4-32

MVA2011 IAPR Conference on Machine Vision Applications, June 13-15, 2011, Nara, JAPAN

Dense Stereo Disparity Maps - Real-time Video Implementation by the Sparse Feature Sampling

Kunio Takaya
Electrical and Computer Engineering, University of Saskatchewan
57 Campus Drive, Saskatoon, SK. Canada S7N 5A9

Abstract

To realize the real time dense stereo disparity map (DDM) running at a video rate of 30 fps, the dynamic time warp algorithm (DTW) is time wise a bottle neck despite its robustness for stereo matching. The DTW method requires to calculate a large similarity matrix $\mathbf{S}$ *of the size* N^2 *for the raster size* N*, if pixel-by-pixel matching is attempted. The computation time to calculate* $\mathbf{S}$ *is significant for real-time systems and embedded hand-held devices. Two methods, coarse quantization method and hump detection method, to reduce* N *by sparse feature sampling are proposed in this paper. Both proposed methods reduce* N *much below the raster size, and create a set of sparse samples without sacrificing the spatial resolution for stereo matching. The size of the sparse set is typically* $N = 30$ *and* $N = 15$ *for each respective method, compared with the raster size* $N = 320$*. Thus, the calculation time of DDM is dramatically improved by more than 100 times. By using the proposed methods, a real-time system was realized on the Windows platform.*

1 Introduction

The dense stereo disparity map, i.e. image based distance measurement has been developed for applications such as robotic vision and video surveillance. Two small video cameras embedded in the hand-held computer or game controller can be turned into a distance image sensor or a range finder. Electrical retina stimulation with the implanted 2D electrode array that has recently been reported is potentially capable for blind people to regain vision with the technology of BMI (Brain Machine Interface). The dense disparity map is definitely one important mode of artificial vision to sense the distance by vision, when such BMI is fully developed. The challenge is to perform frame-by-frame image processing fast enough to keep up with a video rate. [1] The objective of study is to develop a robust and fast system capable of displaying the dense stereo disparity maps continuously at a video rate.

2 Stereo Matching by Dynamic Time Warp Algorithm

Stereo matching in the binocular image pair is to find two corresponding feature points, one in the left image and the other in the right image. The distance between the corresponding points are referred to as stereo disparity that is inversely proportional to the distance to the point in the 3D space. Finding the correspondence of all points in a sequence to those in another sequence is a problem of optimization to keep the error of mismatching to be minimum. The Dynamic Time Warp Algorithm (DTW) [7], or Dynamic Programming (DP) is the algorithm for stereo matching when the pixel points are arranged in a sequence of distance. Due to this constraint, the DTW is particularly useful to find stereo correspondence in the raster scanned 1D raster profiles (waveforms). The DTW is one of the robust algorithms which is known to work even for the image having some occluded objects, meaning that an object is seen from a camera but not from the other. When the search for matching is exhaustively done for each and every pixel in the raster, the process is time consuming as N is the raster size, and N^2 is very large. However, if feature points are sparsely sampled, for example at edges or the peak of humps, N^2 can be made

