Benchmarking of Bootstrap Temporal Stereo using Statistical and Physical Scene Modelling

S. Crossley, N. A. Thacker* and N. L. Seed
Dept. of Electronic and Electrical Engineering
University of Sheffield, Mappin St., Sheffield, S1 3JD, UK
s.crossley@sheffield.ac.uk

Abstract

Temporal stereo vision algorithms can offer improved robustness, however, this can only be delivered after several frames of a stereo image sequence have been processed. We present a new method of bootstrapping temporal stereo which can overcome such start-up problems by applying additional coarse-to-fine pre-processing to the first few images in a stereo sequence. To gauge the performance of temporal bootstrapping, we have employed a new algorithmic evaluation technique that uses statistical and physical scene modelling to produce accurate result errors data. The performance of the bootstrap temporal stereo algorithm, as determined by the automatic evaluation technique, as well the results from real stereo image sequences, are presented.

1 Introduction

In [3], the issue of robustness in stereo vision was tackled with the introduction of a temporal stereo algorithm which used a feedback loop to integrate data over a sequence of stereo image pairs. In this work, we address the problems with robustness that are present in temporal stereo at start-up; when the initial few frames of a stereo sequence are being processed. Our new coarse-to-fine (CTF) bootstrap algorithm allows a temporal stereo algorithm to produce more robust data from the very first frame of an image sequence.

In response to our need to evaluate the stereo bootstrap algorithm, a new stereo evaluation framework has been developed which can test the performance of a stereo algorithm on complex scenes containing multiple objects and motions. By using a combination of artificial scenes generated by image rendering software, and corresponding scene geometry models, any 3-D result produced by a stereo algorithm can be analysed and an accurate outlier count given. The new algorithmic test-bed goes some way to solving the problems of algorithmic benchmarking, because it allows the relative performance of algorithms to be compared directly without the need for human interpretation of image data or results.

In Section 2, we present a short introduction to temporal stereo followed by a

* N.A.Thacker is now at Dept. Medical Biophysics, University of Manchester.

BMVC 1998 doi:10.5244/C.12.35

description of the CTF bootstrap process in Section 3. Section 4, outlines the new stereo evaluation technique and how it was employed in assessing the performance of CTF temporal stereo. Section 5 contains the experimental results for the performance of the bootstrap temporal stereo algorithm on both the artificial and real image sequences, together with a statistical performance evaluation as determined by the automatic evaluation technique.

2 Temporal Stereo

The principal problem in stereo vision is the correspondence problem [2]. Conventional stereo vision algorithms are usually single cue methods, for example matching using edge characteristics, which can often be unreliable due to insufficient disambiguation power. In an effort to improve the robustness of stereo vision, there is growing awareness that temporally integrating stereo match data can increase the probability of finding unambiguous matches by allowing the correspondence task to use two or more image and motion cues. Examples of several temporal stereo techniques which have been suggested are; [3], [20], [9], [19], and [10]. All these works agree that the inclusion of temporal constraints within a stereo algorithm can reduce the probability of a mismatch and, hence, improve the robustness of the 3-D result.

This work follows on from the work in [3]; a stretch correlation based temporal stereo algorithm. The original stretch correlation algorithm [11] used area based correlation to guide the matching of the left and right edge features. The algorithm exploited epipolar geometry, to reduce the correspondence problem to a 1-D search, together with a disparity gradient limit [16]. The temporal stereo algorithm improved on the conventional stretch correlation algorithm by making the epipolar search bands adaptive; calculating search band sizes and locations based on disparity information obtained from the previous image pair. Temporally seeding using disparity values, meant the algorithm could work with much smaller epipolar search bands. As the probability of a mismatch occurring is proportional to the area searched [18], the temporal algorithm worked very well once the stereo result had ‘grown’ to fit the objects in the scene correctly. Outliers caused by the stereo matcher incorrectly pairing left and right image features were reduced, and an additional benefit of reducing computational load was achieved, *typically accelerating the algorithm four fold*.

