ROBUST PEDESTRIAN DETECTION AND TRACKING IN CROWDED SCENES

Philip Kelly, Noel E. O'Connor and Alan F. Smeaton

Centre for Digital Video Processing, Adaptive Information Cluster, Dublin City University, Ireland

Abstract

In this paper, a robust computer vision approach to detecting and tracking pedestrians in unconstrained crowded scenes is presented. Pedestrian detection is performed via a 3D clustering process within a region-growing framework. The clustering process avoids using hard thresholds by using bio-metrically inspired constraints and a number of plan view statistics. Pedestrian tracking is achieved by formulating the track matching process as a weighted bipartite graph and using a *Weighted Maximum Cardinality Matching* scheme. The approach is evaluated using both indoor and outdoor sequences, captured using a variety of different camera placements and orientations, that feature significant challenges in terms of the number of pedestrians present, their interactions and scene lighting conditions. The evaluation is performed against a manually generated groundtruth for all sequences. Results point to the extremely accurate performance of the proposed approach in all cases.

Key words: Pedestrian Detection, Pedestrian Tracking, Stereo, Crowds
PACS: 42.30.Tz

1 Introduction

The vision of Ambient Intelligence (AmI) [1] depicts environments that are able to adapt intelligently to facilitate the requirements of the people present. AmI leverages a networked system of smart devices and sensors, which have been smoothly integrated into the environment to act as a global interface between users and information systems [2]. In this way, the control of the augmented environment becomes action oriented, responding appropriately to the behaviour of the human users present. This promises many benefits for both single individuals and larger groups of people in a variety of application scenarios.

In order for AmI to become a reality, a number of key technologies are required from a variety of disciplines [1]. These include unobtrusive sensor hardware, wireless and fixed communication systems, software design, information fusion, intelligent agents, to cite but a few. In this paper, a focus is made on the requirement for robust detection and tracking of humans in unconstrained scenes. This is a key enabling technology since knowing who is where in a scene *and* what their actions have been allows other layers in an AmI framework to infer beliefs about those people. Consider the example of an automated pedestrian traffic light system. An embedded intelligent system should be able to determine the number of people waiting to cross, whether any special assistance should be flagged for any individual pedestrian (e.g. wheelchair, children or elderly pedestrians), estimate the time needed for everyone to cross, determine the state of traffic flow on the road and ensure each person crosses the road successfully before allowing vehicular traffic to flow. Clearly detecting and tracking the pedestrians is a necessary pre-processing step. However, this poses significant challenges when pedestrian detection and tracking in unconstrained real world crowded environments is considered. For example, just because a person is in the scene doesn't mean that they want to cross the road, however, if the person walks towards the crossroads, stops and waits, then they probably do. RFID tagging is a possible solution for determining this in constrained environments, but cannot help in scenarios where there is no contact with people in a scene until they enter the environment.

Many of the person detection techniques described so far in the literature – see section 2 – make assumptions about the environmental conditions, pedestrian and background colour intensity information, the pedestrian flow, that a person will exist in the scene for a given number of frames, or that a person enters the scene un-occluded. In this paper, a robust pedestrian detection and tracking system for a single stereo camera is presented, which attempts to minimise such constraining assumptions. It is able to robustly handle:

(1) occlusion, even when multiple people enter the scene in a crowd;
(2) lack of variability in colour intensity between pedestrians and background;
(3) rapidly changing and unconstrained illumination conditions;
(4) pedestrians appearing for only a small number of frames;
(5) relatively unconstrained pedestrian movement;
(6) relatively unconstrained pedestrian pose, appearance and position with respect to the camera;
(7) varying camera heights, rotations and orientations;
(8) static pedestrians.

In addition, as the proposed pedestrian detection algorithm uses a simple biometric person model that is defined with respect to the groundplane, the system requires *no* external training to detect and track pedestrians. However, although the proposed system was designed to minimise constraining assump-

tions, a small number of inherent assumptions still exist within the system framework. They include;

(1) that pedestrians in the scene are standing upright with respect to the groundplane;
(2) that all moving objects in the scene (within the volume of interest) are caused by foreground pedestrians;
(3) that pedestrians in the scene are moving at a velocity of less than 3 metres per second.

In addition to this, the system does have a small number of drawbacks on the type of scenario it can survey. These include; (a) that a relatively flat groundplane is present within the scene, where no object of interest is located below this groundplane; (b) the camera must be orientated so that the groundplane is visible in the image plane; and (c) the system is only able to reliably detect pedestrians for a short-medium range, up to a maximum distance of 8 metres from the camera. An area of future work envisioned by the authors includes the investigation of techniques to further reduce these assumptions and limiting constraints.

The main areas of contribution of this paper are twofold. The first lies in the introduction of a novel, non-quantised, plan-view statistic (an overview of such statistics is given in section 2) which incorporates global features into the pedestrian clustering framework of the authors' previous work [3]. The use of this plan-view statistic within this framework significantly improves robustness to both over- and under-segmentation of pedestrians in comparison to [3]. The second main contribution area lies in the robust pedestrian tracking technique that has been developed. Within this area a number of contributions can be identified, which include; (a) a matching technique that incorporates a novel weighting scheme for matching pedestrians to previous tracks; (b) a series of kinematic constraints that model possible pedestrian movement through the scene and that can be used to remove implausible matches of pedestrians to previous tracks; and (c) rollback loops and post-processing steps to increase track robustness to both over-/under-segmentation.

This paper is organised as follows: section 2 gives an overview of the related work in the area of pedestrian detection and tracking techniques and outlines the benefits of stereo information within this area. Section 3 gives an overview of the key components to the overall pedestrian detection and tracking system. Sections 3.1 and 3.2 discuss the details of the proposed approach to pedestrian detection and tracking respectively. In section 4 experimental results (evaluated against a groundtruth) are provided for indoor and outdoor situations at various orientations containing multiple pedestrians at various depths, some with severe occlusion and displaying a large variability in both local and global appearance. Finally, section 5 details conclusions and future work.

2 Related Work

Robust segmentation and tracking of pedestrians within an unconstrained scene is one of the most challenging problems in computer vision. A few of the complicating factors to segmenting people include; the large variability in a person's local and global appearance and orientation [4]; occlusion of an individual by one or several other persons, or objects, especially if the person is located within a crowd; lack of visual contrast between a person and background regions. In addition, unconstrained real-world outdoor environments tend to create further challenges, such as rapidly changing lighting conditions due to varying cloud cover, shadows, reflections on windows, and moving backgrounds.

A significant amount of research literature exists on person detection and tracking. Various techniques for segmenting individual pedestrians have been investigated using traditional 2D computer vision techniques. Unfortunately, few of these, if any, produce reliable results for long periods of time in unconstrained environments [5]. Reasons for this stem from various assumptions regarding the environmental conditions and type of pedestrian flow being violated. For example, techniques, such as [6–10], depend on accurate segmentation of moving foreground objects from a background colour intensity model as a first step in their algorithmic process. This relies on an inherent assumption that there will be significant difference in colour intensity information between people and the background. Other techniques [11–14] use rhythmic features obtained from a temporal set of frames for pedestrian detection, such as the periodic leg movement of a walking human, or motion patterns unique to human beings, such as gait. However, the assumption that a person will be moving (and normally in a predefined direction), means that people standing still, or performing unconstrained and complex movement, or in crowded scenes when legs are occluded, will not be detected. Other techniques, such as [6,7], make an assumption that a person will appear in the scene un-occluded for a given period of time allowing a model of the pedestrian to be built up while they are isolated. In addition, appearance-based techniques often fail when two people get close together, as the algorithm fails to allocate the pixels to the correct model because of similarities in appearance, and tracking is lost. To increase reliability, some systems, e.g. [15], integrate multiple cues such as skin colour, face and shape pattern to detect pedestrians. However, skin colour is very sensitive to illumination changes and face detection can identify only pedestrians facing the camera.

3D stereo information has been proposed as a technique to overcome some of these issues. The use of stereo information carries with it some distinct advantages over conventional 2D techniques [5,16]:

(1) It is a powerful cue for foreground-background segmentation [17];
(2) It is not significantly affected by sudden illumination changes and shadows [18];
(3) The real size of an object derived from the disparity map provides a more accurate classification metric than the image size of the object;
(4) Occlusions of people by each other or by background objects can be detected and handled more explicitly;
(5) It permits new types of features for matching person descriptions in tracking;
(6) It provides a third, disambiguating dimension for matching temporal pedestrian positions in tracking.

However, range information also has its disadvantages; (a) it can be a noisy modality where the standard deviation of the depth value at a pixel over time is commonly of the order of 10% of the mean [5]; (b) it cannot segment foreground objects at the same depth as background regions; and (c) no technique has been developed that returns correct range information in all scenarios, all of the time. However, despite these drawbacks, the authors consider a stereo-based approach the most promising for the envisioned application scenarios.

In the literature, stereo-based range information has previously been applied in pedestrian detection scenarios. In [19] it is applied to augment a background colour intensity model to obtain foreground regions. These foreground pixels are clustered together into blob regions of discrete disparity bounds, and finally the blobs are clustered into people-shaped regions by searching through the space of possible groupings. A similar technique is used in [17] whereby foreground blobs are temporally grouped into a single region if they have similar disparity values, and the grouped region does not exceed the size range of a normal person. However, these techniques are prone to under-segmentation when faced with crowded conditions.

An inherent problem associated with mounting camera systems at oblique angles is that partial occlusion of pedestrians is likely to occur. In [20] this issue is addressed by mounting a stereo camera above a door and pointing it downward, towards the ground. In this approach, 3D points within a 3D volume of interest are selected. The groundplane is then broken up into square segments corresponding to bins in a histogram, and the 3D points are orthographically projected onto the groundplane. The more 3D points that are projected into a given bin, the higher the bin's *occupancy*. To detect people, a threshold is applied to the occupancy map and Gaussians fitted to each peak [20].

This overhead viewpoint, however, does have disadvantages. Firstly, the camera orientation is generally only applicable to indoor scenarios due to the necessary overhead camera placement structures being unavailable in outdoor environments. Secondly, a camera in this point of view generally has a limited

field of view [21], as a maximum height is constrained by a ceiling. This short height can be restrictive as the field of view can be limited unless a wide field of view lens is employed. However, this type of lens can result in significant occlusion problems in all but the central portion of the image [21]. Therefore with overhead camera viewpoints a trade-off exists between the field of view and occlusion. An advantage to using stereo cameras over monocular cameras is that this trade-off can be removed.

