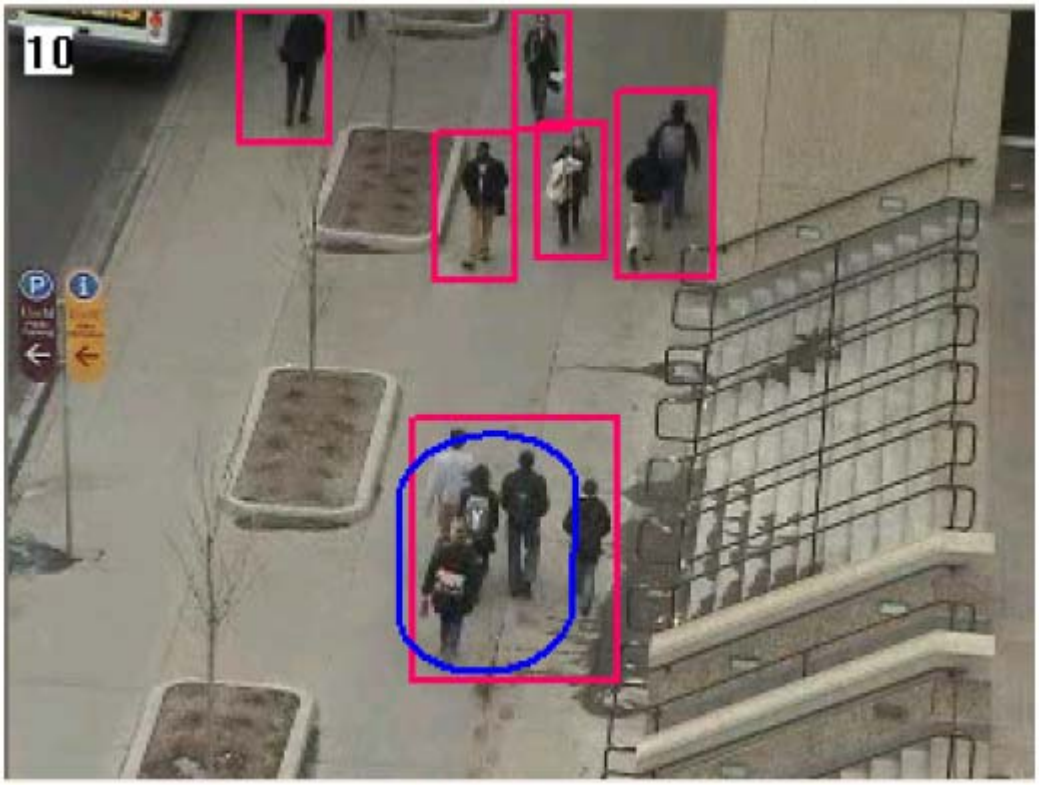


Fig. 10. A sparse group with differing gaps between individuals.

Figure 10 shows a different scenario with a group that is not very dense. The challenging part is that the gaps between individuals are not uniform. Even as a human, it is hard to establish whether this is actually a group of people walking together, or whether they are walking independently albeit in close proximity. Again, the count output gives a good estimate of the actual count.

In contrast to the above figures, Figure 11 distinctly shows three individuals walking independently who are misclassified as belonging to one group. The misclassification is due to spatial proximity of the three people. This might seem to be a situation which would throw off the estimate significantly. However, even in the case where separate entities are misclassified as belonging to a group, the count estimate is based on area of the group as a whole, which is not much different from the sum of the areas occupied by each individual separately. Thus, the count estimate, as can be seen, still remains very good.

In Figure 12 we see three people walking in a single-file line. It is hard to see that these are in fact three individuals from this frame itself, but this can be determined by viewing the entire video. As we have mentioned earlier, there are locations where such alignment might be forced due to physical constraints or obstacles in the region. In contrast with our method based on projected area of the entire group, a conventional identification system which tries to separate independent moving entities would suffer due to the significant overlap of individuals in this case.

In Figure 13, it can be seen that the camera is looking at the scene from a low height, and with a smaller tilt angle. This could potentially cause the count estimate to be wrong, owing to the large projections of blobs on the ground and head planes.