However, [3] also demonstrated that, on start-up, temporal stereo, like simulated annealing algorithms, could get stuck in local minima in the correlation search space, leading to outliers in the 3-D data. The proposed solution was to use a stochastic search element, which applied exhaustive searches to a small percentage of the image blocks being correlated. It was hoped that the stochastic search would allow the algorithm to escape local minima and converge on the correct solution. However, the method proved to be rather crude and unreliable, allowing the re-introduction of ambiguity leading to extra outliers and missing some previous correctly matched data.

3 Coarse-to-Fine Bootstrap Algorithm

The notion of using image data at various scales to overcome ambiguity and increase the reliability of stereo, in particular for correlation stereo algorithms, is not new. [12], [15] and [6], all illustrate stereo algorithms which established gross correspondences first,

moving onto finer scales to refine the result. [14] and [21] both use a translating camera(s) to provide a series of narrow to wide stereo baseline images, from which a CTF estimation of disparity for the scene was made. According to [15]; a CTF approach allows exhaustive searching with the ability to find an optimal matching point. This seemed like the ideal solution for providing the disparity information required during the start-up period for a temporal stereo vision system.

In order to achieve the initial exhaustive search that is required to give a temporal stereo algorithm a good seed point from which to start, the early frames in a stereo sequence can be subjected to an additional CTF image pre-processing stage. Processing the initial few frames at a much coarser scale (for example 30% to 50% of their original size) allows a correlation stereo algorithm to perform much larger epipolar band searches, but without increasing the computational load. Additionally, the temporal stereo algorithm suffers no extra loading because each stereo image pair is only processed once, at a single CTF scale. The CTF scale can then be increased with each successive frame until the 100% image size is reached, and the temporal algorithm can proceed as normal. The benefit of this approach is that a valid 3D result is available from the first frame, and, our analysis seems to support the fact that the result is more robust to outliers than the original temporal result processed at 100% image size. Certainly, this technique is far more robust than the stochastic search method used previously to give increased 'search range' without an excessive computational load.

Factors such as the disparity gradient can still be applied at all image scales: According to [1]; a change in the viewing distance scales both dot separation and disparity and, under experimentation, the disparity gradient seemed to remain constant over a wide variety of scales. This view is also supported by [16] which also says that the region of disparity gradient support should be set to be large enough to provide good disambiguation while not being so large that processing time is spent needlessly.

One of the other advantages of the CTF approach being used at the beginning of the sequence is that a reduction in scale removes some of the low level 'fine' structural detail leaving just the main scene structure. In doing this, we hope to remove some of the low level ambiguity which can often be present in a scene and concentrate on the top level scene structure, thus increasing the probability of matching correlation blocks correctly.

The difference between our application of CTF and the way it is applied in [12] for example, relates to the final objectives of the stereo algorithm being used. [12] outlines a scheme for using an adaptive correlation windowing system in which the correlation window's significant dimension is dependent on the local grey level variance. This is really only applicable to stereo algorithms trying to obtain dense estimates of disparity from the correlation of grey level data, whereas our temporal stereo algorithm is feature based, only returning disparity estimates for edge features. In this work, the CTF bootstrap is just used as a method of moving to the finest image scale (which will contain the richest edge information and produce the most accurate stereo 3-D result) as quickly and robustly as possible.

4 Complex Scene Evaluation

One of the long-standing problems in computer vision is the determination of algorithmic robustness. We would suggest that, not only is there still a large amount of

scepticism about the accuracy and reliability of stereo vision algorithms (leading to a lack up up-take by industry), but the lack of a consistent test-bed for algorithmic evaluation will mean that this situation will persist.

The development of our automatic evaluation tool for complex scenes was as the result of a desire to accurately quantify the performance of stereo algorithms for complex scenes. In [3], an automatic evaluation procedure was outlined to quantify the performance of stereo algorithms using images of planar data. Mismatches made by a stereo algorithm resulted in 3-D data points lying 'off' the plane in 3-D space, and hence an accurate outlier count could be given. However, unless a scene is sufficiently simple to be modelled using flat planes, then the only way to provide a rigorous evaluation of an algorithm's performance is to use images which have accompanying accurate ground truth data.