If the 3D groundplane is calibrated with respect to the stereo rig then 3D points can be orthographically projected onto the groundplane no matter what orientation the camera rig is positioned at, therefore allowing the occupancy map approach of [20] to be applied from stereo cameras mounted at more oblique angles. In this manner, the advantages of mounting the camera at an oblique angle, which maximises viewing volume, *and* that of an overhead view, which simplifies person modeling and reduces occlusions, can be exploited. Occupancy maps, however, have their own problems. For example, [22] illustrates that occupancy maps cannot detect a person far from the camera because the number of 3D points on a distant person is too small to create a single significant peak on the maps. However, the occupancy map is an example of one of a number of *plan-view statistics* that are used in various other techniques [22,5,23,24] where the camera is mounted at oblique angles.

Another type of plan-view statistic, proposed in [22], projects 3D voxels instead of 3D points orthographically on the floor plane, and accumulates the volumes. This allows people farther away from the camera to meet the required threshold to be segmented as a person. However, in crowded situations the peaks often connect, resulting in under-segmentation. The *height map*, another plan-view statistic, is introduced in [5] to complement two of the occupancy map's failings; namely its lack of virtually all object shape information in the vertical dimension, and the decrease in saliency in the occupancy map when the person is partially occluded by another person or object, as far fewer 3D points corresponding to the person will be visible to the camera. The height map is similar to the occupancy map but each bin is a single value, namely the height above the ground-level plane of the highest point within each vertical bin. It is effectively a simple orthographic rendering of the shape of the 3D point cloud when viewed from overhead. New people are detected if their height is over a given threshold *and* their occupancy is over a threshold. However, depending on the height threshold, children may not be detected.

All these techniques can have difficulties when dealing with substantial occlusion of a new pedestrian, where the occupancy count is unlikely to reach the minimal required thresholds. Therefore they tend to introduce assumptions that substantial occlusion does not occur before a person has been detected and added to the tracked list. In addition to this, a resolution must be chosen to quantise the 2D space into vertical bins; the resolution should be small

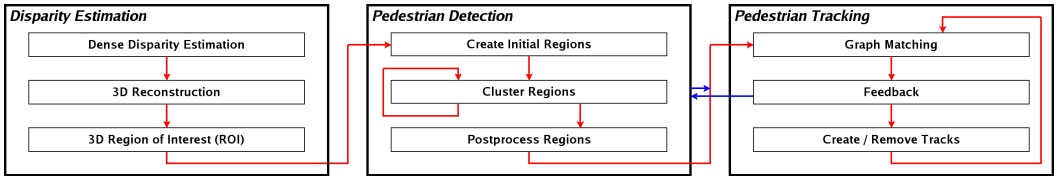


Fig. 1. System overview.

enough to represent the shapes of people in detail but also must consider the limitations imposed by the noise and resolution properties of the depth measurement system. In the proposed technique, a new plan view statistic is applied, which is used as a guide within the clustering process and is not subject to fixed thresholds. Thus it is not subject to the threshold problems outlined above. In addition, the 3D space is not quantised into discrete bins, and so a quantisation resolution does not have to be fixed.

3 System Overview

Figure 1 illustrates an overview of the proposed system. As an input to the pedestrian detection module, a dense disparity map of the scene is required. A wide variety of dense stereo correspondence algorithms have been proposed in the literature, any of which could be used to obtain the required input disparity map. However, a major flaw in many stereo-based techniques is that few attempt to apply the use of scene features and temporal information to obtain the highest *quality* disparity map that is possible within reasonable time constraints. Instead, techniques tend to apply standard stereo correspondence algorithms, and often the reasons for a specific choice of algorithm are not well justified. The disparity estimation technique employed by the authors is that of [3], which describes a dynamic programming based stereo correspondence technique that has been specifically developed for pedestrian surveillance type applications. This technique reduces artifacts in the calculated disparity map via a number of enhancements to the dense disparity estimation algorithm – however, this paper does not focus on the disparity map generation, and interested readers are directed to the relevant papers for further details. In should be noted, however, that a lower quality disparity maps does not inherently mean that the proposed pedestrian detection will fail and as such standard disparity estimation approaches can be employed to generate the input map.

From the input disparity map, a set of 3D points is obtained via triangulation – see figure 1. The disparity map is post-processed to remove artifacts and constrain the 3D points to a volume of interest (VOI). This VOI is defined by a maximum and minimum height with respect to the groundplane in the scene and a maximum distance from the camera. In our experiments all 3D points that are lower than 0.9 meters ($\simeq$ 3 foot) in height, or greater than 2.1 meters

($\simeq$ 7 foot) in height are defined as outside the VOI and so are removed. In addition, all 3D points further than 8 meters from the camera are also defined to be outside the VOI. The VOI is limited to a distance of 8 metres due to a small stereo camera rig baseline of 10cm and the degradation of accurate stereo information beyond this distance. Finally, all remaining disparity points are retained and labelled as *foreground* 3D points. For further information on the post-processing of disparity values readers are directed to [25].

The resultant foreground points are then clustered together into detected pedestrians via an iterative region growing framework. This technique is based on the approach proposed in [3] – an overview of this technique is presented in section 3.1. However, in this paper the technique is augmented with a novel, non-quantised, plan-view statistic that incorporates global features into the pedestrian clustering framework and as such reduces the over- and under-segmentation of pedestrians in comparison to that of [3].

The final stage of the pedestrian detection module – see figure 1 – involves the post-processing of the resultant clustered 3D regions to remove regions (or parts of regions) caused by noise and background objects. This architecture, whereby background-foreground segmentation is implemented *after* pedestrian regions are created, contrasts to many techniques proposed in the literature. For example, in techniques such as [26,8,9] motion segmentation techniques are employed to obtain foreground subtracted pixels, from which hypotheses of pedestrian objects are obtained. However, robust background-foreground segmentation of pixels from background models is not a trivial problem, especially in real-world conditions where rapid changes in lighting conditions can occur. To date there is no background subtraction technique that addresses all the traits required of background models in unconstrained environments. As a result, techniques that are built upon this basis are limited by the success of the underlying flawed segmentation algorithms of motion segmentation. In the proposed methodology, background subtraction techniques are applied to *guide* the final segmentation of the final clustered objects as opposed to being the basis of a technique to obtain those objects. Using this technique the reliance upon the background model is reduced significantly. For further information on the post-processing of the final regions, readers are directed to [3].

With regard to pedestrian tracking (see figure 1), the system initially detects pedestrians in each frame independently. These detected pedestrians are temporally tracked by representing previous tracks and current image pedestrians by a *Weighted Bipartite Graph*, described in section 3.2. A *Maximum Weighted Maximum Cardinality Matching* scheme is then employed, with additional kinematic constraints, to obtain the best match from previous tracks to currently detected pedestrians. A number of separate rollback loops are used to backtrack the pedestrian detection module to various states to further

reduce over-/under-segmentation of detected pedestrians and increase tracking robustness.

3.1 Pedestrian Detection

The technique for pedestrian detection proposed in this paper is based on that described in [3], where an iterative region growing framework is employed. In this paper the clustering algorithm is enhanced to significantly improve robustness to both over and under-segmentation of pedestrians. This is achieved by introducing a new plan-view statistic with a view to imposing more stringent testing upon the clustering of regions in the final iteration of the algorithm.

The pedestrian detection algorithm of [3] clusters 3D points into pedestrian shaped regions by incorporating a simple human biometric model directly into the region clustering process. This model is dependent solely on the position of the groundplane in the scene and the *Golden Ratio*, and therefore the *only* constraint on the orientation of the camera rig in the scene is that the groundplane must be in view. The groundplane must be in view in order for it to be calibrated using the stereo camera coordinate system.

An overview of the proposed pedestrian detection algorithm can be illustrated using the example of the two pedestrians in figure 2(a). Using the technique outlined in section 3, foreground 3D points are obtained – see figure 2(b), where the brighter the colour, the closer the point is to the camera. These 3D points are also illustrated (using their original colours) in figure 2(e) from a *plan-view* orientation, whereby the viewing angle is parallel with the groundplane normal and the 3D points are orthographically projected onto a 2D plane. Note that for ease of illustration, figures 2(f)-(n) are also depicted from a plan-view orientation. However it must be stressed, that this is for illustrative purposes only and the described technique clusters *3D points*, and not 2D points.

As detailed in [3], initial clusters are obtained from the foreground disparities – as illustrated in figure 2(g), where each colour represents a distinct region. Each region, reg, is defined by; (1) reg_h: the regions maximum height above the groundplane; and (2) reg_{cx}: the central axis of the region, which is the 3D line that is parallel to the 3D groundplane normal and runs through the average 3D point in the region. Figure 2(f) illustrates the heights of 3D points above the groundplane, where the brighter the colour, the greater the height.

These initial regions are then iteratively grown in a controlled manner, where the merging of two regions, reg^1 and reg^2, is permitted if $d_{cx}^{12} < \delta$, where d_{cx}^{12} is the Euclidean distance from reg_{cx}^1 and reg_{cx}^2. From this previous inequality it can be seen that the maximum distance permitted between two merging

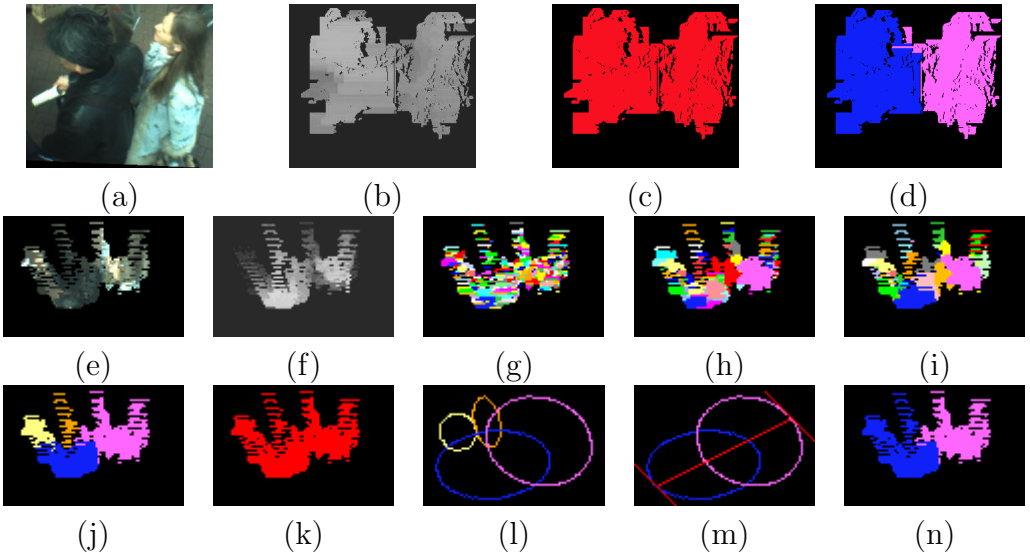


Fig. 2. (a) Image tile showing two pedestrians very close together; (b) Foreground disparity; (c) & (k) Final regions (i.e. 7^{th} iteration) without the under-segmentation test; (d) & (n) Final regions (i.e. 7^{th} iteration) with the under-segmentation test; (e) 3D points from a plan-view orientation; (f) 3D point heights; (g) Initial regions; (h) Region clustering - 1^{st} iteration; (i) 2^{nd} iteration; (j) 6^{th} iteration; (l) Best-fit ellipses of the 4 regions from (j); (m) Region diameter.