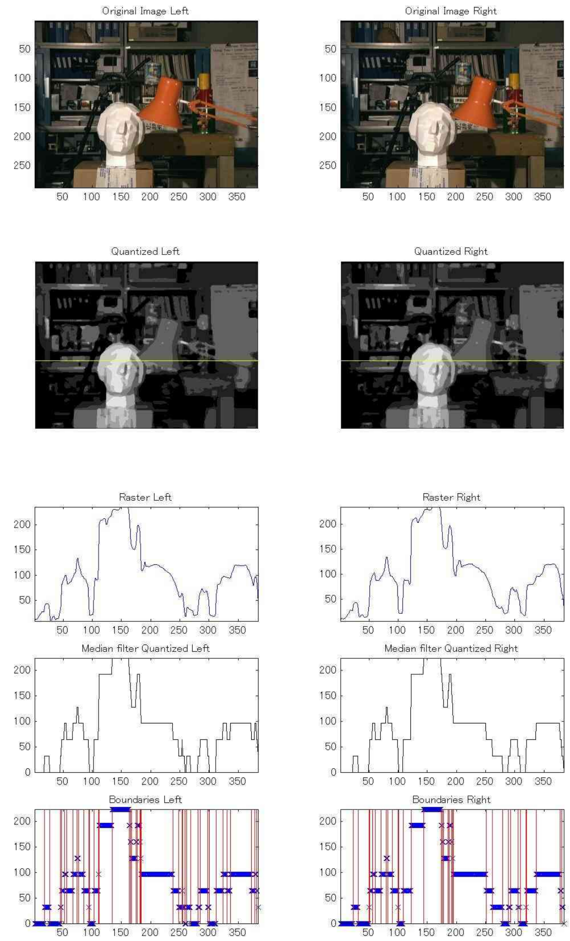

Figure 1. Original and Coarsely quantized images, and waveforms of median filtered images, that of coarse quantization, and transitional points.

.	.	.	.	$I_\ell(n-1)$	$I_\ell(n)$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
$I_r(m-1)$	.	.	.	$C(X_{1\cdots n-1}, Y_{1\cdots m-1})$	$C(X_{1\cdots n}, Y_{1\cdots m-1})$
$I_r(m)$	.	.	.	$C(X_{1\cdots n-1}, Y_{1\cdots m})$	$s(n,m)$ + min. of 3 neighbours
.	.	.	.	.	.

Figure 2. Local decision making in the cost matrix **C** by the DTW dynamic programming

considerably smaller.

In the DTW algorithm [7], the similarity matrix **S** plays the key role in the optimization process. **S** indicates how similar the nth pixel point of the sequence I_ℓ is to the mth pixel point of I_r.

$$s(n,m) = \sum_{k=-L/2}^{L/2} |I_\ell(n+k) - I_r(m+k)|$$

for a window size $L+1$. Stereo matching is the problem to minimize the penalty to match dissimilar points to match all points in I_ℓ and I_r. Define the left pixel array up to the nth element, and the right pixel array up to the mth element as

$$\begin{aligned} X_{1\cdots n} &= \{I_\ell(1), I_\ell(2), \cdots, I_\ell(i), \cdots, I_\ell(n)\} \\ Y_{1\cdots m} &= \{I_r(1), I_r(2), \cdots, I_r(j), \cdots, I_r(m)\} \end{aligned}$$

The element of the cost matrix **C** is given by

$$C(X_{1\cdots n}, Y_{1\cdots m}) = \\ s(n,m) + min \begin{cases} C(X_{1\cdots n-1}, Y_{1\cdots m-1}) \\ C(X_{1\cdots n}, Y_{1\cdots m-1}) \\ C(X_{1\cdots n-1}, Y_{1\cdots m}) \end{cases}$$

$C(X_{1\cdots n}, Y_{1\cdots m})$ is the minimum cost to match the nth pixel point of I_ℓ and the mth pixel point of I_r among all previous matches of $\le n$ and $\le m$. There are only three choices to match $I_\ell(n)$ and $I_r(m)$. Given the costs to match up to $I_\ell(n-1)$ and $I_r(m-1)$, $I_\ell(n-1)$ and $I_r(m)$, $I_\ell(n)$ and $I_r(m-1)$, the smallest of the three costs plus the similarity $s(n,m)$ is the cost to match $I_\ell(n)$ and $I_r(m)$. Back tracking the matrix **C** shown in Fig. 2 from the right bottom corner yields the optimum path that defines the best matching of all pixels. The size of a raster waveform is $N = 320$ for the CIF image. The size of the similarity matrix **S** and the cost matrix **C** is $N^2 = 102,400$. Furthermore, the window size is practically as large as $L = 10$. The required computation for **S** and **C** is about 1 million calculations, which substantially make the implementation of the DTW algorithm difficult in video applications.

3 DTW Implementation for Coarsely Quantized Raster Waveforms

If a raster waveform is coarsely quantized into several levels, pixel values are grouped by the levels. Uniform runs of a continuous quantization level is regarded as originated from the same video object. A run is defined by the starting edge and the ending edge so that the same disparity can be assigned to the whole run length. However, the raster waveform captured from a video camera is affected by various sources of noise such as photon noise, thermal noise, and electronic noise. The source image, generally, requires some sort of denoising. Low-pass filtering removes high-frequency noise, but affects the slope of edges where important feature information, i.e. position, is recorded. If edges are smeared, the stereo disparity to be determined by corresponding edges becomes less accurate. The median filter, statistical nonlinear filtering removes noises such as salt and pepper noise, but the sharpness of edges is not altered.