We have been using the POV-Ray [17] ray tracing package to produce fairly realistic scenes and hence stereo image sequences. Generating artificial images in this way allows precise control over the objects, colour, texture, motion, noise and lighting in a image. Having the POV-Ray image source code, also allowed us to write scene description files describing the precise geometry of each scene in our artificial image sequences. These scene description files could be interpreted by our complex scene stereo evaluation tool and used as ground truth data for assessing the performance of the temporal stereo algorithm.

The evaluation tool modelled the expected errors for the stereo algorithm as stated in [8]:

$$\sigma_Z = \frac{Z^2 \sigma_d}{If}$$

Equation 1, Standard Deviation, σ_Z, of the Error in Z at a Distance of Z

Where σ_Z is the standard deviation of the Z error at a distance Z away from the camera, I is the interocular separation distance, f is the focal length of the camera, and σ_d is the standard deviation of the expected disparity error. The disparity error in Equation 1 is calculated using Equation 2, In Equation 2, σ_d, is affected by the edge feature location accuracy σ_{ef}, the epipolar alignment accuracy, σ_{ep}, and the angle of the edge feature, θ, away from the horizontal.

$$\sigma_d = \sqrt{2\sigma_{ef}^2 + \frac{2\sigma_{ep}^2}{\tan^2 \theta}}$$

Equation 2, Standard Deviation, σ_d, of the Disparity Error

Each point in a stereo result was evaluated by calculating its theoretical σ_Z. If the point in 3-D space was consistent with the geometric scene description to within $3 \times \sigma_Z$, then it was accepted as a good stereo match (i.e. not an outlier).

5 Experiments

5.1 Artificial Scenes with Automatic Stereo Evaluation

The results from two artificial stereo image sequences are presented here; the first is a scene containing some translating cubes, and the second represents the view as seen by an autonomous vehicle moving through a room containing objects. Both image sequences contained homogenous white gaussian noise and shadows, however they did not contain lens distortion or focal blurring. Each of the stereo sequences was processed by the temporal stereo algorithm, once with and once without the benefit of CTF bootstrapping. The 3-D results were compared with the known ground truth data for the scenes using the new stereo evaluation tool and a number of statistics on the temporal algorithm's performance were obtained.

Figure 1, POVRay Cubes

Figure 2, Conventional Temporal Stereo

Figure 3, Temporal with Coarse-to-Fine Bootstrap

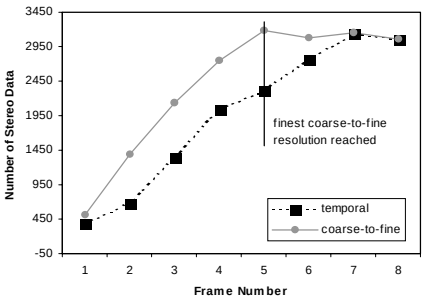

Figure 4, POVRay Cubes: Number of Stereo Data Returned

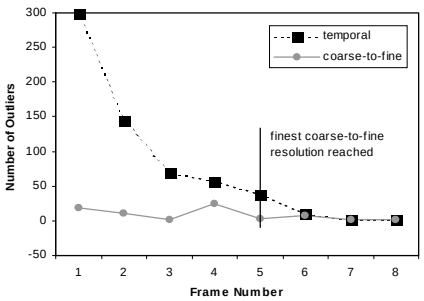

Figure 5, POVRay Cubes: Number of Outliers

The cubes sequence, Figure 1, was analysed using a conventional temporal stereo algorithm, **Figure 2**, and using the same temporal stereo algorithm but including CTF bootstrapping, **Figure 3**. (Initially, the temporal algorithm with CTF started with an image resolution of 30%, incrementing by 20% with each successive frame.)