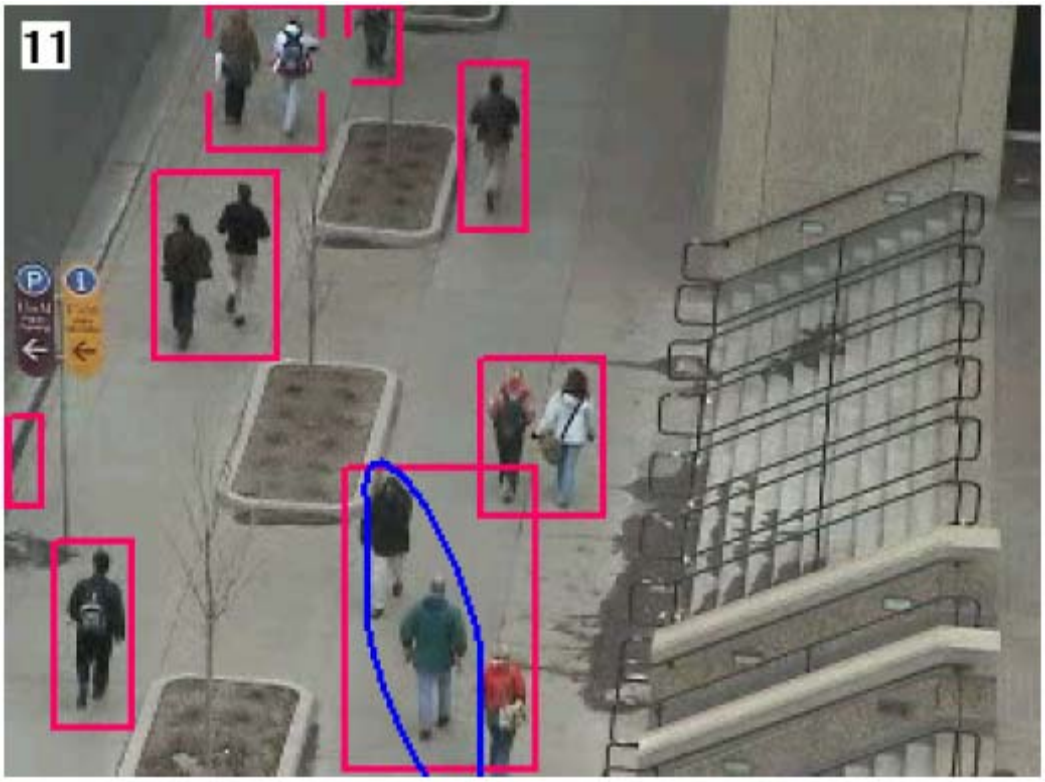


Fig. 11. Individuals walking separately classified as one group.

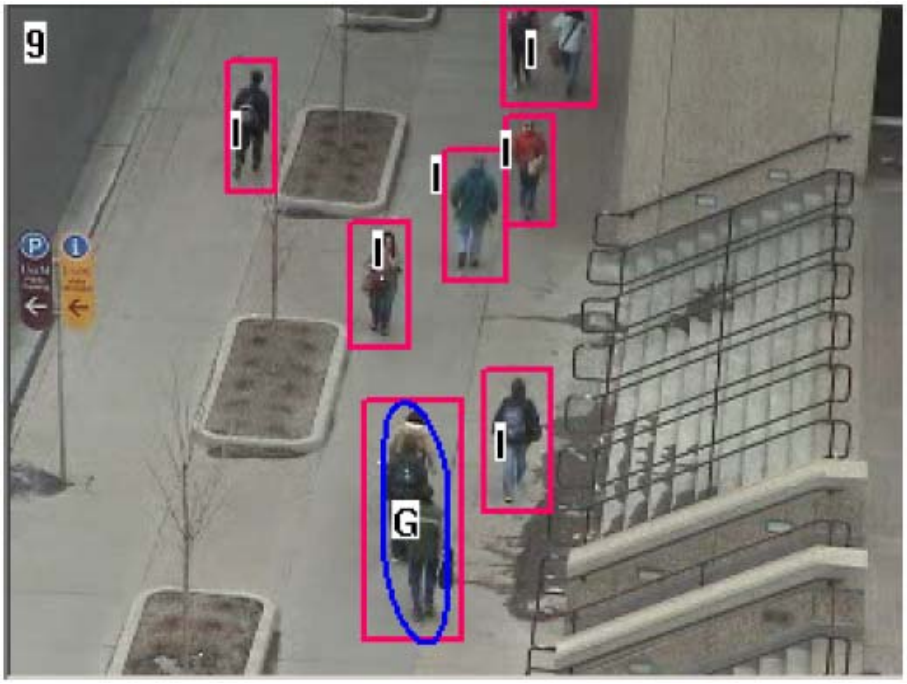


Fig. 12. A group in a single-file line with significant occlusion.

