3D IMAGE ANALYSIS FOR PEDESTRIAN DETECTION

Philip Kelly ⋆**, Eddie Cooke, Noel O'Connor, Alan Smeaton**

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

Abstract *A method for solving the dense disparity stereo correspondence problem is presented in this paper. This technique is designed specifically for pedestrian detection type applications. A new Ground Control Points (GCPs) scheme is introduced, using groundplane homography information to determine regions in which good GCPs are likely to occur. The method also introduces a dynamic disparity limit constraint to further improve GCP selection and dense disparity generation. The technique is applied to a real world pedestrian detection scenario with a background modeling system based on disparity and edges.*

1 Introduction

Many computer vision based applications depend on accurate detection and segmentation of foreground objects as a first step in their algorithmic process. Traditional approaches often fail in unconstrained real-world environments due to dynamic background conditions such as moving backgrounds, changing lighting conditions and shadows, and the variability in a foreground objects local and global appearance. The use of stereo information has been proposed as a means to guide object segmentation, due to it having some distinct advantages over conventional 2D techniques [1].

Many stereo correspondence techniques have been proposed in literature, a taxonomy of many such techniques can be found in [2]. This is a difficult process, however, especially in areas of homogeneous colour or occlusion. To improve results, many algorithms make use of one or more constraints, such as the *disparity limit constraint*, which limits the region where a match for a pixel in one image can occur in a second image. This reduces ambiguities as well as decreasing computational expense.

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

A second technique to improve results from stereo correspondence algorithms is to use highly reliable matched pixels, known as Ground Control Points (GCPs) [3], to help guide results.

In this paper we define a dynamic programming based stereo correspondence technique, using GCPs and *dynamic* disparity limit constraints, that has been developed specifically for pedestrian surveillance type applications. The technique for obtaining GCPs is a 3 stage process; (1) using groundplane homography information, regions are determined in which good GCPs are likely to be found; (2) the best GCP disparities are selected based on the region value built up from neighbouring pixels; (3) background GCPs are found using background disparity and edge models. In addition, we introduce a technique for obtaining a *dynamic* disparity limit constraint to further improve GCP selection and dense disparity generation, in addition to reducing algorithmic complexity.

This paper is organized as follows: Section 2 presents the details of the developed algorithmic approach. Firstly, an overview of homographic transformations is introduced; we then illustrate how GCPs are obtained and dense disparity is generated. In Section 3 we present experimental results from a real world outdoor situation. Finally, Section 4 details conclusions and future work.

2 Algorithm Details

2.1 Groundplane Space

For convenience, we assume that the two input images are rectified via [4]. Rectification aligns the images vertically, so that epipolar lines are parallel. A *second* preprocessing step is then applied that aligns the images horizontally with respect to the groundplane. This technique is based on the application of a groundplane homography [5]. The homography maps the groundplane in one image, I_1, to the groundplane in a second image, I_2, using the equation

$$\mathbf{x} \cong H_{12}\mathbf{x}' \tag{1}$$

where $\cong$ denotes equality up to a scale factor, $\mathbf{x}$ is a point in I_1, $\mathbf{x}'$ is a point in I_2 and H_{12} denotes the plane induced homography from I_2 to I_1. Applying this transformation to the input image I_2 results in the alignment of groundplane points between the two images I_1 and $H_{12}I_2$, whereas points above or below the groundplane do not correspond. Figure 1(a) shows two images I_1 and $H_{12}I_2$ overlaid. Let the space wherein the two images I_1 and $H_{12}I_2$ are be known as *groundplane space*.

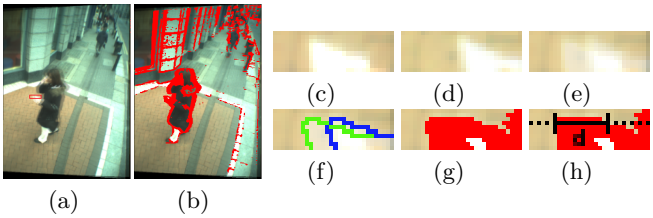

Fig. 1 (a) Groundplane Space; (b) FARs; (c) I_1; (d) $H_{12}I_2$; (e) I_1 and $H_{12}I_2$; (f) Edges; (g) FARs; (g) Disparity

2.2 Foreground Activity Regions (FARs)

GCPs can be found to help guide stereo correspondence techniques resulting in more accurate results. However, false GCPs can severely degrade the final matching results [3]. Introducing stricter constraints on the selection of GCPs could decrease the number of false GCPs, but could also reduce the total number of GCPs, leading to the possibility of an insufficient number of GCPs being available to guide the matching process successfully.