Both Figure 2 and Figure 3 show the same stage in the cube sequence; frame 5, at which point the CTF scale had reached 100%. At this stage both the temporal and the CTF bootstrapped algorithms would have seen exactly the same input. In **Figure 3** however, all of the surfaces of the cubes are covered by the CTF stereo data, whilst the standard temporal algorithm, **Figure 2**, still had not managed to grow its result to cover the extreme far and near points in the scene.

For the temporal only algorithm, there were very few initial correct matches, with the rest of the correct data 'growing' out from the initial correct matches as the sequence progressed. However, with CTF bootstrapping, the stereo algorithm returned correct (if sparse) disparity data for all of the visible cube edges from the first frame onwards. Figure 4, shows the amount of stereo data returned by both algorithms over the sequence (the amount of edge data available in each frame in the sequence remained approximately constant). As the CTF result progressed, the amount of stereo data returned steadily grows because as the image size of the increases, so does the amount of edge data present. In the temporal only case, there is a similar steady growth in the amount of data returned, but this is due to mismatches being resolved and the 'growth' of the result over all areas of the scene. Figure 5, shows how the robustness of the CTF result remained very good throughout the sequence, with only a few (less than 1%) outliers, whereas the temporal only case has to resolve a large number of initially incorrect matches before finally reaching an outlier free result.

Visual inspection of the 3-D results revealed that, where CTF scaling had been used, there was a drop in the accuracy of the stereo result when compared with the best temporal only result (the same frame in the sequence but with image scale = 100%). This is caused by CTF scaling of the input stereo image pair, where the feature location accuracy is being lost by a factor of $\frac{100}{scale}$, where 0% < *scale* ≤ 100%. Using a 50% scaling factor, reduces the number of horizontal and vertical pixels in the image by 50%. However, the image is still defined over the same physical 'camera array', therefore the effective pixel size must have increased by a factor of $\frac{100}{50}$. Because the accuracy of the edge detection process does not change, for example σ_{edge} = 0.1 pixels, then the CTF feature location accuracy must now be; σ_{edge} × *new effective pixel size*, instead of the original σ_{edge} × *true physical pixel size*.

This theory is supported by plotting accuracy histograms for the error data supplied from the automatic evaluation tool. From **Equation 1**, the normalised Z error for a stereo point can calculated using;

$$\frac{\sigma_Z}{Z^2} \rightarrow \frac{\Delta Z}{Z^2}$$

Equation 3, Normalised Z Error

Where ΔZ is the measured Z error returned by the automatic evaluation tool. Normalised Z errors are independent of depth, Z, and remain constant for a particular stereo camera configuration (i.e. throughout a sequence). **Figure 6**, shows a typical normalised Z error histogram, from the cubes sequence. Measuring the σ of the histogram gives us a measure of how accurate the stereo result is. Using this method, we have been able to prove that CTF accuracy does appear to degrade by a factor of $\frac{100}{scale}$. The normalised Z errors can then be re-projected to any depth, Z, and the expected distribution of physical Z errors can be plotted. **Figure 7** shows the distribution of Z errors for the stereo algorithm at a distance of 3*I*, for two CTF scales, 30% and 100%, based on the re-projected normalized Z errors measured from the cubes sequence. In this case the σ of the 30% scale image is 3 times that of the 100% image, which compares favourably with a theoretical value of $\frac{100}{30} = 3.33$.

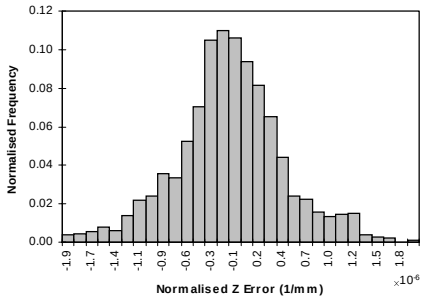

Figure 6, A Typical Normalised Z Error Histogram Distribution from the Cubes Sequence

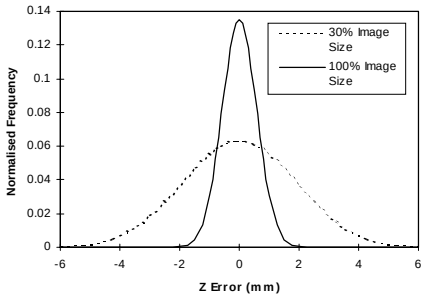

Figure 7, Theoretical Z Error Distribution for Two Coarse-to-Fine Image Scales

A very similar conclusion can be reached when looking at the results for the artificial room sequence, Figure 8.