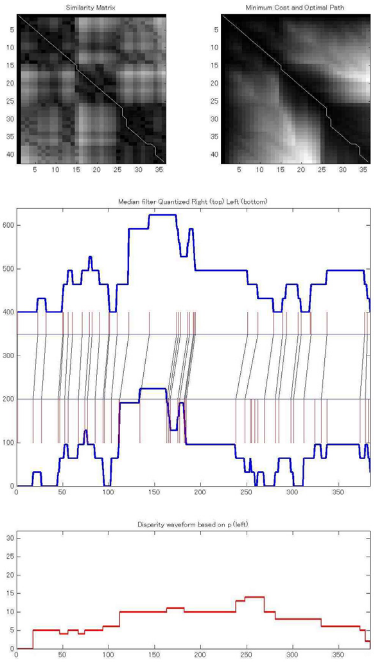

Figure 3. Similarity matrix **S**, cost matrix **C** and the optimal path in white (top), correspondence of all one-to-one matched transitional points (middle), and disparity profile in red (bottom)

2D 5×5 median filter is effective to denoise the source images. The coarse quantization applied to the median filtered image produces a patchy segmented image as shown in Fig. 1. The raster waveforms of the median filtered and coarsely quantized image are shown also in Fig. 1. Then, all transitional points and run-lengths are registered. Transitional points where jump occurs are shown in the bottom row of Fig. 1 with red lines. The number of transitional points is typically about 30, whereas the raster size is 320.

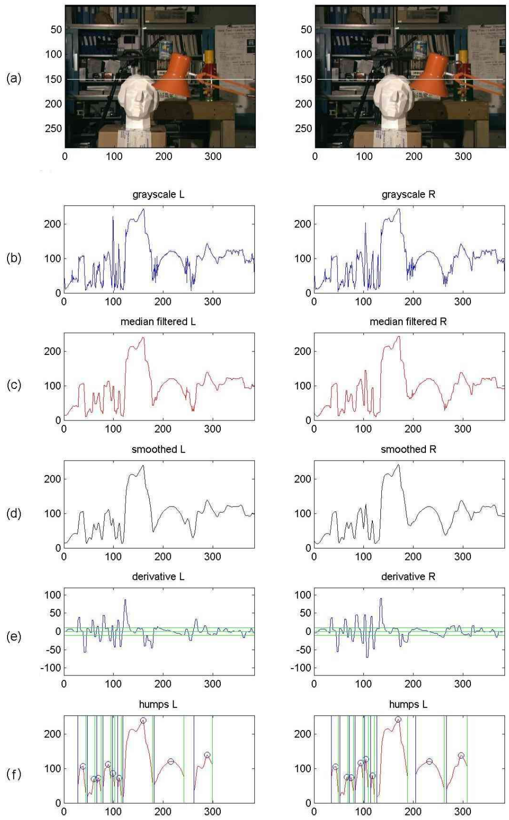

Figure 4. Sequence of digital signal processing (DSP) for hump detection: (a) Original left and right image, (b) Raster waveforms at the cursor line of (a), (c) Processed by 7 point 1D median filter, (d) processed by 5 point FIR smoothing filter, (e) First derivative of waveforms in (d) and the threshold $\pm\epsilon$, (f) Detected humps indicated by a start line in black, a stop line in greed, and a circle at the peak of the hump.