regions is constrained by the parameter δ. The value of δ is obtained by a biometric pedestrian model [25] that is based on the *Golden Ratio*, $\Phi = \sqrt{5} * 0.5 + 0.5 \simeq 1.618$. This parameter Φ can be used to define the approximate proportions of a human body if the height of the person is known [27]. Using Φ and a height value, other points on the human body can be defined, such as the width of the shoulders, |lo|, or the head, |mn|; the distance from the top of the head to the neck, |af|, or the eyes, |ad|. Appendix A illustrates these values with the aid of a diagram and lookup table. In the first stage of the clustering process $\delta = |\text{ad}|$, where the height of a region is defined by reg_h. This initialises δ as a value of roughly 0.05% of the height of the region. Regions that have a central axis within a Euclidean distance δ from reg_{cx} are then merged.

Throughout the clustering process δ is gradually increased from |ad| to |lo|. As such, δ controls the growth rate in the algorithm. By increasing the value of δ slowly, each separate object region can be allowed to grow in isolation and avoid being merged. The iteration from $\delta = |\text{ad}|$ to $\delta = |\text{lo}|$ occurs in seven distinct steps. Seven was chosen since using Φ there are 4 steps to go from |ad| to |lo|, the extra three are halfway between two steps of Φ and are needed to ensure that the regions are not grown too fast, otherwise under-segmentation is more likely to occur. Figures 2(g)-(j) depict the regions at various stages of the process. However, in the 7^{th}, and final, iteration of the clustering algorithm, the technique is prone to under-segmentation if two regions, reg^1_{cx} and reg^2_{cx}, belonging to two different pedestrians are positioned very close together, i.e.

if $d_{cx}^{12} < \delta$. The result of such an event is illustrated in figures 2(c) and (k) where two pedestrians have been clustered into one region. In addition, if for two other regions that belong to a single pedestrian d_{cx}^{12} is slightly greater than |lo|, then they will not be merged and thus over-segmentation will occur. The technique is therefore prone to both over and under-segmentation as the clustering technique is based *solely* on the position of the central axes of regions, without taking into account the global features associated with the regions.

3.1.1 Robustness to Under- and Over-segmentation

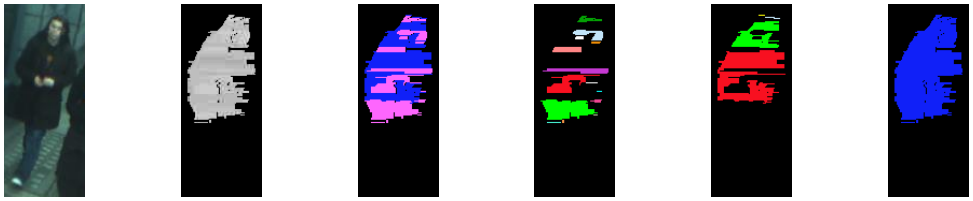
The first contribution of this work is to augment the clustering framework described previously with a novel plan-view statistic that incorporates the required global feature information from regions. This in turn leads to increased robustness to both under- and over-segmentation of pedestrians.

During the final iteration of the clustering algorithm (i.e. when δ is at its maximum value of |lo|), robustness to under-segmentation can be enhanced by invoking an additional constraint on the clustering of two regions. This additional constraint, which will be referred to as the *under-segmentation test*, is designed to compare the global shape of the two regions to determine the possible presence of two people. The under-segmentation test incorporates a novel plan-view statistic that approximates the global shape of each region by a *best-fit ellipse* around the shoulder height of each region (see section 3.1.1.1 for details on how to obtain this ellipse). Figure 2(l) illustrates the best fit ellipses for each of the four regions of figure 2(j) that exist before the final clustering iteration occurs. Using these region statistics, two regions, reg^1 and reg^2, can be merged if two constraints are passed

(1) $d_{cx}^{12} < |\text{lo}|$, which states that two regions can only join if the distance between their centres is less than the shoulder width of a person, and
(2) $\gamma < 2\,|\text{lo}|$, which is defined as the *under-segmentation test*. In this inequality, let γ be the maximum Euclidean distance between two region ellipse points on the line l, where l is a 2D line that passes through the centre of the two region ellipses – see figure 2(m).

The *under-segmentation test* ensures that for two regions to be merged, the distance across the two regions, parallel to their centres, must be less than the combined shoulder widths of two people. This constraint, in addition to the first central axis constraint, creates a powerful pair of clustering constraints that result in a significant reduction in under-segmentation.

Robustness to over-segmentation can be enhanced using similar techniques via an *over-segmentation test*. For example, if from two regions, reg^1 and reg^2, the statistics show that $d_{cx}^{12} > |\text{lo}|$ but $\gamma < 2\,|\text{lo}|$ – then the two regions may



(a) (b) (c) (d) (e) (f)

Fig. 3. Splintering of pedestrians; (a) Image data; (b) Foreground disparity; (c) Over-segmented region; (d) reg_1 sub-regions; (e) reg_2 sub-regions; (f) Merged region.

belong to *either* 1 or 2 people (if it is the latter, then merging the two regions would result in under-segmentation). In order to determine whether the two regions should be merged, further examination of the regions is required. In the proposed approach, the two regions are allowed to merge if the diameter of a second best-fit ellipse, fit to only the 3D points located above shoulder height, equates to the size of a single persons head. Using this approach, two best-fit ellipses (one from each region) are obtained, using Φ to constrain the 3D points used in the creation of the ellipse to those above neck height (i.e. higher than $|\mathrm{aj}| - |\mathrm{af}|$). If the radius of the major axis of both the ellipses are greater than half the width of a head, $\frac{|\mathrm{mn}|}{2}$, then it is determined that two head regions do indeed exist and therefore merging *cannot* occur. Otherwise, the merging of reg^1 and reg^2 is permitted. Using this technique, reg^1 and reg^2, can be merged if

(1) $d_{cx}^{12} > |\mathrm{lo}|$, and
(2) $\gamma < 2\,|\mathrm{lo}|$, and
(3) $\alpha < \frac{|\mathrm{mn}|}{2}$ and $\beta < \frac{|\mathrm{mn}|}{2}$, where α and β are the major axis diameter of the "head region" ellipses from reg^1 and reg^2 respectively.

3.1.1.1 Best-Fit Ellipse To determine the best-fit ellipse of a region, a 3D point set is created using Φ to obtain all 3D points in the region that are at or above a particular height – shoulder height is chosen in the *under-segmentation test* as the best area to fit the ellipse, rather than the region as a whole, as this area is less likely to be perturbed by objects, such as backpacks or outstretched limbs. These 3D points are then orthographically projected onto the groundplane, which removes one degree of freedom from the points. This is similar to the techniques used in the generation of other plan-view statistics such as those used in [22,5,23,24] except that the points are *not* quantised into discrete bins. The best-fit ellipse is then obtained from the resultant 2D point set in a manner similar to that presented in [28].

3.1.2 Dealing with Distant Pedestrians

A prerequisite of the proposed algorithm is good disparity estimation and 3D reconstruction. The more accurate these are, the better the subsequent segmentation. However, most stereo correspondence algorithms (including the one employed by the authors) compute the disparity of a given point to be a discrete value between 0 to n, where n is defined by the disparity limit constraint. 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, u and v, still results in a smooth surface as the Euclidean distance between the 3D position of the points, U and V, is relatively small. However, the farther away an object becomes from the camera, the greater effect a change of disparity will have in terms of Euclidean distance. For example, if the disparity values at u and v were 1 and 0 respectively, then the Euclidean distance from U to V becomes ∞.

Therefore, the farther away the pedestrian is from the camera, the more likely it becomes that the 3D points belonging to a single person will become spread out [3]. In addition, there are fewer 3D points belonging to the pedestrian and therefore the central axis of a clustered region becomes more susceptible to noise. A repercussion of this is that as the distance of a pedestrian from the camera increases, then the likelihood of the two regions, reg_1 and reg_2, belonging to the same pedestrian having either $\gamma > 2\,|\mathrm{lo}|$ or $d_{cx}^{12} > |\mathrm{lo}|$ increases. This can result in over-segmentation of a pedestrian, as seen in figure 3(c).

Solutions to this problem include to; (1) turn off the under-segmentation test for regions at distances greater than a certain distance, $dist_z$; (2) allow an increase in the value for |lo| for regions at distances greater than $dist_z$; or (3) take into account the characteristic appearance of distant over-segmented regions, and merge them appropriately. The first two options both involve an unknown threshold, $dist_z$, and both are subject to causing unnecessary under-segmentation. In this paper, the third solution is adopted and it is observed that, in general, over-segmentation at large distances results in a characteristic *splintering* of regions in 2D image space. In the proposed approach, this splintering is defined to have occurred if; (a) each of the two regions, reg_1 and reg_2, are composed of more than one disjointed sub-regions in *2D* image space – see figures 3(d) and (e) where each sub-region of each of the two regions of figure 3(c) is coloured differently; and (b) the merging of reg_1 and reg_2 would result in two or more of the sub-regions in *each* of reg_1 and reg_2 becoming connected in 2D image space – see figure 3(f) where all the sub-regions of figures 3(d) and (e) are now connected in 2D image space. Using this approach, if two regions, reg_1 and reg_2, are found to be splintered, then the under-segmentation test for the regions is not employed and reg_1 and reg_2 can be merged simply if $d_{cx}^{12} < |\mathrm{lo}|$. The authors have found that this splintering test works as well as either option (1) or (2), but without the need to set any external thresholds.