These extended projections could also cause large errors in the intersected area calculation. This implies that the proposed method would break down at extremely low tilt angles. The count estimate in the figure shows the favorable performance of the system for this small tilt angle of the camera. Note that the people in the lower left of the image are not counted since they are only partially visible, and neither are the two people in the center of the image, who are relatively stationary, and have been mostly modeled into the background.

Fig. 13. A frame with the camera placed low and a small tilt angle.

Figure 14 shows a sparsely populated group with large gaps between people. As a result, we see that the count is slightly overestimated, owing at least to some extent, to the large gaps in the group. This is a potential pitfall with views which are too close with small tilt angles. Again, note that the detection on the bottom left of the image is not counted as an individual. Here, the two people in the center were counted as an individual, since they were moving slightly but occluding each other.

Fig. 14. A sparsely populated group with a small tilt angle. Note that the system slightly overestimates the count owing to big gaps in the group.

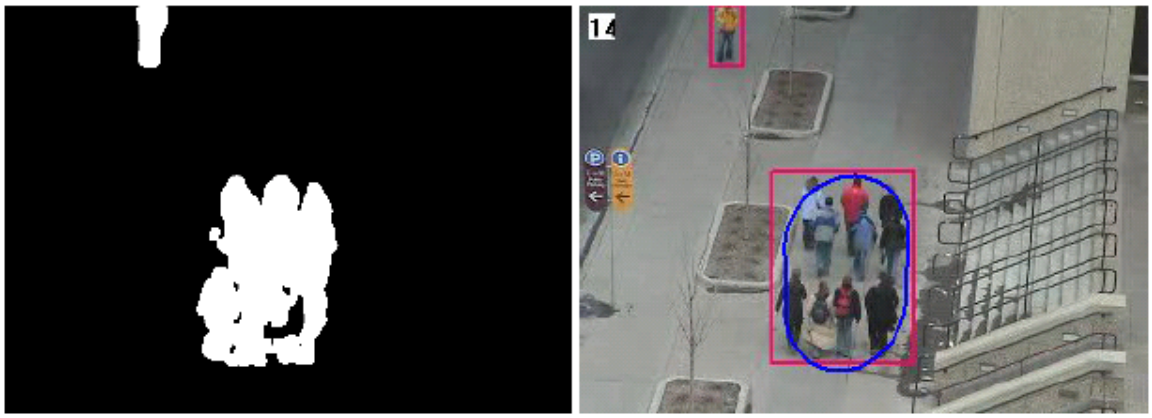


Fig. 15. A group of people with a significant gap.

Figure 15 shows another example of a group which contains a significant gap. Also shown is the corresponding motion segmented image. Note that due to the geometry of the scene, the gap between the two lines of people in the group is not visible in the motion segmentation. If the same group were walking horizontally across the image, the

gap would perhaps become visible. As such, the count is overestimated for this group (estimate 13 instead of the actual 10), since the blob would appear more or less the same even if there were 3 additional people standing in the gap. Nonetheless, the estimate is still reasonably close to the actual count.

