Pedestrian Detection in Uncontrolled Environments using Stereo and Biometric Information

Philip Kelly, Noel E. O'Connor and Alan F. Smeaton
Centre for Digital Video Processing & Adaptive Information Cluster,
Dublin City University, Ireland
kellyp@eeng.dcu.ie

ABSTRACT

A method for pedestrian detection from challenging real world outdoor scenes is presented in this paper. This technique is able to extract multiple pedestrians, of varying orientations and appearances, from a scene even when faced with large and multiple occlusions. The technique is also robust to changing background lighting conditions and effects, such as shadows. The technique applies an enhanced method from which reliable disparity information can be obtained even from untextured homogeneous areas within a scene. This is used in conjunction with ground plane estimation and biometric information, to obtain reliable pedestrian regions. These regions are robust to erroneous areas of disparity data and also to severe pedestrian occlusion, which often occurs in unconstrained scenarios.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*depth cues, stereo, object recognition*

General Terms

Design, Security, Algorithms

Keywords

Pedestrian Detection, Homography, Stereo, Disparity, Biometric Data

1. INTRODUCTION

Pedestrian detection and tracking is an important task for many computer vision based applications, such as security systems, pedestrian density and flow pattern estimation, driving assistants and automated crossroads. However there are many inherent difficulties with extracting individual pedestrians from an unconstrained scene. A few of the complicating factors include; the large variability in a pedestrian's local and global appearance [9] and orientation; occlusion of a pedestrian by one or several other pedestrians, or objects, especially if the pedestrian is located within a crowd; a pedestrian's clothing may also be the same colour as the background region, resulting in difficult segmentation. In addition, pedestrian detection in real world scenarios must address rapidly changing lighting conditions due to the sun becoming occluded and then emerging from behind clouds; the possibility of having moving backgrounds, reflections on windows or from rain puddles on the ground, and shadows cast by pedestrians and other foreground objects.

Various techniques for segmenting individual pedestrians have been investigated using traditional 2D computer vision techniques. Some require a certain camera orientation which is difficult to achieve in outdoor scenarios [13]. Other approaches use rhythmic features, such as the periodicity of human walk, or motion patterns unique to human beings, such as learned gait [10]. These techniques assume a pedestrian is non-static, unoccluded and walking in a particular direction relative to the camera. Shape based approaches, such as [5], try to solve the harder problem of recognizing pedestrians in single images, hence addressing both moving and stationary pedestrians. This is achieved by searching images for pedestrian shapes which are matched to a predefined set of pedestrian templates. The biggest challenge that this problem offers is to model the huge amount of variations in the shapes, pose, size and appearance of humans and their backgrounds. In addition, unless a good estimate of both the size and position of the pedestrian is known then it becomes a very computationally expensive approach. To increase reliability, some systems, such as [12], integrate multiple cues such as stereo, skin color, face and shape pattern to detect pedestrians. However, skin color is very sensitive to illumination changes and face detection can identify only pedestrians facing the camera. These systems illustrate that stereo and shape are more reliable and helpful cues than color and face detection in general situations [6].

The results of many of these 2D based approaches are depreciated in unconstrained real-world environments due to dynamic conditions such as rapidly changing lighting conditions causing shadows, pedestrian occlusion and the large variability in a pedestrian's local and global appearance due to pose, orientation and clothing. 3D stereo information has been proposed as a technique to guide pedestrian detection, as stereo and shape are more reliable and helpful cues than color and face detection in general situations. The use of stereo information carries with it some distinct advantages over conventional 2D techniques [14]: (a) it allows explicit occlusion analysis and is robust to illumination changes; (b) the real size of an object derived from the disparity map pro-

VSSN'06, October 27, 2006, Santa Barbara, California, USA.

vides a more accurate classification metric than the image size of the object; (c) using stereo cameras both stationary and moving objects can be detected.

In this paper we present a significant improvement of our existing pedestrian detection technique descried in [8]. This technique can be applied using a stereo camera system that is located, as most surveillance cameras are, above human height and orientated at approximately a 45 degree angle to look down on pedestrians below, making this technique applicable to both indoor and outdoor scenarios. The technique is based on the use of disparity information, ground plane estimation and basic human biometric information. We can detect pedestrians even in the presence of severe occlusion or a lack of reliable disparity data. We also make reliable choices in ambiguous areas since candidate pedestrians are selected using the disparity of head regions. These are usually highly textured and unoccluded, and therefore more reliable in a disparity image than homogeneous or occluded regions. As the technique uses disparity information, pedestrian characteristics, such as the variability in clothing colour, for example, does not affect the approach. In addition, biometric information, based on the *Golden Ratio* is used to remove regions that do not adhere to a pedestrian's global shape.

This paper extends the work described in [8] with two important contributions. The first is the enhancement of the dense disparity algorithm which is the basis for obtaining 3D information from a given scene and without which the accuracy of the pedestrian detection technique deteriorates significantly. To obtain a disparity map of a given scene, correspondences have to be made between pixels in one image, to pixels in a second image of the same scene taken from a different position. This, in itself, is a difficult process, especially in areas of homogeneous colour with little texture, where the correct correspondence of a single pixel may be ambiguous or non-existent.

The second contribution of this paper involves the clustering of these disparity values into individual pedestrian regions. This technique differs to that described in [8] in that it incorporates a biometric pedestrian model directly into the object clustering process. This technique is less likely to cluster two pedestrians into a single region, can overcome large gaps of erroneous disparity data, has lower complexity and results in better pedestrian segmentation.

This paper is organized as follows: Section 2 gives an overview of the system and the components of this which are the key contributions in this paper. Section 3 presents the details of the developed algorithmic approach. Firstly, we describe improvements to the the dense disparity estimation process; we then illustrate how individual pedestrians are segmented and post-processed. In Section 4 we present experimental results from a real world outdoor situation containing multiple pedestrians at various depths, some with severe occlusion, displaying a large variability in both local and global appearance. Finally, Section 5 details conclusions and future work.