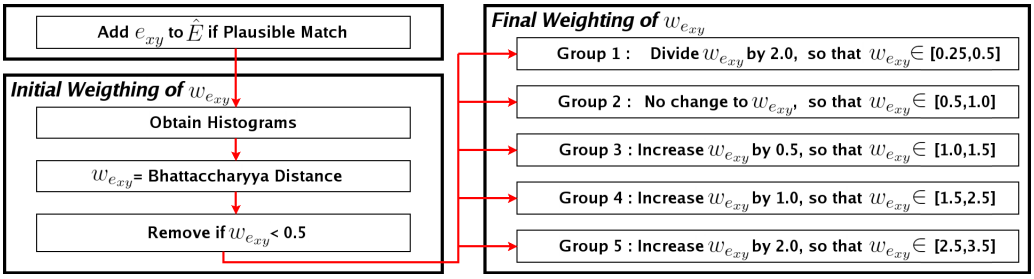


Fig. 4. Creating and weighting e_{xy}.

3.2 Pedestrian Tracking

Let $p_1, p_2, \ldots p_N$ represent the N pedestrians that have been detected in frame i as outlined in the previous section, and $t_1, t_2 \ldots t_M$ represent the M pedestrians that have been temporally tracked up to frame $i-1$. If $M = 0$ and $N > 0$, then each p_x is assigned a new track t_x, where $x = 1 \ldots N$. For all frames where $M > 0$, it is required to update the M tracks to incorporate pedestrian data from frame i. This is achieved by matching the N pedestrians in frame i to the M tracks from frame $i-1$. However, it may not be possible to match all pedestrians to tracks, or vice versa. In addition, a given pedestrian may be more or less likely to be a continuation of a certain track.

This situation can be represented by a *weighted bipartite graph*, $G = (V, E)$ [29]. A graph is bipartite if there exists a partition of the vertex set $V = V_1 \cup V_2$ so that both V_1 and V_2 are independent sets, and an edge, $e_{v_1 v_2} \in E$, can only link $v_1 \in V_1$ to $v_2 \in V_2$. In this scenario, V_1 represents the N pedestrians detected in the current frame i, and V_2 the M temporally tracked pedestrians in frame $i-1$. e_{xy} denotes a match between a pedestrian, p_x, and a track, t_y, where $x = 1 \ldots N$ and $y = 1 \ldots M$. To match pedestrians to tracks, a subset of edges, $\hat{E} \subset E$, is created, and each e_{xy} is weighted to indicate the likelihood of a match between p_x and t_y. If there is no likelihood of a match then $e_{xy} \notin \hat{E}$. Figure 4 illustrates the process in which a single edge e_{xy} is created and weighted. The creation and weighting of edges in $\hat{E}$ is described in the next section.

In order to obtain the best matching of pedestrians to tracks, a *Maximum Weighted Maximum Cardinality Matching* scheme is employed [29]. In graph theory, a *matching* in $G = (V, \hat{E})$ is a subset, S, of the edges $\hat{E}$ such that no two edges in S share a common end node. A *maximum cardinality matching* has the maximum possible number of edges and a *maximum weighted matching* is such that the sum of the weights of the edges in the matching is maximised. The scheme employed therefore maximises the number of pedestrians matched to tracks, while simultaneously obtaining the maximum weighting for those matches. The details of this matching scheme are presented in section 3.2.3. A table of all symbols used in this section is provided in appendix A.

3.2.1 Creating $\hat{E}$

For a correct matching of p_x to t_y, then e_{xy} must be an element of $\hat{E}$ and the weighting of the edge, $w_{e_{xy}}$, should be high enough to ensure that e_{xy} is included in the final path determined by the matching scheme. The existence of the edge e_{xy} in the set $\hat{E}$ is determined solely by the constraints of the physical world. For the following three sections, apart from the thresholds set for comparing histograms, all thresholds are determined from observations of pedestrians' 3D physical movements between frames in test sequence data.

To obtain and weight the edges in $\hat{E}$, the following statistics are obtained from each p_x region in frame i;

(1) $p_x^{3d^i}$: the position of the centre of mass of a detected pedestrian's 3D head region orthographically projected onto the groundplane;
(2) $p_x^{max^i}$ and $p_x^{min^i}$: the maximum and minimum heights above the groundplane of all the 3D points belonging to the pedestrian that are *visible* in frame i and within the required VOI defined in section 3;
(3) $p_x^{c^i}$: the set of HSV colour values of all foreground points belonging to the pedestrian.

Similarly, all t_y have similar statistics in frame $i-1$; $t_y^{3d^{i-1}}$, $t_y^{max^{i-1}}$, $t_y^{min^{i-1}}$ and $t_y^{c^{i-1}}$. In addition, each t_y has three additional statistics;

(1) $t_y^{n^{i-1}}$: the number of frames for which the track has existed;
(2) $t_y^{v^{i-1}}$: the velocity of the track in the previous frame, where $t_y^{v^{i-1}} = \left| t_y^{3d^{i-1}} - t_y^{3d^{i-2}} \right| \times \frac{1}{td_{i-2}^{i-1}}$ and td_{i-2}^{i-1} is the time difference (in milliseconds) between frames $i-1$ and $i-2$;
(3) $t_y^{3d^i}$: the extrapolated position of the track in the current frame, where if $t_y^{n^{i-1}} < 2$ then $t_y^{3d^i} = t_y^{3d^{i-1}}$, otherwise $t_y^{3d^i} = t_y^{3d^{i-1}} + (t_y^{v^{i-1}} \times td_{i-1}^{i})$;
(4) $t_y^{s^{i-1}}$: the track state, which is either *walking*, St^w, *accelerating*, St^a, or *standing*, St^s. A person is considered to be walking if they have either; (a) moved in the same direction for 3 consecutive frames, (i.e. the angles between $t_y^{3d^{i-4}}$,$t_y^{3d^{i-3}}$,$t_y^{3d^{i-2}}$ and $t_y^{3d^{i-1}}$ are greater than 90° in each case), or; (b) moved in the same direction for 2 consecutive frames and $Edist(t_y^{3d^{i-1}}, t_y^{3d^{i-3}}) > t_{noise}$, where $Edist$ is Euclidean distance and t_{noise} is the maximum distance a track's t_y^{3d} is allowed to fluctuate in one frame when they are standing still. In our experiments t_{noise} is set to 0.3 metres, meaning it is expected that a pedestrian's position will fluctuate by up to 0.15 metres from its correct position in any given frame. A person is deemed to be accelerating if $Edist(t_y^{3d^{i-1}}, t_y^{3d^{i-2}}) > t_{noise}$. Finally, a person is standing still if neither of the other states are possible.

To determine whether $e_{xy} \in \hat{E}$, p_x is compared to t_y and an evaluation is made

whether the match is physically plausible. For example, if the time difference, td^i_{i-1}, between frames $i-1$ and i is one second, and $Edist(t_y^{3d^{i-1}}, p_y^{3d^i}) = 20$, where $Edist$ is Euclidean distance in metres, the edge e_{xy} is *not* plausible, as the pedestrian p_x would have to moving at a rate of 72km/h! Therefore, if $Edist(t_y^{3d^{i-1}}, p_y^{3d^i}) > t_{max}$ then $e_{xy} \notin \hat{E}$, where $t_{max} = dist_{max} \times td^i_{i-1}$ and $dist_{max}$ is the *absolute maximum* distance a pedestrian is assumed to be able to be the walk in a second. In this work, a threshold $t_{avge} = dist_{avge} \times td^i_{i-1}$ is also applied, where $dist_{avge}$ is the *average maximum* distance a pedestrian is assumed to walk in a second. In our experiments, $dist_{max}$ and $dist_{avge}$ are set to 3 and 2 metres per second respectively (however it should be noted that the minimum value of t_{max} or t_{avge} should be set to t_{noise} regardless of the value of td^i_{i-1}). This limitation of possible pedestrian movement, where $Edist(t_y^{3d^{i-1}}, p_y^{3d^i}) > t_{max}$ then $e_{xy} \notin \hat{E}$, defines the first kinematic constraint that form a contribution of this work. As a set, these constraints can be used to remove implausible matches of pedestrians to previous tracks.

A second physical constraint is based on a pedestrian's ability to turn while *walking* at a sufficiently large velocity. It is assumed that due to the forward momentum incurred, a pedestrian can only turn a certain angle θ in a single frame. This constraint can be formulated as follows. Let $\vec{a}$ be the vector from $t_y^{3d^{i-1}}$ to $t_y^{3d^i}$ and $\vec{b}$ be the vector from $t_y^{3d^{i-1}}$ to $p_x^{3d^i}$, and let $\theta = cos^{-1}\left[\frac{\vec{a}.\vec{b}}{|\vec{a}||\vec{b}|}\right]$ be the angle between $\vec{a}$ and $\vec{b}$ (obtained using the dot product). From the statistics of a track it can be assumed that a pedestrian is then moving at a high enough velocity if $Edist(t_y^{3d^{i-1}}, p_y^{3d^i}) > t_{noise}$ and either; (a) a person is accelerating very quickly ($t_y^{s^{i-1}} = St^a$ and either $t_y^{v^{i-1}}$ or $t_y^{v^{i-1}}$ is greater than t_{avge}), or; (b) a person is walking and the velocity in the previous frame was greater than that to be expected of a position change due to noise if the pedestrian has suddenly stopped walking ($t_y^{s^{i-1}} = St^w$ and $t_y^{v^i} > \frac{t_{noise}}{2}$). If either of these cases are true then the pedestrian may either stop *or* continue on in roughly the same direction in frame i, i.e. $Edist(t_y^{3d^{i-1}}, p_y^{3d^i}) \leq t_{noise}$ *or* $\theta \leq \theta_{max}$, where θ_{max} is the maximum angle that a walking pedestrian can turn per frame. θ_{max} is set to 60° in all our experiments. This is the second kinematic constraint in this work used to remove implausible matches of pedestrians to previous tracks. However, it should be noted that as the time difference, td, between frames increases this constraint is invalidated. In this work it is assumed that the latency between frames is less than the time taken for a pedestrian to stop walking forward and make a turn greater than θ_{max}.

3.2.2 Weighting $\hat{E}$

In order to obtain the correct matching of p_x to t_y, a weighting, $w_{e_{xy}}$, must be associated with e_{xy}. This value should be high enough to ensure that e_{xy} is included in the final path determined by the matching scheme. As illustrated

in figure 4, the initial weighting of e_{xy} is assigned by a colour histogram comparison measure between p_x and t_y. This value of $w_{e_{xy}}$ is then adjusted by a predetermined amount that forces the weights into five distinct groups of varying importance. This novel weighting scheme for matching pedestrians to previous tracks forms a contribution to this work.