The similarity matrix **S** is constructed from the transitional points. The element of the **S** matrix is modified to include the proximity term not to match pixels of too far apart.

$$s(n,m) = \sum_{k=-L/2}^{L/2} |I_\ell(n+k) - I_r(m+k)| + \alpha(|n-m|)$$

Where, $I_\ell(.)$ and $I_r(.)$ are the median filtered waveorms. Since the number of the transitional points is roughly 10% of the raster size, the calculation time to construct the **S** matrix is much shorter than the case of pixel-to-pixel matching, i.e. $N^2 = 0.1^2 = 1\%$. The similarity matrix **S** and the cost matrix **C** are shown in Fig. 3. The minimum path resulted from the DTW algorithm is shown by the white zig-zag line. Vertical or horizontal line segments mean the multiple matching from a point, likely due to the occlusion. When the nth point of the left image is found to correspond to the mth point in the right image, the disparity is defined as

$$\text{disaprity}(n,m) = m - n$$

In calculating the disparities, the points of multiple matching are ignored. The correspondence of all valid one-to-one matches are shown in Fig. 3. The disparity profile is shown in red in Fig. 3.

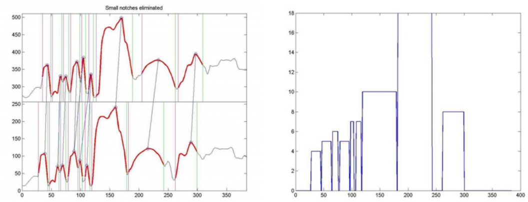

Figure 5. Correspondence of humps in the left and right waveforms, and the calculated disparities

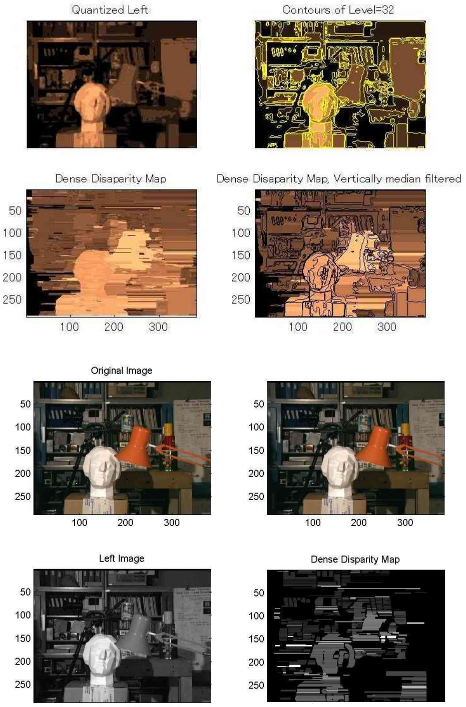

Figure 6. Dense disparity map by the coarse quantization method (top), and by the hump detector (bottom) of an image from the Tsukuba data base.

4 Sparse Correspondence Based on Major Humps in the Raster Profile

Another approach to determine the sparse correspondence of features is to use major humps that exist in a raster waveform. Humps in the raster scan line generally represent image objects. The hump means an upward convex shape defined by a sharp rising and a falling edge. Mathematically, a hump is defined for the waveform $f(t)$ for $t \in [t_1, t_2]$ with the condition that satisfies,

$$\frac{df(t_1)}{dt} > +\epsilon, \quad \frac{df(t_2)}{dt} < -\epsilon$$

ϵ specifies the steepness of the slope. Since the hump here is defined only in terms of the derivatives of the raster waveform, a hump is not necessarily a single modal hump, but it could have multiple modal

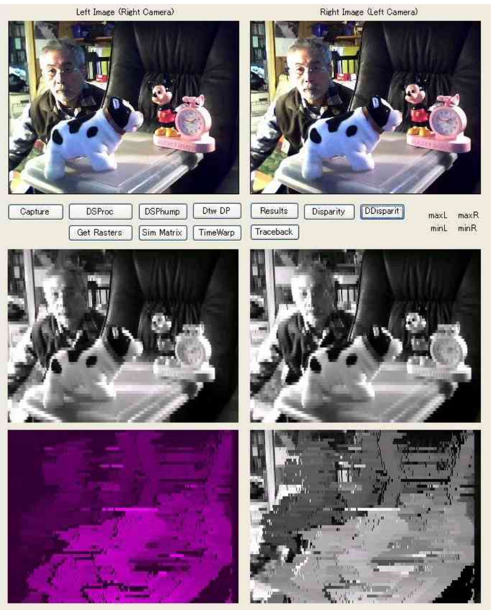

Figure 7. The screen shot of the real-time video system continuously displaying the DDM

points (peaks). Since hump detection uses the derivative waveforms, it is important to denoise waveforms without altering the hump profile. The following digital signal processing (DSP) is applied to successfully detection the hump and extract sparse features. (1) 7 point median filter, (2) 5 point symmetric smoothing FIR filter, (3) differentiate the denoised and smoothed left and right waveforms, (4) apply the threshold ϵ, and (5) register the duration, peak position, peak value of each hump. This sequence of DSP is illustrated in Fig. 4. The humps found by the procedure are shown in (f) of Fig. 4 for the left and right waveform at the raster scan line in white. The thresholds applied to the first derivative, $\pm\epsilon$ are shown in (e) of Fig. 4 with green lines.