2. SYSTEM OVERVIEW

Figure 1 illustrates an overview of the system and the key contributions of this paper, which are highlighted in the diagram. The stereo camera setup involves two digital lenses with a resolution of 640x480 pixels, and a baseline of 10cm. As a starting point to the algorithmic process, Ground Control Points (GCPs) are obtained using the technique defined by us earlier in [7]. A new stage in the disparity estimation process is introduced which is described in Section 3.1. These GCPs are interpolated throughout the image by applying predefined knowledge of the orientation of the 3D position of the GCPs, with respect to the 3D groundplane within the scene. This is where we introduce an extra stage into the disparity estimation process which improves the resultant disparity map, especially in areas where there is a lack of texture. This is a very significant improvement to the system as the pedestrian segmentation process is highly dependent on the quality of the input disparity map and without this some pedestrian regions are likely to be split into more than one region due to horizontal areas of homogeneous texture.

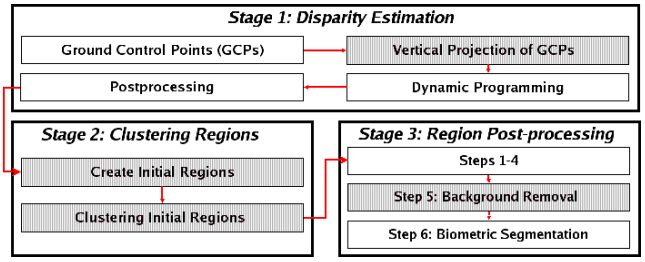


Figure 1: System Overview

The disparity maps are then generated and post-processed as described by us previously in [8]. The second contribution of this paper involves clustering these disparity values into pedestrian regions described in Section 3.2.1. This clustering process differs to that described in [8] in that it incorporates a biometric pedestrian model directly into the object clustering process. This technique is less likely to cluster two pedestrians into a single region, can overcome large areas of erroneous disparity data, has less complexity and results in better pedestrian segmentation.

Finally, in Section 3.2.2, the clustered regions are post-processed into individual pedestrians. As shown in Figure 1, this is a 5-step process. The first four steps remove background regions, which include regions due to noise. The fifth step improves on the post-processing steps set out in [8] to segment pedestrians which are at the same depth as background regions, such as walls. The fifth step applies biometric information to segment any possible regions that may contain two or more pedestrians that exist in individual regions. This final step mirrors a post-processing step defined in [8].

3. ALGORITHMIC DETAILS

3.1 Enhanced Disparity Estimation

The technique used to obtain dense disparity information in this paper is based on the work described in [7]. This details a dynamic programming based stereo correspondence technique that has been specifically developed for pedestrian surveillance type applications. The technique reduces artifacts in the calculated disparity map by integrating certain constraints into the dense disparity estimation algorithm. The two main constraints it invokes are the use of a *dynamic* disparity limit constraint, which limits the region where a match for a pixel in one image can occur in a second image.

This limit is dynamic and changes throughout the matching process. The second constraint involves the use of highly reliable matched pixels, known as Ground Control Points (GCPs) [4], to help guide results. Figure 2(b) shows the GCPs obtained from the input scene shown in Figure 2(a).

This technique, however, can still suffer from disparity estimation errors in large areas of homogeneous colour. Figure 2 illustrates an example scenario. In Figure 2(c) the disparity through the midsection of a pedestrian on the lefthand side is incorrect. This type of artifact occurs as there is no texture within the area of the pedestrians torso. GCPs can therefore not be found within this region, and therefore cannot be used to guide results. In addition, the height of the region of homogeneous color is too great for inter scanline consistency to be enforced by the one-pass dynamic programming based disparity estimation technique. If the inter scanline cost is increased, the dense disparity map will become blocky and results will deteriorate.

We improve the dense disparity estimation technique by applying predefined knowledge of the orientation of the 3D position of the Final Ground Control Points (FGCPs), obtained in [7], with respect to the 3D groundplane within the scene. In general, most surveillance cameras are located above human height and orientated at approximately a 45 degree angle to look down on pedestrians, and the groundplane, below. We obtain the 3D position of this groundplane via a technique described in [8]. By applying the reasonable assumption that all the objects in the scene (i.e. people) are vertical with respect to the 3D groundplane, we can improve the resultant disparity map by extending the FGCPs across regions of homogeneous texture.

An initial step in this process is clustering FGCPs into discrete areas of 3D space. This is a simple implementation of 8 neighbourhood connected components, where two separate FGCPs can be clustered together if they have a Euclidean distance in 3D space of less than a threshold, α. We set α to 25cm. The 3D position of a FGCP point, $p^{3d} = \{x, y, z\}$, is obtained via triangulation, as defined in [11]. This clustering is a necessary step as we do not want to extend the FGCPs from one distinct foreground object into another.