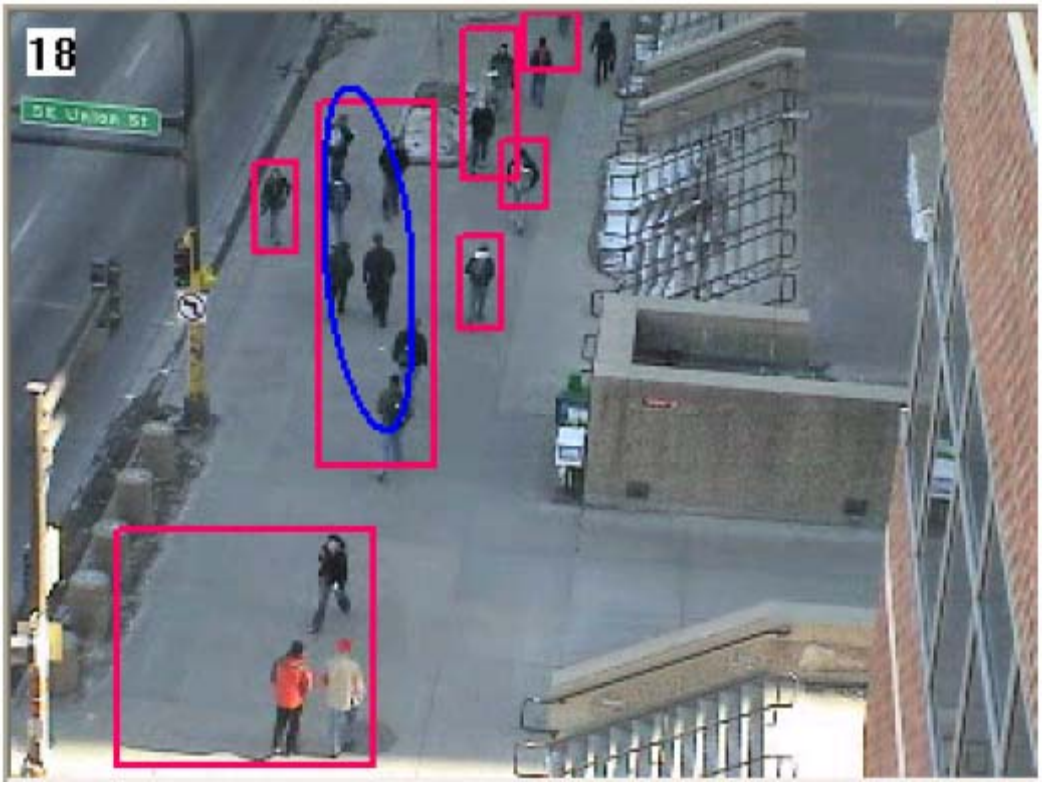


Fig. 16. A frame with a naturally denser scene, and a larger tilt angle of the camera.

Figure 16 shows a frame taken from a camera position farther away from the scene, and with a larger tilt angle than before. Ideally, for a conventional counting system, a tilt angle of 90 degrees (an overhead camera looking straight down on the scene) would be perfectly suited for the objective of counting moving objects independently. It is important to note that this is not true for our system, which has an added advantage of separating objects which are close to human height from other moving objects. This is done using projections on the ground plane and the head plane, and computing the intersected area. In case of an overhead camera, these projection areas would be the same for objects of different heights, making height distinction impossible. We thus need a tilt angle lower than 90 degrees to achieve this differentiation. The count estimate in Figure 16 is a good one, even with this large tilt angle.

Figure 17 shows a scene with significantly different intensity levels in different parts of the image. Shadows are predominant in the high-intensity region of the view, and one large shadow can be clearly seen. Due to the projection and intersection methods, the effect of shadows is reduced as they lie on the ground plane only. The blue contour of the group clearly shows that shadows are not included in the area computation. The resulting count estimate obtained from the system is a reasonable one.