For each hump detected, say ith hump of the left, duration $d_L(i)$ of the hump, peak value $v_L(i)$ are recorded. Similarly, $d_R(j)$, $v_R(j)$ are recorded for the jth hump of the right waveform. The element of the similarity matrix $\mathbf{S}$ in this case is

$$s(k,\ell) = \alpha|d_L(i) - d_R(j)| + \beta|v_L(i) - v_R(j)| + \gamma|i - j|$$

Fig. 5 shows the stereo matching with the DTW algorithm applied to the sparse features of the humps. The numbers of humps found for this particular scan line are 9 for the left, and 9 for the right waveform. In optimization, a set of parameters $(\alpha, \beta, \gamma) = (0, 0.67, 0.33)$ was used. Using the correspondence between the left and right humps shown in Fig. 5, the disparity profile of this scan line is shown in Fig. 5.

The DDM for a pair of stereo images in the Tsukuba database is shown in Fig. 6 for the coarse quantization method (top), and for the hump detector method (bottom). The image of coarse quantization and the DDM are shown with and without contour lines superimposed. The DDM for the hump detector method shows dark borders because no disparity is measured in between the adjacent humps. In the DDM, the greater the disparity, the brighter is the pixel. The closer objects are shown with the brighter intensity.

5 Real-time DDM System

The system is a software based system except that two USB CCD cameras (Webcam) were used. As two identical USB cameras are simultaneously used for video capturing, the driver has to be of type that recognizes different units of the same camera as separate units. Only a few USB cameras such as QuickCam Pro 4000, QuickCam Pro 9000, Watchport/V2 and Microsoft LifeCam VX-700 can be used in the multi-camera applications. The spacing between the two cameras is set between 6 cm to 10 cm. The programs were written in Visual C# 2008 Express (C language). XVideoOCX (Marvelsoft) was used to interface USB cameras to the programs, mainly utilizing its real-time video capturing capability. For video rate control of the developed programs, the timer interrupt caused by the end of frame was used. Fig. 7 is a screen shot of the developed system capable of showing the DDM in real-time at a video rate of about 15 fps, a little short of the aimed target.

References

[1] S. Forestmann, Y. Kanou, J. Ohya, S. Thuering, Real-Time Stereo by Using Dynamic Programming, Computer Vision and Pattern Recognition Workshop, vol. 27, issue 02 June 2004.

[2] A. Klaus, M. Sormann, K. Karner, Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure. ICPR 2006, Vol. 3, pp. 15-18, Hong Kong, Aug. 20-24, 2006.

[3] Christopher M. Christoudias, Bogdan Georgescu and Peter Meer, Synergism in Low Level Vision, Int. Conf. on Pattern Recognition. ICPR 2002, Vol. 4, p. 40150 Quebec City, Aug. 11-15, 2002

[4] S. Birchfield and C. Tomasi, Depth Discontinuity by Pixel-to-Pixel Stereo, The 6th IEEE Int. Conf. on Computer Vision, Bombay, Mumbai, India, pages 1073-1080, January 1998.

[5] M. Gong and Y-H Yang, Fast Stereo Matching Using Reliability-Based Dynamic Programming and Consistency Constraints, Proc. of the 9th Int. Conf. on Computer Vision (ICCV 2003) pp. 610-617. Nice, 2003.

[6] H. Hirschmuller, Stereo Vision in Structured Environments by Constistent Semi-Global Matching, IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR 2006) New York, June, 2006.

[7] Dan Ellis, Dynamic Time Warp (DTW) in Matlab, http://labrosa.ee.columbia.edu/matlab/dtw/