The perpendicular projection of p^{3d} onto the groundplane is then obtained, via the following technique. Let $\{A, B, C, D\}$ represent the equation of the groundplane in 3D, where A, B, and C are the X, Y, and Z components of the surface normal, and D is the distance value. The perpendicular projection of p^{3d} onto a groundplane point, q^{3d}, is obtained by finding the intersection of the line defined by the points $\{x, y, z\}$ and $\{x+A, y+B, z+C\}$, and the 3D groundplane. q^{3d} is defined as $\{x + (A * t), y + (B * t), z + (C * t)\}$, where

$$t = -\frac{A * x + B * y + C * z + D}{\sqrt{A^2 + B^2 + C^2}} \quad (1)$$

If p^{3d} is within the predefined search space defined in [8], i.e. if the height, h, of the point above the groundplane is $height_{min} < h < height_{max}$ and $z < z_{max}$, then each pixel in the image between p^{2d} and q^{2d} is traversed, where p^{2d} and q^{2d} are the projections of p^{3d} and q^{3d} onto the image plane respectively. For each point between p^{2d} and q^{2d} the disparity is interpolated and a *new* FGCP is created. This path traversal is stopped at any point, (x, y), from p^{2d} onwards if

1. (x, y) is the last edge on the path
2. (x, y) is a FGCP
3. $SAD_c((x, y), ((x+d, y)) \geq t_{MaxAccept}$, where $(x+d, y)$ is the point in image 2 where d is the interpolated disparity at the point (x, y), SAD_c is the sum of absolute colour differences and $t_{MaxAccept}$ is a threshold representing the maximum acceptable difference
4. $SAD_g((x, y), ((x+d, y)) \geq t_{MaxAccept}$, where SAD_g is the sum of absolute gradient differences

The third and fourth tests ensure that these extended disparity values do not result in poor correspondence matches. In our experiments, $height_{min}$ and $height_{max}$ are set to 0.9 and 2.4 meters respectively, which represent the minimum and maximum expected pedestrian height above the groundplane. z_{max} is set to 8 meters, due to the image resolution and the degradation of accurate stereo information beyond this distance.

If the second item in the list is the one to cause the traversal of the path from p^{2d} and q^{2d} to stop, then before the stop is made it is tested to see if the two regions can be merged together. The two separate regions can be clustered together if the two neighbouring FGCPs on the path have a distance in 3D space of less than a Euclidean distance of α to each other. Figure 2(d) shows the results from this stage.

The final stage in this process involves traversing the image horizontally with respect to the groundplane and interpolating FGCPs across areas within regions. For each FGCP we obtain the 3D path that is horizontal with respect to the groundplane, then we project this path onto the image plane. This path is traversed from left to right and, if possible, we interpolate FGCPs across areas where there are no *original* FGCPs that lie between the bounds of individual FGCP regions. If, along the path, two FGCPs from the same region are separated by a gap where there are no FGCPs, then it is tested to see if the tests defined in items 2 and 3 of the previous list are satisfied for every point within the gap. If this is true then the disparity is interpolated between these two points. Figure 2(e) shows the results from this stage. Finally, Background Ground Control Points (BGCPs) are obtained using background disparity and edge models and a dynamic programming based disparity estimation technique is applied as described in [7].

This addition to the disparity estimation technique has the potential to greatly improve the resultant disparity map in certain scenarios. For example, Figure 2(f) shows the dense disparity map obtained with this process. Notice how the disparity flows more smoothly vertically and the consistency of the disparity within homogeneous regions is improved.

Unfortunately, if incorrect GCPs are obtained initially then they too will be extended vertically and horizontally. This would decrease the quality of the resultant disparity map significantly under certain conditions. The probability of this scenario occurring is reduced by altering the second post-processing step of FGCPs defined in [7] to make it more stringent in its choice of FGCPs. Previously, this post-processing step assessed a value, val_{col}, which is associated with each FGCP. This value represents the number of neighbouring pixels around the FGCP pixel that agreed with the FGCPs choice of disparity for that point. Therefore, the higher the val_{col} value, the more likely that the FGCP match was a good one. The post-processing step looked at each FGCP's val_{col} and compared it to the val_{col} value of all other possible disparities that the FGCP could

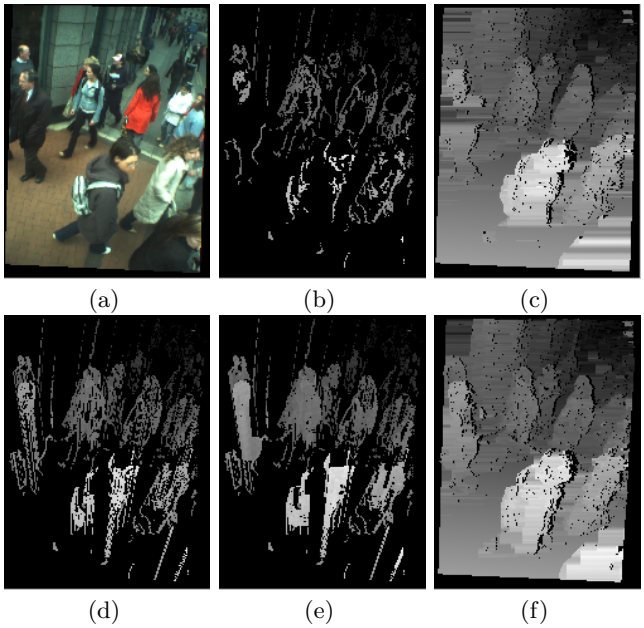


Figure 2: (a) Input Image; (b) FGCPs; (c) Original Disparity; (d) Extend FGCPs Vertically; (e) Extend FGCPs Horizontally; (f) Improved Disparity

be matched to. The post-processing step then removed any FGCP that could be matched to a second possible disparity that has val_{col} value equal to half its val_{col} value. In the new system, this step has been altered so that any FGCP that can be matched to *any* other second possible disparity, is removed. The resultant FGCPs are thereby more likely to be true, and due to the technique used to obtain the FGCPs, there are still numerous FGCPs that can be used to drive the disparity estimation technique.

3.2 Enhanced Pedestrian Detection

Before individual pedestrians are detected, the disparity is post-processed, as described in [8]. This is a two stage process. The first step removes disparity points that are outside a predefined search space mentioned in Section 3.1. The second post-processing step removes erroneous areas from the resultant disparity map. After post-processing has been completed, we refer to the remaining disparity points as *foreground* disparities. Figure 3(1a) displays an example of these foreground disparities. In Figure 3, the first column represents points on the 2D image plane, the second column displays these points in 3D from above, in a birds-eye view, and the third column displays the same 3D points but from a front elevation, so that the normal of the groundplane is vertical in the resultant image. In this front elevation, the brown line represents the detected 3D groundplane. The image scene that the disparities were obtained from in Figure 3 can be seen in Figure 4(a).

