Abstract - This paper presents a new approach for fast and accurate stereo correspondence. Some of the efficient and robust implementation aspects of the stereo matching algorithms based on thresholding of correlation values of previous pyramid structures that reduce the search area for finding the best matching point are addressed. The disparity for each scan line is chosen by selecting the position that gives the maximum correlation coefficient from the pyramid structure. However, sometimes this maximum value may be overridden by a spurious maximum. Ambiguities mainly come from image noise, lack of intensity variation or geometric distortion. For disambiguation, the adaptive window theory is applied to avoid mismatch in featureless area. Adjusting threshold parameter at each level of the pyramid structure reduces computational time and increases the accuracy for the next pyramid level. Both synthetic and real image sets have been tested and good results have been obtained.

Keywords: stereo matching, disparity, pyramid structure, correlation, adaptive window.

1 Introduction

The different perspectives of two eyes lead to slight relative displacements of objects in the two monocular views of scene. At the same time, the same object will give rise to a pair of different images. The 3D shape and location of the object can be recovered by fusing these stereo pair. Stereopsis is a useful method for machine perception. The objective of scene analysis by stereopsis is to recover the three-dimensional (3D) location of features from their projection in two-dimensional (2D) images [1]. For recovering 3D shape using stereo, the correct and fast estimation of disparities is the most difficult problem. Given a point in the left image, the problem is to find its corresponding point in the right image, which has additional image variations. Differences of stereo pair might be caused by occlusion of objects, specular reflections, which move independently of the surfaces of objects, sensor noise and various other causes. For establishing image correspondences, many different algorithms have been proposed. Based on [2] and [3], we classify the current stereo algorithms into two categories: feature-based and area-based. Feature based stereo techniques employ simple, geometric primitive such as line segments and planar patches [1,4]. Such models are appropriate for simple, well-constructed environments consisting of man-made objects. These techniques have generally been more successful overall as the matching process is much faster than the area-based techniques and there are fewer feature points to be considered [5]. Despite the more disambiguating power of features as compared to intensity values, intensities are still needed to resolve remaining ambiguities. The area-based stereo techniques compare the gray levels (intensity) of image patches to find the corresponding pixel in the other image and represent depth at each pixel in the image. The simplest method of such comparisons is correlation [6,7]. Many area-based algorithms have been researched but in most cases, they have proved to be unsatisfactory as they produce poorly defined matches. They have been found to be highly sensitive to distortion in gray level and geometry, and are computationally very expensive.

This paper presents a new approach for correct and fast calculation for stereo matching. We address some of the efficient and robust implementation aspects of the stereo matching algorithms by using fast correlation and adaptive windowing techniques in a multi-resolution scheme [8,9]. This paper has been organized in the following sections: Section 2 briefly describes the stereovision setup and geometry used. Section 3 describes the fast correlation method in area-based technique. Section 4 includes the method for estimating disparity with adaptive window scheme, pyramid structure and Gaussian low-pass filtering. Section 5 gives the results obtained through the proposed algorithm and explains the entire depth estimation procedure in the form of an algorithm.

2 Stereo System

One of the basic problems to be solved in stereovision is shown in figure 1. A 3D feature in the scene is viewed by the stereo head in the 2D image plane. Loss of depth information results when a number of 3D points project onto the same 2D point. To recover this lost information, two images taken from different perspectives are required. The simplest model shows two identical cameras separated only in the y direction by a baseline distance D and having a known focal length f. The image planes are
A point $P$ defined by its coordinates $(x,y,z)$ in the real world, will project as the corresponding 2D image coordinates $(x_l,y_l)$ and $(x_r,y_r)$ onto the left and right images respectively. $O$ is the origin, which coincides with the image center in the left camera. The perspective projection can be defined through simple algebra as,

$$x = x_l \cdot \frac{D}{d}$$  \hspace{1cm} (1)

$$y = y_l \cdot \frac{D}{d}$$  \hspace{1cm} (2)

$$z = f \cdot \frac{D}{d}$$  \hspace{1cm} (3)

The disparity $d$ is the displacement between the locations of the two features in the image plane.

$$d = \|x_l - x_r\|$$  \hspace{1cm} (4)

These equations provide the basis for deriving the 3D depth from stereo images.

### 3 Fast Correlation

The stereo head used in this work is calibrated and has a fixed base-line distance $D$ that generates parallel epipolar geometry and hence the search surface is reduced from two dimensions to one dimension. As shown in figure 1, given two images which form in the retinal planes, we want to find out which point in one retinal plane corresponds to a point in the other retinal plane. Given a point in one retinal plane, it may a priori be put in correspondence with any point in the other retinal plane. To solve this problem, intensity-based area correlation technique has been applied.