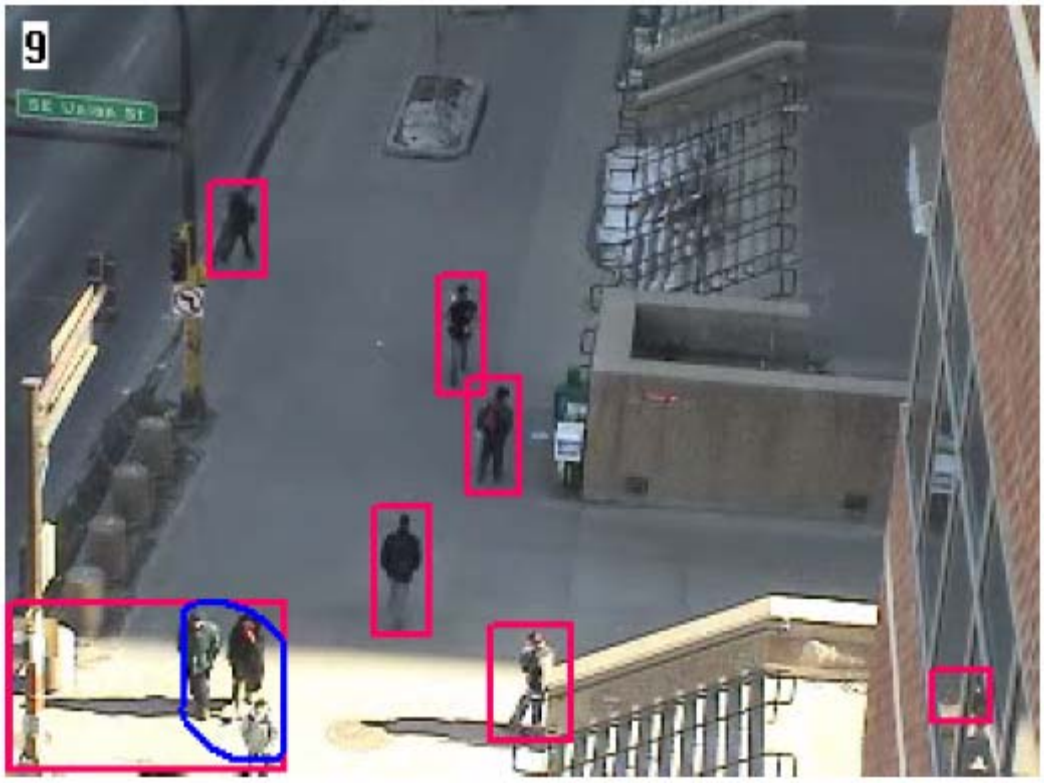


Fig. 17. A frame showing drastically different intensity levels in different parts.

10.2 Accuracy of Modal Estimate

As discussed earlier, a count history is maintained for each group of people for the duration of the time that it remains together. At each instant, the estimated count is then the mode of all the estimates in the count history. Here we quantitatively examine the accuracy of the proposed methods using this modal estimate. Table 1 shows the average modal count estimates for groups of different sizes. Results are shown for both the heuristic- and shape-based methods. The count estimates for a given group size are averaged over all occurrences of groups of that size, where the mode is taken over the history from the entire lifetime of the group. It can be seen that the shape-based method consistently outperforms the heuristic approach, especially for larger groups. During the testing, there were a few cases where groups of 2 or 3 people were miscounted as a result of people who were not moving together being erroneously grouped together as one, since they appeared to be one blob. The ellipse fitting (shape-based) method is more prone to this error as is reflected in the table.

Both methods suffer from a common limitation. During evaluation, it was noticed that two groups of sizes 4 and 7, respectively, were overestimated to be almost twice the ground truth. The problem here was that these two groups were far away from the camera near the horizon. This is because the per-pixel error increases with distance from the camera, since the distance between two neighboring pixels is greater in the world at this distance. We cannot get an accurate area measure in this region. This problem is suppressed by using a region-of-interest approach, where we do not consider any groups in regions beyond a certain distance from the camera.

Another possible approach for handling this problem is to incorporate the distance of the group from the camera into the probability priors while tracking. Then estimates further away from the camera are weighted so that they have a lesser influence on the count than the estimates close to the camera.

Table 1

Average estimated counts of both methods based on the modal estimate for groups of various sizes.

Group Size	No. of Groups	Heuristic Count	Shape Count
2	46	2.46	2.60
3	29	2.80	3.82
4	18	4.45	4.23
5	19	5.34	5.36
6	11	6.55	6.49
7	6	8.20	7.81
8	4	9.04	8.85
9	5	9.7	9.48
10	4	8.90	10.43
11	3	12.5	11.63

Table 2

Average per-frame instantaneous errors of larger groups over their lifetime.