In [8] a technique for detecting pedestrians was described in which foreground disparities were clustered together into regions. Each region was then compared to a pedestrian model based on the golden ratio. Depending on the outcome of this comparison, the region was either split into multiple pedestrians, left intact or removed. This technique is very dependent on the quality of the disparity map which should be smooth and detailed; continuous and even surfaces should produce a region with smooth disparity values with their boundaries are precisely delineated, while small surface elements should be detected as separately distinguishable regions [15]. However, it is not easy for a stereo algorithm to satisfy these two requirements at the same time. Algorithms that can produce a smooth disparity map tend to miss the details and those that can produce a detailed map tend to be noisy. In our system, we wish to obtain a relatively smooth disparity map as we are not too concerned with small subtle differences in disparity because our ultimate application is detecting and counting pedestrians.

The smoothing effect, in certain scenarios, can cause problems to the technique proposed in [8] as it assumes a disparity map is obtained with well-defined object boundaries. To highlight how these conditions occur we take, for example, the section of the 3D points in Figure 4(a), shown in Figure 4(b) observed from a birds-eye view. The majority of these points belong to the pedestrian labeled **A** in Figure 4(a). In order to segment this pedestrian, all the 3D points belonging to the pedestrian should be clustered together into one distinct region, as shown in Figure 4(d). However, due to the smoothing effect, the disparities may change gradually from 1 disparity level to another at object boundaries. Figure 4(e) displays some of these *smoothed* points in red.

These smoothed points can have adverse effects on the technique proposed in [8]. If an object of high disparity is above a second object of low disparity in the image and there exists a set of *smoothed* points from the high disparity object to the low disparity object, where by each jump in the set of *smoothed* points is below the maximum distance threshold to merge the two regions, then both rules for joining the two regions would be adhered to and the two regions would be incorrectly joined. This is a scenario that can occur relatively frequently when pedestrian density is high.

The technique proposed in this paper incorporates the biometric pedestrian model described in [8] into the object clustering technique. An advantage of this is that the technique is less likely to cluster two pedestrians into a single region. It is not affected by *smoothed* points between two distinct objects, and so is less dependent on an ideal disparity map. In addition the process is able to cope with regions of poor disparity data better than that of [8], and this improvement results in one less post-processing step. The complexity of the overall clustering process is reduced when compared to that of [8]. This is due to it not being necessary to recalculate the average disparity for a region within the biometric bounds as described in [8] every time two regions are merged.

In order to cluster a single pedestrian into a distinct region, we would like to obtain the central axis of the region and based on a predefined model, cluster all points within a certain distance of this axis. The central axis is that which runs through the center of a pedestrian region and is parallel to the 3D groundplane normal. This idea can be represented in 3D by a cylinder that is centered on the pedestrian's central axis and has a radius of r. Any 3D point that is inside of this cylinder is a point that belongs to the pedestrian. Figure 4(f) illustrates this idea, where the circle represents the cylinder as seen from a birds-eye view. For each cluster, the position of the central axis is very important. If it does not run through the center of a cluster then it is

very likely that the object can become clustered into two or more separate regions as the cylinder will not be correctly positioned around the object. Figure 4(e) illustrates this potential problem. The central axis of each cylinder is therefore constantly repositioned to the center of the cluster of points it represents throughout the clustering process.

The value for r can be obtained using the biometric pedestrian model, as described in [8]. The height above the groundplane can be used to define the proportions of a human body by applying the *Golden Ratio*, Φ, ($\Phi = \sqrt{5} * 0.5 + 0.5 \simeq 1.618$) [2]. Using Φ and a humans height various other points on the human body can be defined, such as the width of the shoulders, d_{ws}, or the head, d_{wh}; the distance from the top of the head to the neck, d_{hn}, or the eyes, d_{he}. Using this process, each pedestrian will have a different value r, which is solely based on the pedestrian's height, thereby eliminating the need for any other externally defined thresholds.

3.2.1 Region Clustering

The clustering of the foreground 3D points is a two stage process. The first stage is clustering via an 8 neighbourhood connected components algorithm. Each region is initialised using a single 3D point, p. The central axis for this region is set through p, parallel to the groundplane normal defined by the vector $\{A, B, C\}$. The height above the groundplane of this point, p_h, is stored as the regions maximum height, reg_h, and the point where the central axis cuts the groundplane is stored as the regions average groundplane point, reg_{gp}.

To test if another single pixel, q, is allowed to merge with the region, reg, we first obtain the maximum height contained in q and reg. This height value is used to define a 3D distance value, β, using the biometric pedestrian model. β is used as the radius of the cylinder for the region reg. q is allowed to merge with reg if

- in 2D, q neighbours a pixel, p_{reg}, of *identical* disparity in reg
- in 3D, $dist(q, p_{reg}) \leq \beta$
- in 3D, $dist(q, reg_{cx}) \leq \beta$

where $dist$ is the perpendicular Euclidean distance, and reg_{cx} is the region's central axis. If q is allowed to merge with reg then p_h is updated if necessary, and reg_{gp} is updated so that the average groundplane point takes into account the point where q is perpendicularly projected onto the groundplane.