The properties of the groundplane space can be applied to find areas, called Foreground Activity Regions (FARs), in which depth discontinuities are likely to occur. These regions are typically a good place to extract a large number of strong GCPs as they tend to occur in areas of high texture. FARs occur due to foreground objects being above the groundplane, and therefore they do not match in groundplane space, see Figure 1(a). For an illustrative example, take Figures 1(c) and (d), which are corresponding sections from I_1 and $H_{12}I_2$ respectively. As white region in the two images is part of the pedestrians hand, which is foreground, the hand in I_1 and $H_{12}I_2$ does not correspond, see Figure 1(e). This can be more clearly seen by viewing the position of the edge points in I_1 (in green), with respect to the corresponding edge points in $H_{12}I_2$ (in blue) in Figure 1(f). The key property is that, in groundplane space, neither the two sets of edges nor *any* pixel between the two edges match in both colour intensity and gradient information. These pixels are denoted in red in Figure 1(g). Let these regions of non-correspondence be called FARs. An important aspect is that, in general, if there is a jump in disparity, then this disparity discontinuity is incorporated within a FAR.

Once FARs are determined, the FAR is then searched for matching GCPs between images I_1 and $H_{12}I_2$. However, an advantage to obtaining FARs to determine GCPs, instead of just regions or points of high texture within an image, is that a *dynamic* disparity limit constraint is determined for each GCP match. For a *static* disparity limit, the closer an object is to the camera rig, the larger the disparity limit should be. This is an ill-posed problem as the distance of the object to the camera is not determined until the object's depth is obtained. The constraint is therefore usually fixed at the largest expected disparity that should occur in the scene. Using the FARs regions, it is possible to roughly estimate the *maximum possible* disparity. In general, if there is a jump in disparity within a FAR then a match for every point in the FAR will be at a disparity less than the width of that FAR. Observing Figure 1(h) notice how the width of the FAR, d, is greater than that of the disparity of the two edges. The search for GCPs in each FAR is implemented using this variable disparity limit constraint.

2.3 Multiple Initial Ground Control Points (MIGCPs)

Obtaining GCPs from FARs is a three stage process; (1) *multiple* initial GCP matches are found at various disparities; (2) the best choice of disparity for these GCP matches are determined; (3) control point regions are determined.

The first step involves finding Multiple Initial Ground Control Points (MIGCPs). In general it is more difficult to determine the accurate disparity of a point that has homogeneous rather than heterogeneous neighbours. For this reason, we ensure that the matches for the MIGCPs are instantiated by edges that have an edge gradient greater than a threshold, t_sg, which is vertically oriented with respect to the scanline. We refer to a point that meets this criteria as a maximum vertical edge, max_{ve}.

To determine the MIGCPs for a given FAR on a particular scanline, each max_{ve} within the FAR of I_1, called max^1_{ve}, is compared to each max_{ve} in $H_{12}I_2$, called max^2_{ve}, that is within a disparity of the width of the FAR from max^1_{ve}. The two max_{ve} are compared using the sum of absolute square distances (SAD) in their RGB colour intensities. If the SAD is less than a threshold, $t_{MaxAccept}$, then a possible match between the two max_{ve} exists. If this is the case then the process of obtaining additional collaborating information is undertaken. Initially the number of the collaborating value, val_{col}, is 1. Collaborating information is gathered by moving the current point in each image left by a single pixel along the scanline. The new image points fails to be collaborating the max^1_{ve} and max^2_{ve} match if

1. The current point in either image is outside of the FAR or homogeneous in colour with its neighbours;
2. $SAD(curr_1, curr_2) \geq t_{MaxAccept}$, where $curr_1$ is the current pixel in I_1 and $curr_2$ is the current pixel in $H_{12}I_2$

If neither of the above is true, then val_{col} is incremented by one, as another pixel match agrees with the choice of MIGCP match. A similar technique is then applied from max^1_{ve} and max^2_{ve} in the opposite direction along the scanline. The greater the value of val_{col}, the more likely the match is of being correct.

2.4 Final Ground Control Points (FGCPs)

However, determining GCPs based on horizontal matches alone, as is done in MIGCP, can result in false matches. This second step finds collaborating pixels vertically across scanlines, enforcing interline consistency for selecting GCPs. The technique to achieve this can be shown in the following example.