In order to obtain the initial value of $w_{e_{xy}}$, a normalised histogram for the pedestrian, $p_x^{h^i}$, is created using the hue value from the HSV colour values in $p_x^{c^i}$ using 3D points that lie only in the *overlapping* height region between $p_x^{max^i}$ to $p_x^{min^i}$ and $t_y^{max^{i-1}}$ to $t_y^{min^{i-1}}$. A similar histogram, $t_y^{h^{i-1}}$, is created for the track. $w_{e_{xy}}$ is determined by obtaining the Bhattacharyya distance [30] between the corresponding $p_x^{h^i}$ and $t_y^{h^i}$. Thus, the value of $w_{e_{xy}}$ will lie between 0 and 1. Finally, if $w_{e_{xy}} < 0.5$ then e_{xy} is removed from $\hat{E}$ as the colour match is deemed to be too weak for a true match between a p_x and t_y to exist.

This weighting, $w_{e_{xy}}$, is then altered to force e_{xy} into one of five distinct weighting groups, whereby the *higher* the value of e_{xy}, the *greater* the importance of that weight, and therefore the *greater* the probability of it being chosen for the final matching. These groupings exist in order to reward good matches, established tracks and penalise more implausible, but not impossible, matches. As illustrated in figure 4:

- In *Group 1*, the weight is actually decreased by 50% of the original value. This decrease is made in order to discourage plausible but unlikely matches. e_{xy} is part of this group if a person is walking or accelerating ($t_y^{s^{i-1}} = St^w$ or St^a) and either; (a) $Edist(t_y^{3d^{i-1}}, p_y^{3d^i}) > t_{avge}$, this discourages the system from attempting to make large jumps in distance, as they rarely occur; (b) $Edist(t_y^{3d^{i-1}}, p_y^{3d^i}) > t_{noise}$ and $\theta > \frac{\theta_{max}}{2}$, as if a person is accelerating or walking then these changes in direction are unlikely to occur; or (c) if a pedestrian has walked the same direction for 3 or more consecutive frames and $\theta > \theta_{max}$, even if $Edist(t_y^{3d^{i-1}}, p_y^{3d^i}) < t_{noise}$, as the previous history of the track indicates that the angle of the track should be continuous even with respect to noise.
- In *Group 2*, the weighting remains as it is – this is the default group.
- *Group 3* rewards a good match in coverage between p_x to t_y in other areas, besides histograms. So if the overlapping height regions are large (i.e. $\geq 50\%$ overlap), then $w_{e_{xy}}$ is incremented by 0.5. Note that e_{xy} may be a member of this group *and* groups 4 or 5 at the same time – if this is the case, then *both* group increments are added to $w_{e_{xy}}$.
- In *Group 4*, a weighting increment is added to $w_{e_{xy}}$ to ensure that older, more established, tracks have priority to be matched with pedestrians. This ensures that established tracks are not left without a match, while new tracks, which may have been initialised due to noise, have been given a match. This is achieved by determining if $t_y^{3d^i}$ is close to $p_x^{3d^i}$ within a more

constrained set of thresholds of angle and distance (simply half the previous thresholds). If this is true and $t_y^{n^{i-1}} = 2$ then $w_{e_{xy}}$ is incremented by one. As previously outlined, if e_{xy} is also a member of group 3 then the total increment between the two groups will be 1.5.

- In *Group 5*, a similar increment to that in group 4 is added to $w_{e_{xy}}$, but if $t_y^{n^{i-1}} > 2$, the weighting is increased by two (leading to a total increment of 2.5 if e_{xy} is also a member of group 3). In this way, tracks that have existed for 3 or more frames have priority over those that have existed for 2 frames, and tracks that have existed for 2 or more frames have priority over those that have existed for only 1 frame.

3.2.3 Maximum Weighted Maximum Cardinality Matching Scheme

After $\hat{E}$ has been created and weighted, the matching algorithm is invoked. The matching scheme technique applied in this work – illustrated in figure 5(a) – is based on Berge's Theorem [31], which states that a matching S in G is maximum *iff* there is no augmenting path, P. In graph theory, a *path* is the list of vertices of a graph where each vertex has an edge from it to the next vertex and an *augmenting* path is one with alternating free and matched edges that begins and ends with free vertices. If such a path is discovered then the cardinality of the matching S can be immediately increased by one, simply by switching the edge membership along P. As such, the proposed matching scheme algorithm is initialised with an empty set of matches and then solves the problem by iteratively searching for the augmenting path [29] with the maximum weight. If an augmenting path is found then the edge membership along P is switched. If no augmenting path is found then M is guaranteed to have maximum cardinality with maximum weight, and by traversing through the path the matches of pedestrians to tracks are obtained. This algorithm is a classical solution to the N-to-M association problem using bipartite graphs.

Within the pedestrian tracking module of this work, an alteration to this N-to-M matching scheme algorithm is made that enforces the physical constraints of real-world pedestrian tracking to be taken into account within the matching framework. When creating $\hat{E}$ (see section 3.2.1) two kinematic constraints are enforced, which ensures that all single edges $e_{xy} \in \hat{E}$ are physically plausible, however these constraints do not ensure that *pairs* of edges are physically plausible. Take for example figure 5(b), where t_1 is a track traversing the scene from left to right, and t_2 is a second track that is travelling parallel to t_1. In frame i, t_1 or t_2 can be matched to either p_1 or p_2, as each match is physically plausible. However, if t_1 is matched to p_2 and t_2 to p_1, then t_1 and t_2 must pass through, or *crossover*, the same physical space between frames $i-1$ and i. Depending on the time difference between the two frames this may be physically impossible, as is the case in our experiments, where the latency between frames difference is typically less than half a second. As such, pairs of

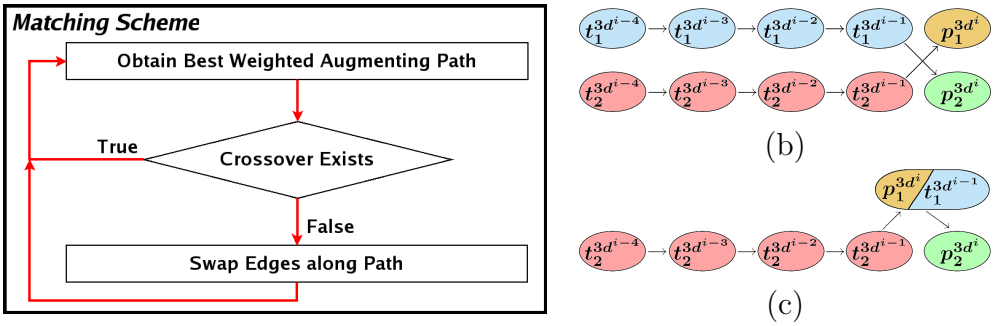


Fig. 5. (a) Matching scheme; (b) Crossover; (c) Near crossover.

edges of this type are not allowed to coexist in a legitimate matching. As such, a constraint is imposed that in a legitimate matching, no two *physical* track segments between frames i and $i-1$ may be within a distance of 10 cm of each other. This eliminates all crossovers, and near crossovers, such as that in figure 5(b), where $t_1^n = 1$ and $p_1^{3d^i} \approx t_1^{3d^{i-1}}$, where although no actual crossover has occurred, the two track segments must again occupy the same physical space at some stage between i and $i-1$. This limitation of possible pedestrian movement defines the third, and final, kinematic constraint applied.

3.2.4 Pedestrian Detection Rollback Loops

After pedestrians have been assigned to tracks, rollback loops are used to backtrack the pedestrian detection module in an attempt to find or extrapolate lost tracks, and to further reduce over- and under-segmentation. The rollback scheme, which forms a contribution to this work, employs three separate loops; the first two aim to locate all lost tracks using two different techniques (the second of which also reduces under-segmentation); the third is designed to reduce over-segmentation. Each rollback loop is now be examined in turn.

If t_y is unmatched then the first rollback loop attempts to find the missing pedestrian, p_x, in the current frame. The post-processing stage of the pedestrian detection module (see section 3) declares a region as noise if it falls below certain thresholds, such as the minimum number of pixels or if it does not span a pre-defined range of heights. However, scenarios such as severe occlusion could force p_x below these thresholds. To retain this lost region, the tracking module backtracks the pedestrian detection module to just before post-processing occurs and reapplies post-processing at half the original thresholds. If any *new* regions emerge, which may be a feasible continuation of the lost track, then the weighted bipartite matching scheme is reiterated. If t_y becomes matched then that new pedestrian region remains and all other regions added by this module are removed.

If t_y is still unmatched, a second rollback loop is employed that makes the

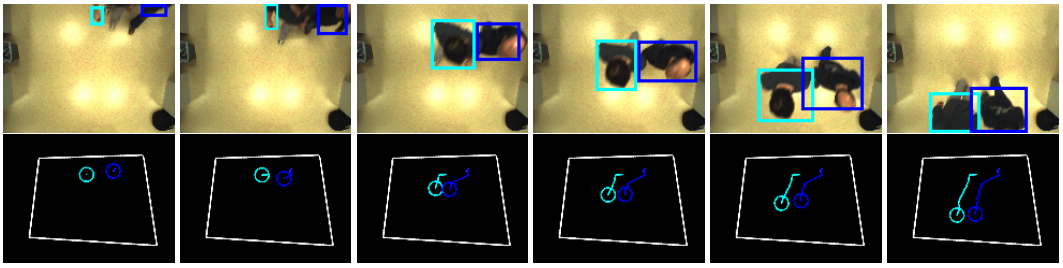
assumption that under-segmentation of pedestrians occurred resulting in two tracks, t_1 and t_2, competing for the same region, p_1. The rollback loop backtracks the pedestrian module to before the final iteration of region clustering, which is then skipped and the regions are post-processed. If p_1 has been segmented into 2 distinct regions, p_1a and p_1b, where the orientation of $t_1^{3d^i}$ to $t_2^{3d^i}$ is similar to that of $p_1a^{3d^i}$ to $p_1b^{3d^i}$ then a possible match may exist. In this approach, the maximum difference in orientation is set at $\pm 22.5°$, therefore allowing a total range in orientation difference of 45°. It is believed that this value allows enough variation in orientation, while simultaneously avoiding the case of incorrect matches from the rollback loop. As in the first rollback loop, if the re-segmentation is successful the weighted bipartite matching scheme is reiterated. However, if the re-segmentation is not possible but an attempt was made, i.e. p_1 exists whereby it can be matched to either t_1 and t_2 but it could not be segmented into two regions, then it is assumed that p_1 actually contains 2 pedestrians, and the unmatched track is extrapolated.