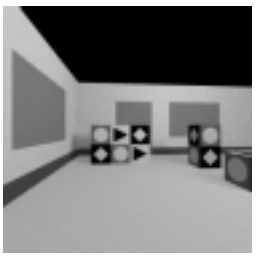

Figure 8, POVRay Room

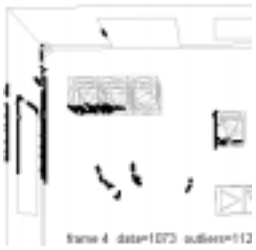

Figure 9, Conventional Temporal Stereo

Figure 10, Temporal with Coarse-to-Fine Bootstrap

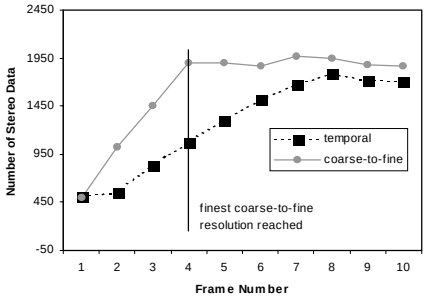

Figure 11, POVRay Room: Number of Stereo Data Returned

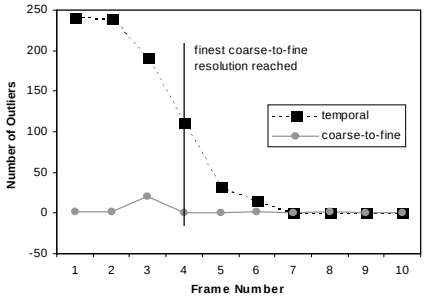

Figure 12, POVRay Room: Number of Outliers

This particular sequence involved moving the 'viewer' through an artificial room containing several obstacles at a range of depths. Once again, the CTF algorithm was able to 'size-up' the scene in the first frame of the sequence with the stereo evaluation tool reporting that very nearly all the data was correctly matched with less than 1% outliers. For the room sequence, the 100% image size for the CTF algorithm was reached at frame 4, and the 3-D results for that frame can been seen in Figure 9 and Figure 10. There are still a large number of outliers present in the conventional temporal

result, Figure 9. Indeed, for this particular sequence, the temporal result never manages to 'catch-up' to the CTF bootstrapped result entirely, see **Figure 11**, although most of the outliers have disappeared by frame 7, see **Figure 12**.

5.1 Real Scenes

The CTF bootstrap algorithm's performance on real image sequences also proved to be very good. Two sequences were tested, Figure 13 shows an image from a sequence of real translating cubes, and Figure 16 shows an image from a sequence of a train moving along a track.

Figure 13, Real Cubes

Figure 14, Conventional Temporal Stereo: Frame 1

Figure 15, Temporal with Coarse-to-Fine Bootstrap: Frame 1

In the case of the cubes, the coarse processing of the first frame was able to cover enough of the disparity search range to return data over almost all of the cubes' surfaces. The CTF bootstrapped stereo algorithm was then able to track all of the cubes throughout the sequence without producing any noticeable gross outliers. For the conventional temporal algorithm, just as with the synthetic cubes sequence, the algorithm had to grow the correct result over several frames (approximately 6), before it eventually caught up with the CTF bootstrapped algorithm.

Figure 16, Real Train

Figure 17, Conventional Temporal Stereo: Frame 1

Figure 18, Temporal with Coarse-to-Fine Bootstrap: Frame 1

For the train sequence, there was an apparent complete lack of outliers even at the most coarse resolution (50%). We have attributed this to the fact that most of the outliers in the conventional algorithm came from matching the fine detail on the side and top of the train. At the coarse scale, these fine details were lost and the algorithm was able to match more robustly the major structural features on the train and the track. This gave the CTF temporal algorithm an excellent seed point from which to start, and was able to process the whole sequence with very few outliers. The temporal only algorithm on the

other hand initially contained several large groups of outliers, which were all eventually all resolved over the first 5 frames.