To test if the region, reg^1, is allowed to merge with a second region, reg^2, we first obtain the maximum height contained in reg^1 and reg^2. This height value is used to define β. reg^1 is allowed to merge with reg^2 if

- in 2D, a pixel in reg^1, p^1_{reg}, neighbours a pixel in reg^2, p^2_{reg}
- in 3D, $dist(p^1_{reg}, p^2_{reg}) \leq \beta$
- in 3D, $dist(p^1_{reg}, reg^2_{cx}) \leq \beta$, or vice versa
- in 3D, $dist(reg^1_{cx}, reg^2_{cx}) \leq \beta$

if reg^1 is allowed to merge with reg^2 then p^1_h is updated if necessary, and the reg^1_{gp} is updated so that the average groundplane point takes into account reg^2_{gp}.

In the first stage of the clustering process β is initialised using the height value obtained and the biometric distance d_{he}. This initialises β as a value of roughly 0.05% of the height of a pedestrian. This small distance is intended to create a large number of small regions. Clusters are initialised in this way to avoid two areas of separate objects, that are separated by a small Euclidean distance in 3D, to become merged. By starting with a low value for β, and gradually increasing this from d_{he} to d_{ws} (see below) we can force each separate object region to grow in isolation and avoid being merged together.

The second stage of the clustering process increases β in 3 steps, $\beta = d_{wh}$, $\beta = d_{hn}$ and $\beta = d_{ws}$. It allows the merging of regions using *only* the 3D criteria as described in stage one. In this way two regions can be merged even if they do not appear directly beside each other in 2D. Instead, the perpendicular projection onto the groundplane for each point in each region is then obtained, as described in Section 3.1. This path is traversed, and any two regions can be tested and possibly merged together if they both appear on that path. The reason for implementing this stage in this way is two-fold. Firstly, the disparity map can still obtain erroneous areas, which if large enough can possibly cut a pedestrian in two, as described in [8]. Using this technique even if this occurs the clustering of the pedestrian's points can be achieved. Secondly, even if two pedestrians occur at the same depth, their heads are normally separated and create separate regions first, by then following these regions vertically down the rest of the pedestrians body will be separated in isolation to the second pedestrian and this will result in better segmentation.

This process has to be slightly altered, however, for objects that occur further than a particular distance, $dist_z$, from the camera. The further away an object is from the camera, the smaller the disparity. The disparity estimation process obtains disparities, d, where $d \in \mathbb{Z}_+$. This means that if the disparity changes within an object then the disparity difference has to be ≥ 1. When the object is close to the camera, a change in disparity of 1 between two pixels, p^{2d} to q^{2d}, will still result in a smooth surface as the Euclidean distance between the 3D position of the points, p^{3d} to q^{3d}, will be relatively small. However, the farther away the object becomes, the larger the Euclidean distance. For example, if the disparity values at p^{2d} and q^{2d} were 1 and 0 respectively, then the Euclidean distance from p^{3d} to q^{2d} becomes ∞. For pedestrians who are greater than $dist_z$ from the camera, a change of 1 in disparity becomes greater than the value of β. This means that although the disparity change is very small, the pedestrian cannot merge properly. To combat this the Euclidean distance resulting from a disparity change of 1 at every point is tested, if the distance is greater than β, then β is redefined as that distance, up to a threshold of twice the original value of β.

3.2.2 Postprocessing

The final stage of our process to detect pedestrians involves post-processing regions using the first 4 post-processing steps as described in [8]. These steps remove noise and background regions which contain no foreground objects. However, in [8], pedestrians at relatively the same depth as background objects, such as walls often fail to be segmented properly, especially when two or more pedestrians become clustered into the same region. This scenario is also possi-

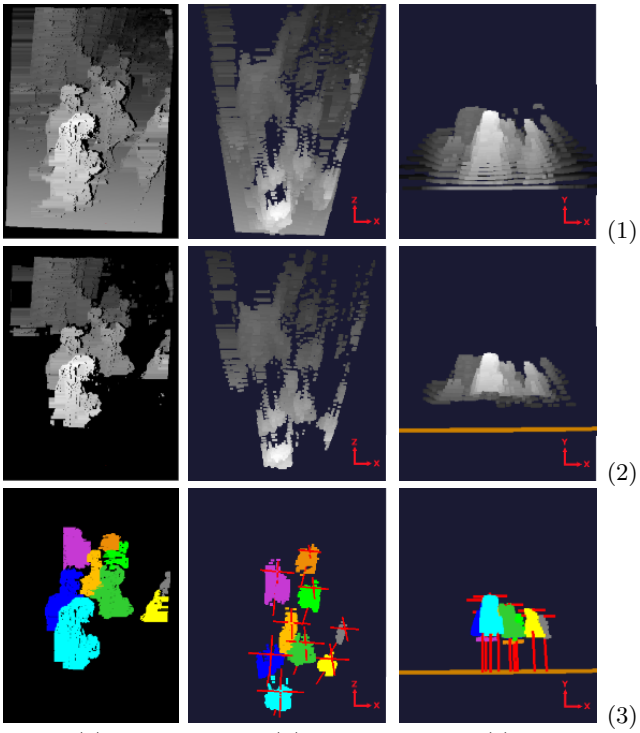


Figure 3: (Row 1) Input Disparity; (Row 2) Foreground Disparity; (Row 3) Detected Pedestrians; (Col a) 2D; (Col b) 3D Birdseye View; (Col c) 3D Front View