The third, and final, rollback loop is designed to reduce over-segmentation by examining all *unmatched* pedestrians. The pedestrian detection module is backtracked to before the final stage of merging regions and the under-segmentation test is turned off. The final clustering stage and post-processing is re-iterated and it is determined whether the unmatched pedestrian region has become merged with a second region. If it does, then the region is considered to be over-segmented and two regions remain merged.

3.2.5 Track Post-processing

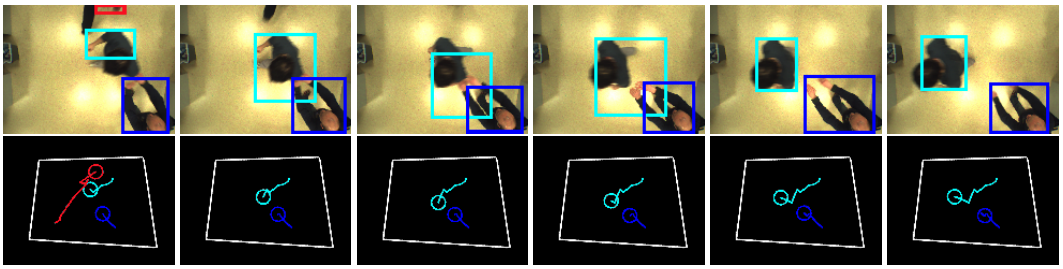
The final stage (and final contribution) of the tracking framework is designed to post-process tracks with a view to increasing track stability with respect to pedestrian over-segmentation problems. If the tracked pedestrian t_1 is over-segmented in frame i as p_1a and p_1b, then a choice has to be made whether to match p_1a or p_1b. Let t_1 choose p_1a and let t_1b be the new track initiated by p_1b. Each separate choice will affect t_1's statistics in frame $i+1$, meaning that a bad choice could end the track prematurely, however the new track from the choice not taken may still exist. If this is the case, then the wrong choice was made. This type of occurrence can be rectified by flagging possible over-segmentations and the resultant choice in frame i. Then if t_1 is discontinued before t_1b *and* the two separate tracks have not diverged *or* t_1b has not demonstrated that it is a stable track by being able to reach a walking state, t_1 is allowed to "steal" the track of t_1b. If this scenario occurs, the history of t_1 is replaced by that of t_1b for the duration of t_1b's lifespan.

Finally, all unmatched tracks that have not been explicitly extrapolated in the second rollback loop, are removed and all unmatched pedestrians are assigned new tracks. In addition, every matched track is updated to incorporate the



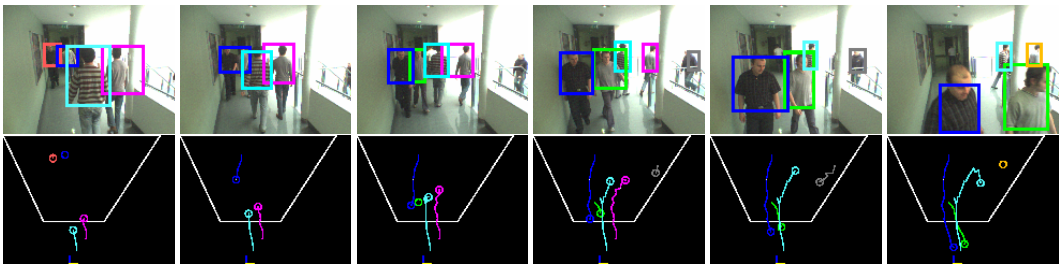
(a) (b) (c) (d) (e) (f)

Fig. 6. *Overhead* sequence, frame numbers; (a) 342; (b) 344; (c) 347; (d) 349; (e) 351; (f) 355.



(a) (b) (c) (d) (e) (f)

Fig. 7. *Overhead* sequence, frame numbers; (a) 186; (b) 187; (c) 188; (d) 189; (e) 190; (f) 191.



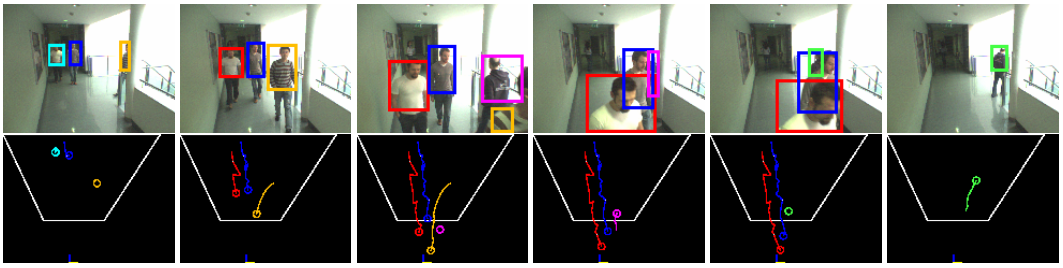
(a) (b) (c) (d) (e) (f)

Fig. 8. *Corridor* sequence, frame numbers; (a) 289; (b) 293; (c) 298; (d) 301; (e) 304; (f) 308.

new data from frame i.

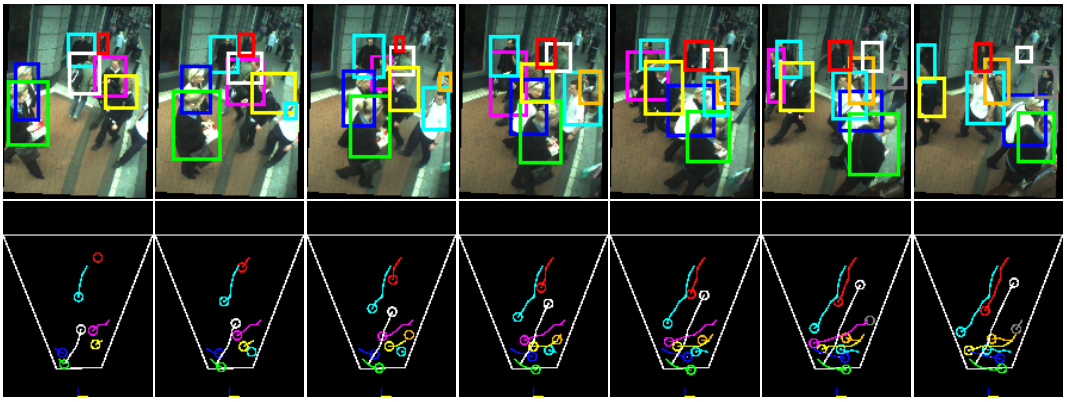
4 Experimental Results

The proposed technique has been quantitatively evaluated against 5 test sequences of resolution 640×480 captured between 2-6.5Hz. The sequences cover three different scenarios, with varying camera height, camera orientation and environmental conditions. The experimental sequences were chosen to test the proposed technique extensively in several areas, such as disparity estimation, foreground segmentation, pedestrian detection and tracking. None of the test



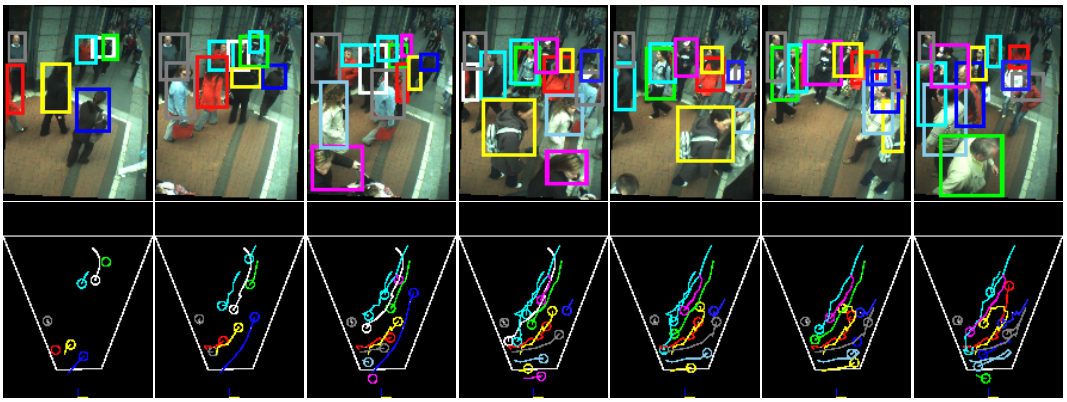
(a) (b) (c) (d) (e) (f)

Fig. 9. *Corridor* sequence, frame numbers; (a) 216; (b) 225; (c) 234; (d) 238; (e) 239; (f) 247.



(a) (b) (c) (d) (e) (f) (g)

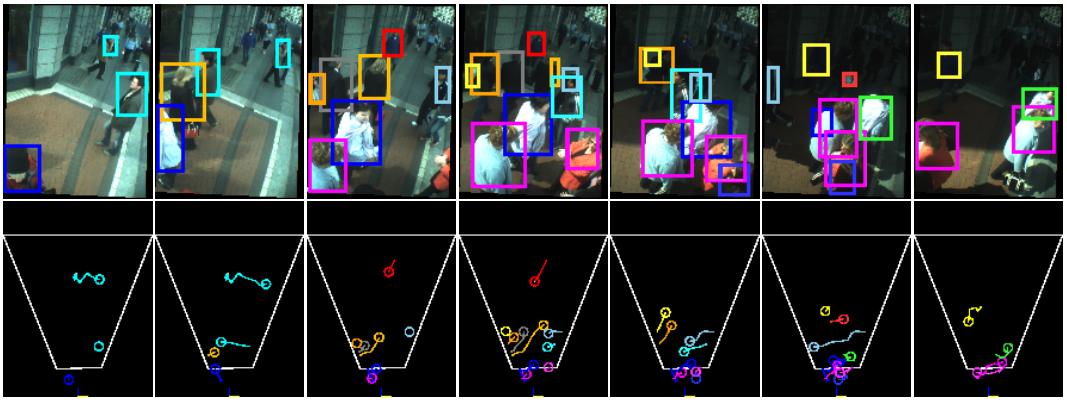
Fig. 10. *Grafton* sequence 1, frame numbers; (a) 009; (b) 010; (c) 011; (d) 012; (e) 013; (f) 014; (g) 015.



(a) (b) (c) (d) (e) (f) (g)

Fig. 11. *Grafton* sequence 2, frame numbers; (a) 017; (b) 021; (c) 024; (d) 027; (e) 029; (f) 031; (g) 035.

sequences were used in development of the proposed algorithms. Figures 6-12 give illustrative examples the sequences, which were specifically chosen to illustrate both the success and possible failings of the proposed approach. In each of these figures there are two rows of images. In the top row, each detected pedestrian is enclosed by a bounding box of a certain colour. Directly



(a) (b) (c) (d) (e) (f) (g)

Fig. 12. *Grafton* sequence 3, frame numbers; (a) 225; (b) 228; (c) 231; (d) 233; (e) 235; (f) 238; (g) 241.