Group Size	No. of Frames	Heuristic- Error Per-Frame	Shape- Error Per-Frame
8	332	1.44	1.17
9	530	1.51	1.30
8	372	1.04	0.85
11	354	1.81	0.72
9	383	0.70	0.83
10	156	1.53	1.24
10	224	1.86	1.03

10.3 Error in Per-Frame Instantaneous Estimates

From the instantaneous per-frame count given by each method for each group, the error in each frame was computed based on the ground truth. The averages of all these errors over the lifetime for larger groups in 4 different sequences are shown in Table 2. Note that these are the actual instantaneous group count estimates at each frame, before any mode was taken. It is encouraging to see that neither method has an average per-frame error of more than 2, even for groups with up to 10 people. This validates the use of area as the basis for estimating the counts of groups. The graph in Figure 18 shows the frame-by-frame instantaneous estimates for a group of size 11 over the lifetime of the group. Both methods show fairly accurate instantaneous estimates, except for a sharp peak around frame number 50, caused by a temporary spreading of the group. Once again, the shape-based estimator is more accurate on the average than the heuristic-based one.

10.4 Sensitivity to Calibration

As mentioned earlier, both count estimation methods require a calibrated camera. It is therefore important to know how sensitive the methods are to the accuracy of the calibration. Tests were conducted in order to measure this for both the heuristic- and shape-based methods for a group of 11 people. Results can be seen in Figure 19. The per-frame instantaneous count estimates were recorded for 2 sets of incorrect camera calibrations, as well as with the correct calibration. In the first case, the camera was calibrated with distances in the world recorded as 20 cm longer than their actual values. In the other case, the camera was calibrated with distances in the world recorder as 20 cm shorter than the actual values.

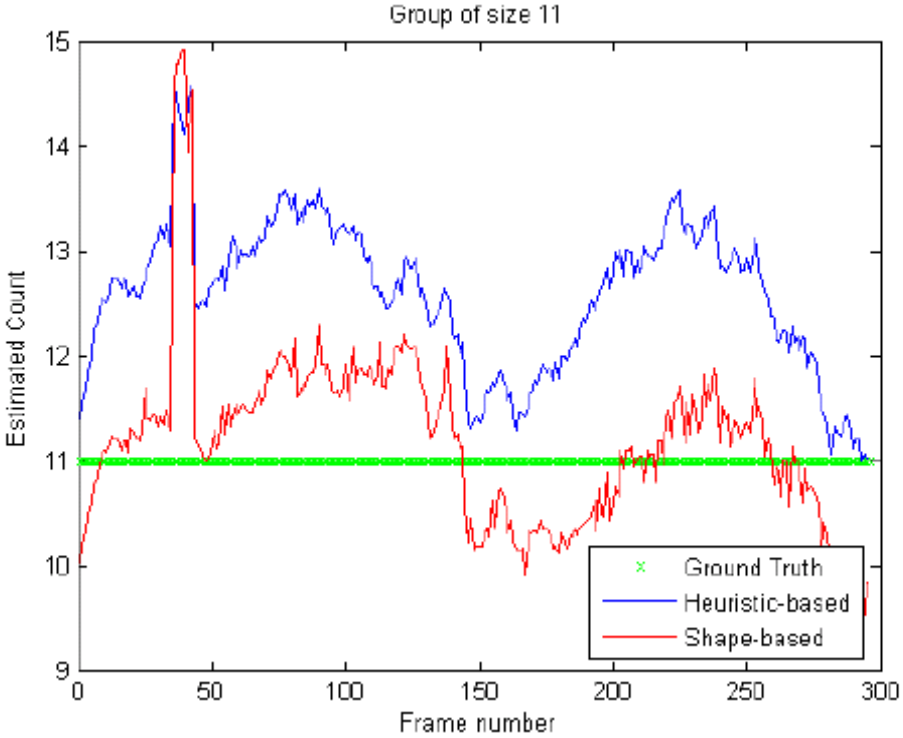


Fig. 18. Plot of actual per-frame instantaneous counts over the lifetime (284 frames) of a group of 11 people.