Let $f_{mn}$ be the intensity value of an $M$ by $N$ image $f$ at position $(m, n)$ and also have similar definition for a second image $g$. The principle of the correlation techniques is shown in figure 2.

In order to define the coordinates of the pixel in right image that matches the pixel of coordinates $(m,n)$ in left image, we consider a rectangular window of size $K$ by $L$, centered at $(m,n)$ and compute its correlation $\gamma_{ij,d}$ with the second intensity image.

The normalized cross-correlation of two windows can be written as follows:

$$\gamma_{ij,d} = \frac{\sigma_{ij,d}(f,g)}{\sigma_f \times \sigma_{ij,d}(g)}$$  \hspace{1cm} (5)

Where, covariance,

$$\sigma_{ij,d}(f,g) = \sum_{m=-K}^{i-K} \sum_{n=-L}^{j-L} (f_{mn} - \tilde{f})(g_{m+n,d} - \tilde{g})$$

$$= \sum_{m=-K}^{i-K} \sum_{n=-L}^{j-L} f_{mn} g_{m+n,d} - (2K+1)(2L+1)\tilde{f} \cdot \tilde{g}$$  \hspace{1cm} (6)

and variance

$$\sigma_f^2 = \sum_{m=-K}^{i-K} \sum_{n=-L}^{j-L} (f_{mn} - \tilde{f})^2$$

$$= \sum_{m=-K}^{i-K} \sum_{n=-L}^{j-L} f_{mn}^2 - (2K+1)(2L+1)\tilde{f}^2$$  \hspace{1cm} (7)

$$\sigma_g^2 = \sum_{m=-K}^{i-K} \sum_{n=-L}^{j-L} (g_{mn,d} - \tilde{g})^2$$

$$= \sum_{m=-K}^{i-K} \sum_{n=-L}^{j-L} g_{mn,d}^2 - (2K+1)(2L+1)\tilde{g}^2$$  \hspace{1cm} (8)

and $d$ is the shift along epipolar lines: $K$ and $L$ define the correlation window size. It can be seen from this equation that the covariance between $f$ and $g$, and the variance of $f$ and $g$ at different positions in the image need to be evaluated. The variances and covariance of points within the window are calculated using Equations (6), (7) and (8). Therefore we have a fast way to obtain the variance and covariance of the input images for the calculation of the cross-correlation using only simple computations that are to be used as a measure of similarity between
matching candidates from the left and right image. In our case, the correlation is performed along the epipolar line. If for any point in the left image, the search window is assumed to be within 0 to z in the right image, then the value d, in equation (5) varies from 0 to z. The traditional way of obtaining the correlation is to fix a point in the left image and vary d in the range of 0 to z to calculate the correlation coefficients.

In our algorithm for fast correlation, we first fix on one particular d at all the points in the left image and calculate the cross-correlation between the left image and the shifted right image. The right image is shifted by an amount d. Then we increase the number of d by 1, and repeat the process of cross-correlation calculation until the value of d has gone through 0 to z. We store the correlation coefficients in the correlation matrix, which was calculated as with the above procedure. The correlation matrix consists of all possible matches between points in the left and right images.

4 Estimating Disparity & Uncertainty

4.1 Adaptive windowing scheme

As Barnard and Fischler [10] point out, “a problem with correlation matching is that the window size must be large enough to include enough intensity variation for matching but small enough to avoid the effects of projective distortion”. If the window is too small and does not cover enough intensity variation, it gives a poor disparity estimate, because the signal (intensity variation) to noise ratio is low. If, on the other hand, the window is too large and covers a region in which the depth of scene points (i.e. disparity) varies, then the position of maximum correlation or minimum Sum of Squared differences (SSD) may not represent correct matching due to different projective distortion in the left and right images [11]. For this reason, a window size must be selected adaptively depending on local variations of intensity and disparity. However, most correlation or SSD based stereo methods in the past have used a window of fixed size that is chosen empirically for each application. In our algorithm, the window size is starting from 3 by 3 and its increased until it covers enough intensity variation.

Ideally the best match point can be detected by maximum correlation coefficient in each scan line. However, several problems occur with this technique.

- If the maximum of the correlation function is not well defined, the disparity is not very accurate.
- Sometimes this maximum value may be overridden by a spurious maximum.
- If the intensity values in the window are all same (no features in window), we cannot define the maximum correlation function because the correlation value is 0.
- Computationally very expensive
- The window size must be large enough to include enough intensity variation for matching, at the same time small enough to avoid the effect of projective distortion.

Ambiguities mainly come from image noise, lack of intensity variation or geometric distortion. For disambiguation, we apply the adaptive window theory and adjust the threshold level at each level of pyramid structure.