Table 1
Experimental results overview.

Sequence	*Groundtruth*	*Detected*	*Correct*	*Precision*	*Recall*
Grafton 1	666	620	577	93.0	86.6
Grafton 2	754	692	669	96.7	88.7
Grafton 3	457	388	362	93.3	79.2
Grafton Total	1877	1700	1608	94.6	85.7
Overhead Total	657	626	592	94.5	90.1
Corridor Total	1027	822	763	92.8	74.3
Total	3561	3148	2963	94.1	83.2

(a) (b) (c)

Fig. 13. Missed groundtruth persons; (a) *Overhead* sequence; (b) *Corridor* sequence; (c) *Grafton* sequences.

beneath this row are plan-view images corresponding to the scenes in the top row. In these plan-view images, the white lines indicate the bounds of the scene, the position of detected pedestrians in that frame are illustrated by a circle of the same colour as their bounding box, and tracks are depicted as "tails" from the centre of the circle to previous positions in the scene. All of results sequences, bar the *Grafton* sequences for legal reasons, are available to

view on-line at [32].

The first sequence, which will be referred to as the *Overhead* sequence, see figures 6 and 7, was set in an indoor scenario with the camera positioned at around 3 metres above the ground. The camera was then orientated back towards the groundplane. The camera rig in this point of view has a limited field of view and due to its proximity with the groundplane it does not encounter significant occlusion problems. The lighting conditions in the scene are stable, but brightly illuminated with a highly reflective ground surface. The sequence consisted of 418 images, captured at a frame rate of $\approx$ 6.5Hz ($\approx$ 1.1 minutes).

The second sequence, which will be referred to as the *Corridor* sequence, see figures 8 and 9, was set in an indoor setting with the camera positioned just above 2 metres from the ground. The camera is orientated at 30 degrees towards the groundplane. Again, the lighting conditions are stable, however the scene's illumination is more challenging than that of the *Overhead* sequence as it is brightly illuminated on one side, and dark on the other side, due to skylights in the corridor. This can cause a lack of texture in those areas, as will be described later. In addition, the scene contains a staircase on the right hand side, where people can descend and ascend at will. The sequence consists of 697 images, captured at a frame rate of $\approx$ 5.3Hz ($\approx$ 2.3 minutes). For these two sequences volunteers were recruited and asked to walk around in front of the camera. No restrictions or instructions were provided as to where people could go, what they could do or what they could wear.

The third, fourth and fifth sequences, see figures 10, 11 and 12, will be referred to as the *Grafton* sequences. These three separate image sequences were taken from a camera mounted at 2.5 metres above the groundplane with a 45 degree angle on a traffic light pole on a busy pedestrianised shopping street in Dublin city centre. All the sequences contain pedestrians from the general public walking during their daily routine. These sequences were deliberately chosen to contain challenging segments – groups of people walking in multiple directions or standing still and rapidly changing lighting conditions. Altogether, the three sequences consist of 330 images, captured at a frame rate of $\approx$ 2Hz ($\approx$ 3 minutes). The first two *Grafton* sequences exhibit constant illumination conditions that minimise shadows cast and background illumination changes. The illumination conditions in *Grafton* sequence 3, see figure 12, are constantly changing. To illustrate the severity of these conditions, the illumination changes between figure 12(c)-(e) occurs in just under 5 seconds. The *Grafton* sequence 3 has 3 differing lighting conditions in its 60 second duration.

Each of the five sequences were manually groundtruthed by positioning a separate bounding box around each person in the image. In the evaluation process, a person is defined as someone who has a section of their body above

the waist, no matter how small, visible in the image. If all that can be seen of a person in the image is an outstretched hand or a backpack then they *are* counted as being present. However, if just a leg or foot is present, then they are *not* counted. The only other constraint is that a groundtruth was not created for people who are further than 8 metres from the camera. This constraint is necessary in the *Grafton* sequences as pedestrians can be seen for over a hundred metres. Placing bounding boxes around all of these people and evaluating against them would introduce significant of noise into the evaluation process. The distance of 8 metres is chosen as the cutoff point as the proposed system removes all 3d points greater than this distance from the camera as the disparity map quality degrades rapidly after this point. In effect, the values in table 1 are the precision and recall values for people within an 8 metre distance of the camera.

In table 1; the second column, *Groundtruth*, represents the number of people present in the groundtruth data; *Detected*, represents the number of distinct regions the proposed algorithm detected in the sequence and; *Correct*, represents the number of *Detected* regions that correctly overlapped with the *Groundtruth*. A correctly segmented pedestrian is defined as a region that overlaps a groundtruth area by 50% or more. It is acknowledge that this percentage is relatively low, but this work is more interested in detecting pedestrians than detecting the correct number of pixels corresponding to a person. As such, this percentage threshold was chosen. In table 1, *Precision* and *Recall* values are also given, where *Precision* is the percentage of *Correct* with respect to *Detected*, and *Recall* is the percentage of *Correct* with respect to *Groundtruth*. Analysis of where in the image sequences the proposed technique failed to correctly detect pedestrians can be simplified by obtaining the centroids of the bounding boxes of all the groundtruthed people that did not have a match in the data, see figures 13(a),(b) and (c), where the centroids are depicted as red dots. In these figures, these points are overlaid onto a sequence image where there is no foreground activity to provide a visual cue to "problem" areas in the image sequences. The results of each test sequence will now be discussed.

The robustness of the proposed technique to cope with two people in close proximity whilst being able to avoid over-segmentation is illustrated in the *Overhead* sequences of figures 6 and 7. The tracks in both sequences are coherent and are not lost, even on close interaction. The recall of this sequence was the highest of all tested. A factor in this was the close proximity of the people to the camera resulting, in general, in good disparity estimation. Analysis of figure 13(a) reveals that all but two of the pedestrians missed are positioned around the boundaries of the scene, at the points where people enter and exit the scene. This is not surprising for two reasons; (1) the disparity is less likely to be well formed around the edges of the image and; (2) when a person enters the scene, the first portion of their body that enters the scene is likely to be a hand or their lower torso, followed shortly by their head and shoulders. There-

fore, when entering and exiting the scene the regions observed by the camera are lower to the ground and clustering of regions with the golden ratio will result in a lower absolute value of δ, which controls the maximum clustering distance in the pedestrian detection module. This means that large foreground regions will not be created until the shoulders and head enter the scene, and may result in the region being removed by the pedestrian detection module's post-processing steps.

Figures 8 and 9 illustrate some issues with the proposed tracking technique. As the two people detected in figure 8(a) approach the camera, they squeeze together and merge as one in figure 8(b) for 5 frames and one of the tracks is lost at this point. The techniques does not have any explicit full-occlusion handling, so the track of a person is lost in figure 8(e) and a new one is started for the same person 4 frames later when they emerge from behind a pillar.

In figure 9, a track is lost again due to a large, but not full occlusion. In figure 9(c), the person on the right (surrounded by a pink bounding box) walks away from the camera, but in figure 9(d), the left hand side of their body is fully occluded. This artificially forces their centre of gravity to the right fooling the system into believing the person to be turning left. The track is then lost in the next frame as this manoeuvre is considered impossible by the tracking system. Whilst the person interactions in this *Corridor* sequence are not very challenging, the sequence is interesting in terms of the distance at which these occur (greater than that of any of the other sequences) as well as the lighting conditions. The right hand of the scene is very bright whilst the left is much darker. Therefore, if people wearing brightly coloured clothes are on the right there is little texture information and vice versa on the left. This lack of texture has a degrading affect on the quality of the disparity and the pedestrian detection post-processing, so 3D regions in these areas are clustered less effectively and are more likely to be removed in post-processing. These issues are not unique to the proposed technique as other techniques that rely on disparity, foreground segmentation or edge gradients would be similarly affected. People on the stairs tend to be missed as the regions here are closer to the groundplane and will therefore, as in the *Overhead* sequences, be removed by the pedestrian detection module's post-processing steps. The missed groundtruth pedestrians depicted in figure 13(b) confirms this. Other missed groundtruths tend to congregate around the 8 metre mark as expected.

Finally, the *Grafton* sequences depicted in figures 10-12 illustrate the robustness of the detection and tracking techniques when subjected to unconstrained crowded conditions. They all depict multiple pedestrians travelling in various directions being tracked robustly. In these sequences, up to 13 people are successfully tracked concurrently. For example, in figures 11(e)-(g) a pedestrian (surrounded by a yellow bounding box) makes a u-turn in the sequence and is successfully tracked. A bad track does occur on the right between frames

(d) and (e), when one pedestrian (surrounded by a grey bounding box) leaves the scene in (e) and another person enters the scene at similar location in the same frame. Figure 12 demonstrates the same issues that were present in the *Corridor* sequences, whereby although shadows do not cause a problem to system precision (notice how the precision in table 1 remains stable regardless of the lighting conditions), the recall is affected by the strong shadows caused by buildings. These shadows result in a lack of texture, and therefore cause tracks that, up to that point, have been stable to become lost. Figure 13(c) backs up the observations made for other sequences, whereby the vast majority of pedestrians missed tend to congregate either around the 8 metre mark or at the scene boundaries.

5 Conclusions and Future Work

In this work a technique for the robust detection and tracking of humans in crowded scenes was presented. The approach is sufficiently generic to be applicable to many different camera placement and orientation scenarios. It is acknowledged that there are some outstanding questions with the groundtruth and evaluation process in this work, such as:

- When is a person exactly 8 metres from the camera? For 2D evaluation, an imaginary bounding line is drawn across the groundplane based on measurements taken from the scene, but this is not ideal.
- How accurate are the 3D statistics obtained from each pedestrian, such as height, velocity and 3D position?
- When is a person "in the scene"? The evaluation process was implemented on the right camera image. Due to the offset of the left stereo image, people to the far right of the right image who are groundtruthed may not appear at in the left image.
- How accurate is the system for a maximum distance of 5,6,7 or 8 metres? Does the system's performance degrade gradually or is there a threshold distance after which there a large drop off in performance?