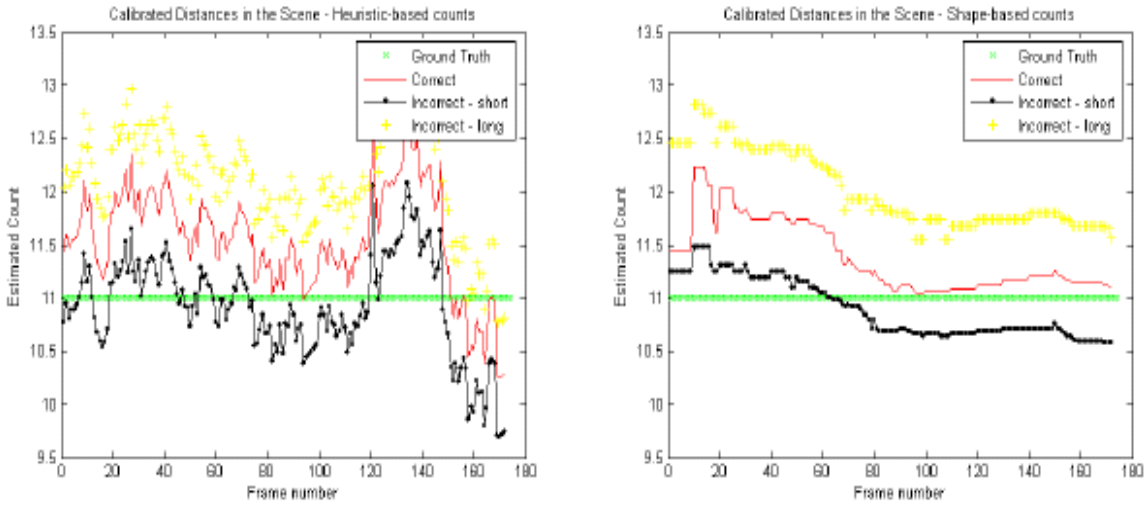


Fig. 19. Instantaneous per-frame counts for the heuristic- and shape-based methods with both correct and incorrect distance measurements used for calibration.

As can be seen, the shape-based method is less sensitive to incorrect calibration, but per-frame count estimates from both methods are within 20% error of the ground truth, which shows that the counting methods are not very sensitive to the accuracy of calibration. Recall, also, that these are the instantaneous per-frame estimates, before any mode is taken. This shows that even if accurate scene measurements are not available for calibration, we can still obtain fairly accurate results of group count from measurements in the vicinity.

10.5 Differentiating Groups, Individuals, and Vehicles

The threshold on the area to differentiate between groups and individuals for all the sequences was slightly below twice the average area occupied by an individual in that particular sequence. Results from 3 sequences show that 460 out of 468 individuals were correctly counted as individuals while the remaining 8 were classified as groups of 2. This test set of individuals included 29 people on skateboards and bikes. Also, 4 groups of two people walking together were classified as individuals owing to their closeness and orientation with respect to the camera. This shows that the adaptive threshold used in the experiments works well for distinguishing individuals from groups.

According to our framework, large blobs could correspond to either groups or vehicles. Large blobs that move faster than a velocity threshold are classified as vehicles, while slower moving blobs as groups. Figure 20 shows an image of a scene where a vehicle in the upper left corner is appropriately ignored, while the group on the lower left side of the image is accurately classified as such. Notice that the total count estimate of 6 is printed in the upper left corner, and is close to the actual count (7).

Fig. 20. A frame showing a group accurately labeled as such, and a vehicle that is not counted as a group.

10.6 Bounds-Based Count Estimates

The graph in Figure 21 shows the per-frame instantaneous count estimates produced using the bounds-based counting method. As can be seen, these produce more conservative estimates of the group count. The upper bound is very close to the ground truth for a short period, which is when the group has bunched up together temporarily and attains a density close to the assumed value. The spike before frame 50 appears as a result of a temporary spreading within the group as it turns in the scene. Alpha trimming (equal number of upper and lower members of the dataset are removed) can be used to calculate the trimmed mean of the upper and lower bounds to reduce errors caused due to this.

10.7 Note on Computational Cost

Finally, we examine the running time of the proposed methods. Results can be seen in Table 3. Processing speed is given in frames per second. Here we have evaluated the processing speed for scenes with different numbers of groups present simultaneously. The numbers in the first column of the table that are not in parentheses are the number of groups simultaneously present in the scene. The numbers in parentheses reflect the total number of people in the scene. These experiments were performed for the heuristic-based method, and the shape-based method both with and without optimization. When optimization was not used, the initial solution for the elliptical cylinder was used for the final estimate.