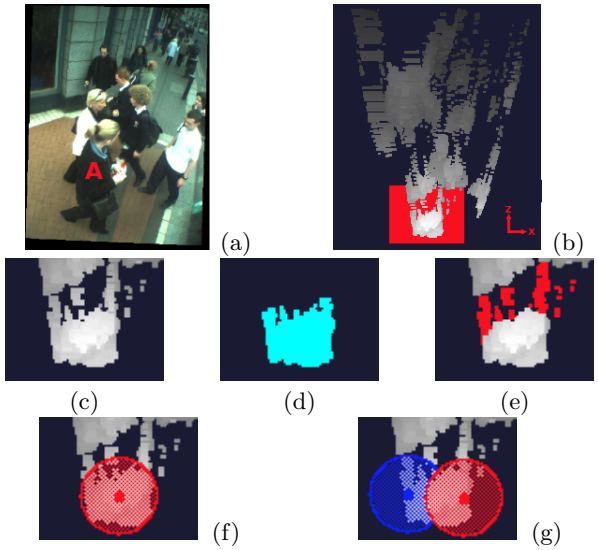


Figure 4: Region Clustering; (a) Input Scene; (b) 3D Birdseye View; (c) 3D Birdseye View Section; (d) Required Clustered Region; (e) Region Boundary Smoothed Points; (f) Biometric Clustering; (g) Incorrect Central Axis

ble using the technique set out in this paper. To separate out these foreground objects we make use of the background edge model described in [8]. For each edge in a region, it is classified, using the background edge model, as either a foreground or a background edge. If the total percentage of background edges is greater than 25%, then the biggest gap, horizontal to the groundplane, between two foreground edges within the region is obtained and removed. If the single region is split into two regions, then two new regions are created. This process is continued until the total percentage of background edges within each region is less than 25%. This post-processing step greatly reduces the number of pedestrians who cannot be detected as they are at the same depth as large background objects.

The final post-processing step is uses biometric information to segment multiple pedestrians who exist in an individual region, as described in [8]. This occurs within regions where β has been altered and increased.

4. EXPERIMENTAL RESULTS

Example results of this pedestrian detection process can be seen in Figure 3 (Row 3). In these images, each differently coloured region represents an individual pedestrian. In Figures 3 (3b) and (3c), a red line, vertical to the groundplane is present for each region, and this represents the central axis of each region, the red lines parallel to the groundplane which are connected to each central axis illustrate the final size of β for each region.

Table 1 displays an overview of results for 1000 images at various pedestrian densities and compares them to the results for the same dataset obtained using the technique proposed in [8]. The images are taken from two separate image sequences from a stereo camera mounted on a traffic light pole on Grafton Street, a busy pedestrianised shopping street in Dublin city center. The first sequence was taken at 10am on a Summer's morning and has relatively low pedestrian flow density, containing, on average 1.17 pedestrians per frame, with a maximum of 7 pedestrians in 1 given image. The second sequence was taken at lunchtime two days later, and the pedestrian flow density is increased to an average of 3.42 pedestrians per frame, with a maximum of 7 pedestrians in 1 given image frame. On both occasions the weather is dry with constant cloud cover. These weather conditions do minimise shadows cast and background illumination changes, however tests on other conditions indicate that our approach is able to cope well with these type of illumination changes, see Figure 7 for an illustrative example.

In Table 1, the first column represents the total number of pedestrians that appear in a given image, for example, $1-3$ indicates that in a given image there are between 1 and 3 pedestrians present, whereas >6 indicates that there are more than 6 pedestrians in each image. n_{ped} represents the total number of pedestrians that exist within these images. For computing precision and recall values, a correctly segmented pedestrian is defined as a region that contains at least the pedestrian's head and no substantial area of a second pedestrian. All other regions that are detected, such as those due to prams, bicycles, etc are labeled as a false positive region. n_{mult} show the number of pedestrians that were not segmented correctly due to the region containing more than one pedestrian and n_{over} represents the number of pedestrians oversegmented. Total [8] displays the results obtained from the same dataset using the system as described in [8].

From these results it can be seen that there are large im-

	n_{ped}	Precision	Recall	n_{mult}	n_{over}
1-3	964	94.10	96.05	8	48
4-6	1186	94.10	94.18	26	37
>6	521	93.89	91.36	15	14
Total	2671	94.06	94.31	49	99
Total [8]	2671	84.88	82.78	291	69

Table 1: Results Overview

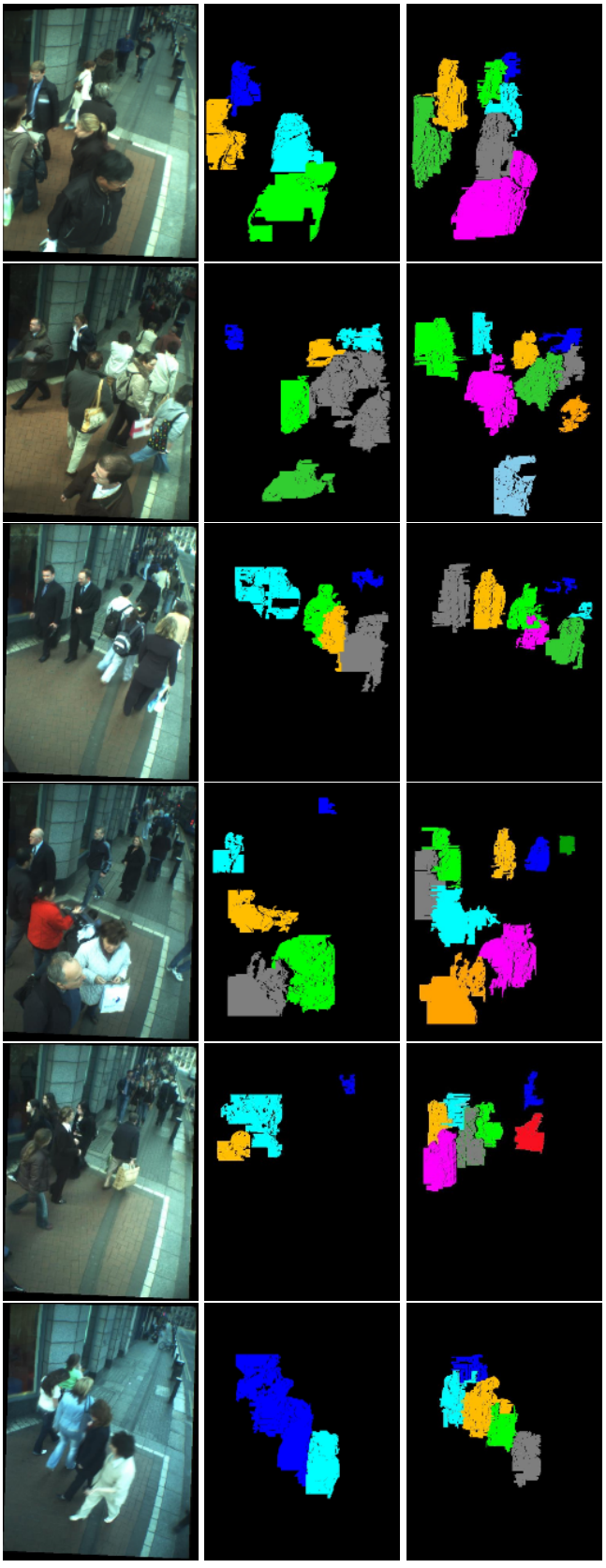
provements in the numbers of pedestrians detected throughout all three categories of pedestrian density flow. These results show an increase of both precision, from 84% to 94%, and recall, from 82% to 94%. In addition, using the technique in this paper only 32%, or 49, of the pedestrians who have *not* been segmented correctly can be attributed to the pedestrian becoming merged with another pedestrian. This is compared to 63%, or 291, in [8]. This decrease in n_{mult} indicates that the technique presented in this paper is able to segment the pedestrians more robustly than the work described in [8]. Figure 5 illustrates examples where correct pedestrian segmentation has been achieved, where this was not possible previously.