To help answer some of these questions, in future work it is planned to groundtruth the system against a 3D Vicon infrared motion analysis system [33]. In this work, objects other than pedestrians are not detected, such as push prams, buggies or bicycles. In fact, the system removes all points under 0.9 meters in height above the groundplane. Varying this feature should be investigated to determine the ideal threshold for a given application scenario, in order to detect different kinds of pedestrians. Future work may also include the fusion of more than one technique for pedestrian detection and tracking, based on estimating the distance of the pedestrian from the camera and switching to another approach, such as appearance-based tracking [6,7], when

appropriate. In addition, the robustness of the technique should be evaluated against a higher image capture frame rate. Using this higher frame rate would also allow the introduction of a fourth pedestrian state, whereby a pedestrian can be standing, accelerating, walking or running.

Finally, techniques to reduce the computational complexity of the proposed algorithms should be investigated. In our experiments, the proposed system was implemented in un-optimised C++, designed in a highly object-oriented framework, and run on a 2GHz laptop. In general, the overall processing time for each frame varies – the more pedestrians within the frame and the more foreground disparity points need to be clustered. This leads to longer processing times. On average the processing of a single 640×480 pixel frame takes between 10–20 seconds. Obviously this is far from real-time processing. However, throughout the system development, the algorithmic design took precedence over complexity, which was rarely addressed. Apart from optimising code, a number of research paths exist that would maintain the main algorithmic features, but decrease complexity.

Acknowledgements

This material is based on works supported by Science Foundation Ireland under Grant No. 03/IN.3/I361.

A Appendix

The height above the groundplane can be used to define the proportions of a human body by applying the golden ratio Φ, ($\Phi = \sqrt{5} * 0.5 + 0.5 \simeq 1.618$) [27]. Figure A.1(a) shows how a body is segmented using Φ. Let |aj| be the Euclidean distance between the horizontal lines a and j. Therefore, |aj| is the height of a human body. Using Φ and |aj| various other points on the human body can be defined. In figures A.1(a) and (b); $|\text{ai}| = \frac{|\text{aj}|}{\Phi}$, $|\text{ah}| = \frac{|\text{ai}|}{\Phi}$...$|\text{ab}| = \frac{|\text{ac}|}{\Phi}$ [27] and |mn| is equivalent to |ae|. Similarly $|\text{lo}| \equiv |\text{ag}|$ and $|\text{kp}| \equiv |\text{ah}|$. Distances of interest are outlined in table A.1. Various parameters and employed notation for pedestrian detection and tracking are collected in tables A.2 and A.3 respectively.

Table A.1
Biometric distances overview.

Distance	Meaning
|aj|	the height of the human body
|ac|	the distance from the head to the forehead
|ad|	the distance from the head to the eyes
|mn|	the width of the head
|af|	the distance from the head to the base of the skull
|lo|	the width of the shoulders
|ah|	the distance from the head to the navel and the elbows

Table A.2
Pedestrian detection symbols overview.

Symbol	Meaning
Φ	the golden ratio $\Phi = \sqrt{5} * 0.5 + 0.5 \simeq 1.618$
δ	controls the rate of growth in the clustering process
l	the 2D line that passes through centre of the two region'd best fit ellipses
γ	the maximum Euclidean distance between two region ellipse points on l
reg_{cx}	the central axes of the region, which is the 3D line that is parallel to the 3D groundplane normal and runs through the average 3D point in the region
d_{cx}^{12}	the Euclidean distance, from reg_{cx}^{1} and reg_{cx}^{2}

Fig. A.1. Golden ratio; (a) Vertical; (b) Horizontal.

References

[1] P. Remagnino, G. Foresti, Ambient intelligence: A new multidisciplinary paradigm, in: IEEE Transactions on Systems, Man and Cybernetics, Vol. 35, 2005, pp. 1–6.

[2] M. Vallée, F. Ramparany, L. Vercouter, A multi-agent system for dynamic service composition in ambient intelligence environments, in: International

Conference on Pervasive Computing, 2005, pp. 157–182.

[3] P. Kelly, N. O'Connor, A. Smeaton, Pedestrian detection in uncontrolled environments using stereo and biometric information, in: ACM International Workshop on Video Surveillance and Sensor Networks, 2006, pp. 161–170.

[4] B. Leibe, E. Seemann, B. Schiele, Pedestrian detection in crowded scenes, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, 2005, pp. 878–885.

[5] M. Harville, Stereo person tracking with adaptive plan-view templates of height and occupancy statistics, International Journal of Computer Vision 22 (2004) 127–142.

[6] A. Senior, Tracking with probabilistic appearance models, in: ECCV workshop on Performance Evaluation of Tracking and Surveillance Systems, 2002, pp. 48–55.

[7] A. Elgammal, L. Davis, Probabilistic framework for segmenting people under occlusion, in: IEEE International Conference on Computer Vision, Vol. 2, 2001, pp. 145–152.

[8] T. Zhao, R. Nevatia, Bayesian human segmentation in crowded situations, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, 2003, pp. 459–466.

[9] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, S. Maybank, Principal axis-based correspondence between multiple cameras for people tracking, in: Pattern Analysis and Machine Intelligence, Vol. 28, 2006, pp. 663–671.

[10] J. Rittscher, P. Tu, N. Krahnstoever, Simultaneous estimation of segmentation and shape, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, 2005, pp. 486–493.

[11] C.-J. Pai, H.-R. Tyan, Y.-M. Liang, H.-Y. Liao, S.-W. Chen, Pedestrian detection and tracking at crossroads, in: International Conference on Image Processing, Vol. 2, 2003, pp. 101–104.

[12] B. Heisele, C. Wöhler, Motion-based recognition of pedestrians, in: International Conference on Pattern Recognition, Vol. 2, 1998, pp. 1325–1330.

[13] R. Cutler, L. Davis, Real-time periodic motion detection, analysis, and applications, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, 2000, pp. 781–796.

[14] S. Niyogi, E. Adelson, Analyzing and recognizing walking figures in xyt, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1994, pp. 469–474.

[15] T. Darrell, G. Gordon, M. Harville, J. Woodfill, Integrated person tracking using stereo, color, and pattern detection, International Journal of Computer Vision 37 (2000) 175–185.

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

[17] L. Zhao, C. Thorpe, Recursive context reasoning for human detection and parts identification, in: IEEE Workshop on Human Modeling, Analysis, and Synthesis, 2000.

[18] I. Haritaoglu, D. Harwood, L. Davis, w^4s: A real time system for detecting and tracking people in 2.5 d, in: European Conference on Computer Vision, 1998, pp. 877–892.

[19] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, S. Shafer, Multi-camera multi-person tracking for easyliving, in: IEEE Workshop on Visual Surveillance, 2000, pp. 3–10.

[20] D. Beymer, Person counting using stereo, in: Workshop on Human Motion, 2000, pp. 127–131.

[21] M. Harville, Stereo person tracking with adaptive plan-view statistical templates, in: Workshop on Statistical Methods in Video Processing, 2002, pp. 67–72.

[22] K. Hayashi, M. Hashimoto, K. Sumi, K. Sasakawa, Multiple-person tracker with a fixed slanting stereo camera, in: IEEE International Conference on Automatic Face and Gesture Recognition, 2004, pp. 681–686.

[23] M. Harville, L. Dalong, Fast, integrated person tracking and activity recognition with plan-view templates from a single stereo camera, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2002, pp. 398–405.

[24] R. Munoz-Salinas, E. Aguirre, M. Garcia-Silvente, A. Gonzalez, People detection and tracking through stereo vision for human-robot interaction, in: Lectures Notes on Artificial Intelligence, 2005, pp. 337–346.

[25] P. Kelly, E. Cooke, N. O'Connor, A. Smeaton, Pedestrian detection using stereo and biometric information, in: International Conference on Image Analysis and Recognition, 2006, pp. 802–813.

[26] X. Liu, P. Tu, J. Rittscher, A. Perera, N. Krahnstoever, Detecting and counting people in surveillance applications, in: IEEE Conference on Advanced Video and Signal Based Surveillance, 2005, pp. 306–311.

[27] Golden number.
URL `http://www.goldennumber.net`

[28] K. Sobottka, I. Pitas, Extraction of facial regions and features using color and shape information, in: International Conference on Pattern Recognition, Vol. 3, 1996, pp. 421–425.

[29] Z. Galil, Efficient algorithms for finding maximum matching in graphs, ACM Computing Surveys 18 (1986) 23–38.

[30] T. Kailath, The divergence and bhattacharyya distance measures in signal selection, in: IEEE Transactions on Communication Technology, Vol. 15, 1967, pp. 52–60.

[31] C. Berge, Two theorems in graph theory, Proceedings of the National Academy of Sciences of the United States of America 43 (9) (1957) 842–844.

[32] Test sequences.
URL `http://www.eeng.dcu.ie/~kellyp`

[33] Vicon 3d system.
URL `http://www.vicon.com`

Table A.3
Pedestrian tracking symbols overview.

Symbol	Meaning
G	the weighted bipartite graph $G = (V, E)$
E	the weighted bipartite graph G's edges, each edge is a match from a pedestrian x to a track y
$\hat{E}$	a subset of the weighted bipartite graph G's edges
e_{xy}	$e_{xy} \in E$ and *possibly* an element of $\hat{E}$, it is a match from a pedestrian x to a track y
$w_{e_{xy}}$	the weighting associated with e_{xy}
p_x	pedestrian number x in frame i
$p_x^{3d^i}$	the position of the centre of mass of a detected pedestrian's 3D head region orthographically projected onto the groundplane in frame i
$p_x^{max^i}$	the maximum height above the groundplane of the pedestrian in frame i
$p_x^{min^i}$	the minimum height above the groundplane of the pedestrian in frame i
t_y	track number y in frame $i-1$
$t_y^{c^{i-1}}$	the set of HSV colour values of all foreground points belonging to the pedestrian in frame $i-1$
$t_y^{3d^{i-1}}$	the position of the centre of mass of a tracked pedestrian's 3D head region orthographically projected onto the groundplane in frame $i-1$
$t_y^{max^{i-1}}$	the maximum height above the groundplane of the pedestrian in frame $i-1$
$t_y^{min^{i-1}}$	the minimum height above the groundplane of the pedestrian in frame $i-1$
$t_y^{c^{i-1}}$	the set of HSV colour values of all foreground points belonging to the pedestrian in frame $i-1$
$t_y^{n^{i-1}}$	the number of frames for which the track has existed
$t_y^{v^{i-1}}$	the velocity of the track in frame $i-1$
$t_y^{3d^i}$	the extrapolated position of the track in frame i
$t_y^{s^{i-1}}$	the track state, which is either *walking*, St^w, *accelerating*, St^a, or *standing*, St^s
$Edist$	Euclidean distance
td_{i-1}^i	the time difference between frames i and $i-1$