Let Figure 2(a) represent the set of MIGCPs in I_1, so for example on scanline **1** there is 1 MIGCP at pixel $\mathbf{1C}_1$, where the subscript 1 represents I_1. Let Figure 2(b) represent the possible match disparities in $H_{12}I_2$ for the MIGCPs in (a), therefore, for example, the MIGCP in scanline **1** has 3 possible matches at $\mathbf{1A}_2$, $\mathbf{1C}_2$ and $\mathbf{1F}_2$, where the subscript 2 represents $H_{12}I_2$. For each MIGCP in I_1 it is checked to see if there exists one or more MIGCPs in the previous scanline within a distance of 1 pixel. In this example, for $\mathbf{2B}_1$ there is a MIGCP in the previous scanline at $\mathbf{1C}_1$. If more than one exists, the MIGCP that has the closest colour intensity and gradient information to $\mathbf{2B}_1$ is chosen. For each possible disparity match of $\mathbf{2B}_1$; determine if $\mathbf{1C}_1$ has a corresponding possible disparity match in $H_{12}I_2$ in the same vicinity. In this case, the match $\mathbf{2B}_1 \rightarrow \mathbf{2E}_2$, $\mathbf{1C}_1$ has a corresponding match from $\mathbf{1C}_1 \rightarrow \mathbf{1F}_2$. However, the match $\mathbf{2B}_1 \rightarrow \mathbf{2B}_2$, has two corresponding matches from $\mathbf{1C}_1 \rightarrow \mathbf{1A}_2$ and $\mathbf{1C}_1 \rightarrow \mathbf{1C}_2$. In this case the pixel from $\mathbf{1A}_2$ or $\mathbf{1C}_2$ that is closest in colour intensity and gradient information to $\mathbf{2B}_2$ is chosen, assume it is $\mathbf{1C}_2$. If a corresponding match *is* found then the val_{col} for the two pixels in I_1 for the *matched disparity* are added together. In this case the val_{col} for matching $\mathbf{2B}_1 \rightarrow \mathbf{2E}_2$ and $\mathbf{1C}_1 \rightarrow \mathbf{1F}_2$ are added together, as are the val_{col} for matching $\mathbf{2B}_1 \rightarrow \mathbf{2B}_2$ and $\mathbf{1C}_1 \rightarrow \mathbf{1C}_2$, see Figure 2(c).

This process continues to the next MIGCPs in the image. In scanline **4**, there is only one match to $\mathbf{4C}_1$, namely $\mathbf{4F}_2$. This means that the chain in $H_{12}I_2$ containing **1C**, **2B** and **3A** cannot continue. However the chain containing **1F**, **2E** and **3E** grows longer to contain **1F**, now the val_{col} of any of the pixels in this chain is the sum of all the initial MIGCP val_{col} in the chain. By using this technique, longer chains have larger values for val_{col}. FGCPs are then found by obtaining the *highest* val_{col} for each pixel. Finally, in post processing, we remove all FGCPs that are not bidirectional, all FGCPs where the highest val_{col} is not at least *twice* that of the second highest and all FGCPs that do not have a chain that spans at least 3 scanlines. In addition we use dynamic programming to enforce the ordering constraint in FCGPs. These post processing steps ensure that most ambiguous FGCPs are removed.

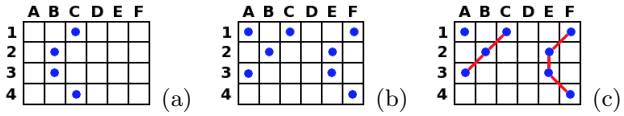

Fig. 2 FGCPs (a) I_1; (b) $H_{12}I_2$; (c) $H_{12}I_2$ links;

2.5 Ground Control Point Regions

At this point it is possible to obtain more GCPs by extending the FGCPs across regions of homogeneous texture. An initial step for this process is to cluster FGCPs into regions of homogeneous disparity using connected components, where two separate regions, r_1 and r_2, can merge if

$$\sqrt{(r_1^d - r_2^d)^2} \leq \frac{r_x^d * \theta}{\sigma} \tag{2}$$

where r_1^d is the disparity of r_1, r_2^d is the disparity of r_2, r_x^d is the lowest disparity of the two regions and θ is the maximum disparity difference allowed for every σ pixels of disparity that exists in r_x^d. In our experiments $\theta = 1.5$ and $\sigma = 10$, allowing 1.5 pixels disparity difference when $r_x^d = 10$, 3 pixels disparity difference when $r_x^d = 20$, and so on. This allows pixels closer to the camera to cluster together easier, allowing for fluctuations in the objects surface. Two separate FGCPs regions can then be merged, if they are separated by a region of homogeneous texture. If this is the case then all pixels between the two FGCPs regions are also classified as FGCPs.