5. CONCLUSIONS AND FUTURE WORK

In this paper we presented refinements leading to a significant improvements in an existing pedestrian detection technique. This technique is able to extract multiple pedestrians, of varying orientations and appearances, from challenging real world scenes even when faced with large and multiple occlusions. The technique is also robust to changing background lighting conditions and effects, such as shadows, see Figure 7. The technique applies an enhanced method from which reliable disparity information can be obtained even from untextured homogeneous areas within a scene. This is used in conjunction with ground plane estimation and biometric information, to obtain reliable pedestrian regions. These regions are robust to erroneous areas of disparity data and severe pedestrian occlusion, which often occurs in unconstrained scenarios.

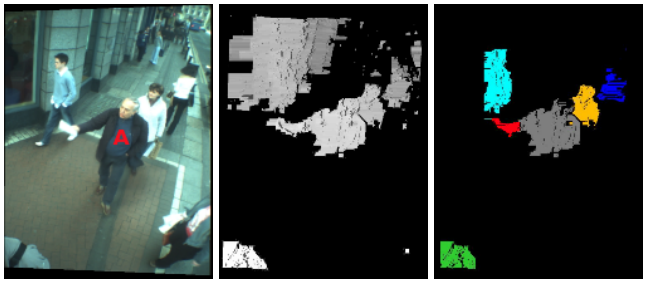
As seen in section 4, 32% of the pedestrians who have *not* been segmented correctly can be attributed to the pedestrian becoming merged with another pedestrian which is where our approach fails. Improvements to the biometric segmentation could address this. However, it is unrealistic to define and search for every possible orientation of a pedestrian in a single region, with another object. The use of other biometric information, such as skin colour, could be used for aiding this segmentation. Another possible technique could be to use face detection algorithms, set to obtain a high value for recall and using the 3D stereo information in cooperation with a biometric model to validate correct face regions. However a difficulty with this is that weather conditions sometimes mean pedestrians wear hats or hooded jackets or carry umbrellas thus concealing much of their faces, and the positioning of the cameras at a 45 degree angle above the pedestrian means a clear shot of the face is often not present.

The technique proposed in this paper does increase the total number of ovesegmentations on pedestrians by just over 40%. This increase occurs as the biometric model is included in the clustering process, and therefore the clustering process adheres to this model more rigidly than that described in [8]. A typical case of oversegmentation of a pedestrian can



(a) (b) (c)

Figure 5: Detection Improvements (a) Input; (b) Segmented Pedestrians using previous detection system ; (c) Segmented Pedestrians using this system



(a) (b) (c)

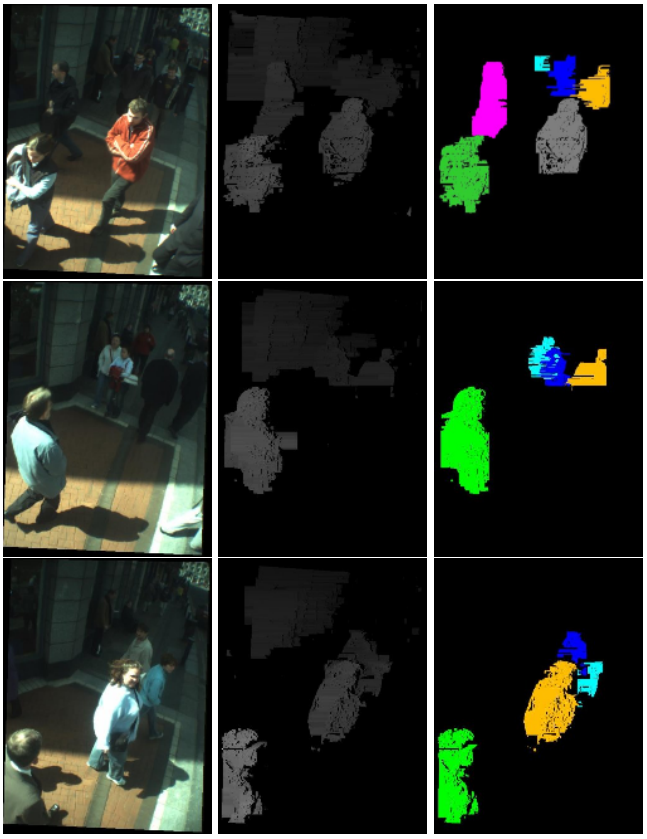
Figure 6: Oversegmentation (a) Input; (b) Foreground Disparity; (c) Segmented Pedestrians