Fig. 21. Plot of upper and lower bounds on counts over the lifetime (284 frames) of a group of 11 people.

Table 3
Frame rate for each of the methods over some crowded scenes.

No. of Groups (Total No. of People)	Heuristic-Based (fps)	Shaped-Based w/o Optim. (fps)	Shape-Based w/ Optim. (fps)
1 (10)	30	30	25
2 (15)	30	27	20
5 (33)	25	21	11

Otherwise, when optimization was used, the method chosen was LBFGS. Uniform priors were used for all shapes in computing the cost function in these experiments. In most cases, the optimization of the cost function in Equation (1) converges within 5 iterations. As can be seen in Table 3, for the shape-based method, the frame rate does drop as the number of groups in the scene increases. However, the processing speed is independent of the sizes of these groups. It does not matter if there are 2 groups of 3 people each or 2 groups of 20 people each – the system's run-time would remain the same. The heuristic-based method, on the other hand, is less affected by the number of groups in the scene, since it is less computationally expensive.

CHAPTER 11

CONCLUSIONS AND FUTURE WORK

We have presented a system that can estimate the number of people in a group in urban environments. We use novel projection methods to compute the area that represents each blob in world coordinates. This method reduces problems due to moving objects which are not similar to human height. It also takes care of the differences between the projected area of people at different distances from the camera. Another novel aspect as compared to conventional methods is the treatment of groups as single entities rather than treating individuals separately. This approach makes the method relatively invariant to occlusion between people in a group. To make the technique robust and stable, we have added extensions to an existing Kalman filter tracker using the concept of history of estimates.

Two different methods, one based on a heuristic learned during training and another based on shape models, have been proposed for estimating group size and tested on various scenes. Motion trajectories of these crowds could also be collected for further data analysis. The system is capable of counting and tracking people without being significantly affected by occlusions, group merges, or splits. The system is view-point invariant, as we use accurate camera calibration methods. Results have been shown to be promising.

Systems that count the number of people in crowded scenes are a relatively recent development in the computer vision literature. As such, there are still several issues that need to be addressed. The system proposed here should be extended in order to deal with stationary people in a scene. This would require a different type of background segmentation which is not based on motion. Also, problems near the horizon are currently avoided using a region-of-interest mask. Methods to improve the accuracy in these regions should be investigated, such as use of weighted estimates.

Currently, the system sometimes classifies slow-moving vehicles as groups of people in certain cases where the motion segmented blob is of a particular shape. Methods for more reliable classification between crowds and vehicles in such circumstances should be explored. Also, the problem of shadows needs to be addressed more explicitly. Due to the projection and intersection technique, which is used for estimating the shape occupied by a blob on the ground plane (Section 4), the effect of small shadows is often reduced in the current method. However, when the shadows appear larger than the people themselves, the count estimation is affected.

In this system, we use the blobs obtained after foreground segmentation to estimate the number of people present. However, given only a blob, one cannot with certainty supply an accurate estimate of the count of people. The same blob could correspond to a different number of people based on their spatial orientation and density, as demonstrated in some of the examples. As such, methods need to be explored in order to adaptively estimate the density of a group based on its shape and other cues from the video sequence.

[15] S. Heshka and Y. Nelson, “Interpersonal Speaking Distance as a Function of Age, Sex, and Relationship”, *Sociometry*, vol. 35, no. 4, pp. 491-498, December 1972.

[16] J. C. Baxter, “Interpersonal Spacing in Natural Settings”, *Sociometry*, vol. 33, no. 4, pp. 444-456, December 1970.

[17] D.E. Thompson et al, “Interpersonal Distance Preferences”, *Journal of Nonverbal Behavior*, vol. 4, no. 2, pp. 113-118, 1979.

[18] W. Daamen and S. P. Hoogendoorn, “Experimental Research of Pedestrian Walking Bahavior”, *Transportation Board Annual Meeting 2003*, pp. 1-16, National Academy Press, Washington, DC.

[19] O. Masoud and N. P. Papanikolopoulos, “Using geometric primitives to calibrate traffic scenes”, In *Proc. IEEE International Conference on Intelligent Robots and Systems*, pp. 1878-1883, September 2004, Sendai, Japan.