2.6 Background Ground Control Points (BGCPs)

Background objects, such as walls, do not, in general, change position within an input image. If a sequence of images of the same scene is used, background disparity and edge models can be built. The initial background disparity model can be set to have the same disparity as the groundplane space. The model will then be updated over time to incorporate the background objects such as walls, bollards, etc. These background models can be used to eliminate the need for searching for GCPs that arise from background objects. To implement this, we first stop looking for GCPs that occur due to background objects by using the background disparity model to detect FARs instead of the groundplane space. Background objects will therefore *not* cause FARs, and thus no GCP will be searched for in these regions. *Background* GCPs (BGCPs) are then determined using background maximum vertical edges, $bmax_{ve}$, which are strong vertical edges that do *not* appear as foreground in either the background edge or disparity model. If two $bmax_{ve}$

are separated vertically by a region of homogeneity that does not contain a FAR then all points between the background maximum vertical edges are defined as a BGCP region. A BGCP region can propagate downward to the next scanline *iff* there is no FAR anywhere inside the BGCP region on the next scanline. Finally a propagated BGCP region can extend left and right until a FAR or a maximum vertical edge is reached. Figure 3(d) shows results from obtaining FGCPs and BGCPs.

2.7 Dense Disparity Estimation

The GCPs can be used as a basis to guide various different stereo correspondence algorithms. We used them in conjunction with a single pass dynamic programming based approach, allowing GCPs to have zero cost in the matching process. It optimises the scanlines by minimising an energy function that is similar to [3];

$$E(d(x,y_1)) = \sum_x C(x,y_1,d(x,y_1)) + \sum_x (\lambda(x,y_1)\rho(d(x,y_1) - d(x+1,y_1)) + \lambda(x,y_1)\rho(d(x,y_1) - d(x,y_1-1)))$$

where y_1 is the current scanline, ρ is the Potts model and $\lambda(x,y_1)$ is a weight function, set to 30 in our experiments. A cost added for both a vertical *and* a horizontal difference in disparity to help enforce inter scanline consistency. In addition the idea of a dynamic disparity limit constraint is used. The limit for a given scanline is found as the maximum of the previous lines dense disparity, the current and next lines background model and GCP disparity.

3 Experimental Results

Figure 3(e) shows results of the dense disparity estimation technique using two cameras from a Digiclops [6] stereo camera rig. The end goal of this research is to count pedestrian numbers in a scene. It is clear from the results that the disparity information obtained will greatly aid pedestrian segmentation, but this information, in itself, is not enough. Take for example, Figure 3(e), row 3, where a man has roughly the same disparity as the wall. We are are therefore unable to separate the pedestrian by stereo information alone. It is also important to also notice that only the pedestrians head and hands appear as FARs. This is due to the pedestrians torso and the background having the same colour and therefore there are no edges or indication of a disparity jump. This problem can also occur if one pedestrian is occluded by another wearing the same coloured clothes. The result of this would be an undefined region between the two pedestrians as there would be no visible disparity discontinuity.

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

Fig. 3 (a) Groundplane Space; (b) FARs; (c) FGCPs; (d) FGCPs and BGCPs; (e) Disparity

4 Conclusions and Future Work

This paper described a technique for dense disparity matching designed for pedestrian detection type applications using GCPs and a dynamic disparity limit constraint. In future work, the use of temporal data along with the improvement of the dense disparity algorithm to enforce better inter scanline consistency would increase the accuracy of the dense disparity data. Current work includes the use of disparity and biometric information, such as height/width, being applied to segment objects, even at the same depth, into separate objects and the classification of that object as pedestrian or otherwise.

References

1. L. Zhao and C. Thorpe. Stereo and neural network-based pedestrian detection. *IEEE Trans. on Intelligent Transportation Systems*, 1(3):148–154, September 2000.
2. D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. Journal of Computer Vision,*, 47(1/2/3):7–42, April 2002.
3. B.T. Choi C. Kim, K.M. Lee and S.U. Lee. A dense stereo matching using two-pass dynamic programming with generalized ground control points. In *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2005.
4. A. Fusiello, E. Trucco, and A. Verri. A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*, 12(1):16–22, 2000.
5. P. Kelly, P. Beardsley, E. Cooke, N. O'Connor, and A. Smeaton. Detecting shadows and low-lying objects in indoor and outdoor scenes using homographies. In *IEE Conf. on Visual Information Engineering*, April 2005.
6. http://www.ptgrey.com/products/digiclops/index.html.