4.2 Thresholding using Pyramid Method

It has been shown that a multi-resolution or pyramid data structure for stereo matching is much faster than one without multi-resolution, as the search range in each level is small. Besides fast computation, a more accurate disparity map can be obtained by exploiting multi-resolution. The upper levels of the pyramids are ideal to get an overview of the image scene. The details can be found down the pyramid at higher resolution.
During the process of projecting the disparity map from the current level of the pyramid to the next (if current level is not level 0), the image size is scaled up by a value of $\tau$ (reduction ratio), and the disparity value by the same $\tau$. The disparity value where the position $(i, j)$ of the new image is not a multiple of $\tau$ is obtained by linear interpolation. In our case, we applied Gaussian Pyramid structure with reduction ratio of 2. Therefore, disparity range of next level is a twice that of the previous level.

If we have a correct disparity value at the current level of pyramid structure, the disparity value at the next level can be defined with above method (figure 5, left). But if we have a wrong estimation of the disparity value at the current level, the resulting estimation at the next level is always incorrect. In most of the image cases, we cannot apply the above method directly because of the ambiguity of disparity estimation. In our algorithm, we define the threshold value for distinguishing the estimation quality. The procedure of Gaussian pyramid is the low-pass filter the original image to obtain a reduced version of the image. Therefore the neighboring pixels in the low-resolution image are more correlated than pixels in the high-resolution image. At each stage of the Gaussian pyramid, the distribution of maximum correlation coefficients is different. Most of the maximum correlation values at the lowest resolution stage are distributed between values of 1 to 0.9 and they decrease with increasing image resolution (figure 6).

We defined threshold value with the average of maximum correlation coefficients at each level multiplied with $\alpha$, multiplication factor. There is an inverse relationship between the disparity map image quality and the processing time based on the value of $\alpha$, the multiplication factor. If the maximum correlation value at present stage is greater than the threshold value, expand the disparity value for the next pyramid structure directly (figure 5 right). The search space for picking up the maximum correlation value at next stage is much reduced. If the maximum correlation value at present stage is smaller than the threshold value, we assume that matching is incorrect and that the particular pixel has an ambiguity because of the image noise, lack of intensity variation or geometric distortion. In this case, we apply the Gaussian estimation to these particular points for checking the spatial surface continuity and noise smoothing.

Gaussian filters are a class of linear smoothing filters with the weights chosen according to the shape of the gaussian function. It removes noise effectively from a normal distribution. These filters are effective low-pass filters from the perspective of both the spatial and frequency domains. The kind of filter used here is the two-dimensional Gaussian filter. These filters are rotationally symmetric, which means that the amount of filtering performed will be the same in all directions. The property of rotational symmetry implies that the Gaussian Smoothing Filter will not bias subsequent edge detection in any particular direction.

The applied 2-D Gaussian filter is defined as

$$d_{new}(i, j) = \sum_{m=-2n}^{2} \sum_{n=-2m}^{2} mask(m, n) d(i + m, j + n)$$

The mask $(m,n)$ is normalized, and the symmetric two dimension Gaussian mask is defined as

$$\sum_{m=-2n}^{2} \sum_{n=-2m}^{2} mask(m, n) = 1$$

4.3 Estimation disparity for ambiguous points

In this section we will see as to how the filtering of the initial estimated disparity map is done. As explained in the previous section we also have the image showing the ambiguous points in the initial disparity map image. In order to reduce the computation time the gaussian filtering is applied on those points in the disparity map image that show ambiguity.

5. Result

The new approach of correct and fast calculation for stereo matching has been tested on both synthetic and real image sets with good results. Figure 7 explains the whole procedure of proposed algorithm. Three different kinds of
images have been tested here. The three sets have been tested at three different resolutions.

The first image set (figure 8) is a real outdoor image data of the pentagon obtained from the image database at the computer vision homepage. The original image size was 256x256. The value of the multiplying factor for the average maximum correlation value used here is 1.05, which helps in determining and extracting the ambiguous points from the disparity map image. Though this increases the computational time the results prove to be very good for depth estimation. The second image set (figure 9) used is a real indoor image set of size 117x613. These images were taken using a stereo camera setup. The images are not from a controlled environment. But the proposed algorithm for depth estimation gave good results. Finally, the third image set (figure 10) used is a synthetic stereo pair of the size 256x256. The multiplying factor has a lesser value here, of 0.95, which reduces the computation time. The results obtained are comparable to those obtained by other proposed algorithms.

Figure 7. Diagram of Proposed algorithm

Figure 8. Example of pyramid structure and disparity estimation (thresholding with $\alpha=1.05$)

Figure 9. Example of pyramid structure and disparity estimation (thresholding with $\alpha=1$)
Figure 10. Example of pyramid structure and disparity estimation (thresholding with $\alpha=0.95$)

References