be seen in Figure 6, where the pedestrian labeled **A**, has his arm outstretched. This arm is at a distance greater than the radius of the the cylinder surrounding the pedestrian cluster, and therefore gets segmented as a separate region. This poses a problem for the technique as increasing the radius increases the possibility of including a second pedestrian within one region. This type of oversegmentation is also possible if pedestrians are wearing large backpacks, or are hunched over. In addition, we notice that if pedestrians are in a wheelchair, the biometric model may not result in good classification, as the model assumes a pedestrian is standing upright, therefore wheelchair pedestrians can be oversegmented. Future work could improve upon the clustering process so that it uses the cylinder based biometric model for initial clustering upto a certain point, but to incorporate a more sophisticated pedestrian model into a final clustering step.

Depending on the scenario, the detection of objects other than pedestrians, for example push prams, buggies or bicycles, within a scene is not required. In this paper, these objects tend not to appear as detected objects as before individual pedestrians are detected, the disparity is post-processed, and all points that are defined as outside a predefined search space, are removed. This post-processing step removes regions that we consider to be irrelevant to our search for pedestrians, such as groundplane points, which includes shadows cast by pedestrians and other objects. At the moment the system removes all points under 0.9 meters (around 3 foot) in height above the groundplane. However, applying this post-processing technique can also remove small children or people in wheelchairs if they are below the 0.9 meter threshold. This threshold should be investigated to determine the ideal threshold for a given scenario, to increase recall without an adverse effect on precision.

The use of temporal data should also be investigated. This information could prove very useful as an additional cue for pedestrian segmentation. Currently, each frame of a video sequence is treated independently. The temporal information that can be obtained using a video sequence could also be employed as a feedback into the dense disparity estimation technique to further improve results, possibly helping to detect a percentage of missed pedestrians.

Finally, our work should be benchmarked against other techniques and there are two possibilities for this. ETISEO [1] is an evaluation campaign to evaluate vision techniques for video surveillance applications. The video data used is single and multi-view surveillance and the ground truth



(a) (b) (c)

Figure 7: Robustness to Shadows (a) Input; (b) Foreground Disparity; (c) Segmented Pedestrians

is annotations and classifications of persons, vehicles and groups. The tasks include detection and tracking of physical objects, and event recognition. PETS (Performance Evaluation of Tracking & Surveillance) [3] also evaluates object detection and tracking for video surveillance using multi-view/multi-camera surveillance video the task is event detection for events such as luggage being left in public places. However, while participation in such open, metrics-based evaluation campaigns would allow us to compare our techniques against those of others more directly, the tasks in PETS and ETISEO are not pedestrian counting but person tracking and event detection using the richness of a multiple camera surveillance setup, whereas our task is pedestrian counting using just one camera location. Nonetheless we do plan to try our techniques on other image data and doing so on the data used in ETISEO and/or PETS would allow some element of comparison to take place.

6. ACKNOWLEDGMENTS

This material is based on works supported by Science Foundation Ireland under Grant No. 03/IN.3/I361. The authors would also like to express their gratitude to both Mitsubishi Electric Research Labs (MERL) and Dublin City Council (DCC) for their support and contribution to this work.

7. REFERENCES

[1] http://www.etiseo.net, 2006.

[2] http://www.goldennumber.net, 2006.

[3] http://www.petsmetrics.net, 2006.

[4] B. C. C. Kim, K.M. Lee and S. Lee. A dense stereo matching using two-pass dynamic programming with generalized ground control points. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1075–1082, 2005.

[5] D. Gavrila and J. Giebel. Shape-based pedestrian detection and tracking. In *IEEE Intelligent Vehicles Symposium*, pages 215–220, 2002.

[6] D. I.Haritaoglu and L. Davis. W4S: A real time system for detecting and tracking people in 2.5D. In *European Conf. on Computer Vision*, pages 877–892, 1998.

[7] P. Kelly, E. Cooke, N. E. O'Connor, and A. F. Smeaton. 3D image analysis for pedestrian detection. In *Image Analysis for Multimedia Interactive Services*, pages 177–180, 2006.

[8] P. Kelly, E. Cooke, N. E. O'Connor, and A. F. Smeaton. Pedestrian detection using stereo and biometric information. In *Int. Conf. on Image Analysis and Recognition*, 2006.

[9] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 878–885, 2005.

[10] C. Pai, H. Tyan, Y. Liang, H. Liao, and S. Chen. Pedestrian detection and tracking at crossroads. *Pattern Recognition*, 37(5):1025–1034, 2004.

[11] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis and Machine Vision, Second Edition*. PWS Publishing, 1999.

[12] M. H. T. Darrell, G. Gordon and J. Woodfill. Integrated person tracking using stereo, color and pattern detection. *Int. Journal of Computer Vision*, 37(2):175–185, 2000.

[13] K. Terada, D. Yoshida, S. Oe, and J. Yamaguchi. A counting method of the number of passing people using a stereo camera. In *Int. Conf. on Image Processing*, volume 2, pages 338–342, 1999.

[14] L. Zhao and C. Thorpe. Stereo and neural network-based pedestrian detection. *IEEE Trans. on Intelligent Transportation Systems*, 1(3):148–154, 2000.

[15] C. Zitnick and T. Kanade. A cooperative algorithm for stereo matching and occlusion detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(7):675–684, 2000.