One issue that does arise when using CTF with temporal stereo, is how to choose the best starting scale and rate of change for the CTF scale parameter. From our experience of using the CTF bootstrapping, a good starting scale for the CTF was around 30% to 50% of the original image size. However, the selection of an appropriate starting scale will depend upon the composition of the scene being viewed, and the size of the epipolar search bands used in the temporal stereo algorithm. As far as the rate of increase for the CTF scale goes, this also appeared to be a fairly arbitrary figure. Indeed for many of our test scenes, experiments have shown that often just a single frame processed at a coarse scale was sufficient to seed the temporal matching at the finest image resolution. However, we suspect there may be a trade off in the rate of CTF scale increase, and the amount of searching (computational load) that the temporal algorithm has to do in processing subsequent frames. Increasing the CTF scale too quickly, may result in gross disparities from a large area of the previous stereo result being used to seed searches on much finer scaled images, leading to large epipolar search bands being used unnecessarily. Gradually introducing the finer scaled images, should minimise this effect, and, for scenes containing large depths of field, give the temporal algorithm a chance to refine and resolve stereo matches at the coarser scales before moving onto the finer image scales.

6 Conclusions

Any search-band-based stereo correspondence algorithm will be sensitive to the parameter range selected, which cannot be known a priori. Temporal matching schemes, making use of simple consistency, alleviate this problem so that after some period of time the correct solution will be achieved regardless of the initial assumptions. In this work we have demonstrated that a coarse-to-fine bootstrap, preceding temporal stereo, will allow the matching process to home in on the best set of matches quickly and as robustly as possible. Also, bootstrapping the temporal matching process in this way should not have an effect on the execution speed of a stereo vision algorithm because the re-scaling of the input images could be done at the same time as the pre-processing stage of the stereo algorithm itself.

A statistical model of the theoretical accuracy of the stereo algorithm has been used, together with an autonomous evaluation tool for stereo vision to prove the robustness of the coarse-to-fine stereo bootstrapping process. The evaluation test-bed has also demonstrated the viability of benchmarking the performance of complex stereo algorithms for difficult scenes, providing they can be modelled successfully.

If you would like to evaluate our temporal stereo algorithm including the coarse-to-fine bootstrapping, it is available, together with the stereo evaluation tool, from:

`http://www.shef.ac.uk/~eee/esg/research/tina.html`

References

[1] Burt P. and Julesz B. Modifications of the Classical Notion of Panum's Fusional Area. *Perception*, 1980, Vol. 9, pp. 671-682.

[2] Barnard S.T. and Thompson W.B. Disparity Analysis of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 1980, Vol. 2, No. 4, pp. 333-340.

[3] Crossley S. Lacey A.J. Thacker N.A. and Seed NL. Robust Stereo via Temporal Consistency. *Proc. of the British Machine Vision Conference*, 1997, pp. 659-668.

[4] Forstner W. 10 Pros and Cons Against Performance Characterisation of Vision Algorithms. *1996 EPSRC Computer Vision Summer School*, University of Surrey, Guildford.

[5] Forstner W. Diagnostics and Performance Evaluation in Computer Vision. *Proc. Performance versus Methodology in Computer Vision*, 1994, pp. 11-25.

[6] Grimson W.E.L. Computational Experiments with a Feature Based Stereo Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1985, Vol. 7, No. 1, pp. 17-34.

[7] Haralick R.M. Performance Characterisation Protocol in Computer Vision. *CVGIP Image Understanding*, 1994, Vol. 60, No. 2, pp. 245-249.

[8] Harris A.J. Thacker N.A. and Lacey A.J. Modelling Feature Based Stereo Vision for Range Sensor Simulation. *Proc. of the European Simulation Multiconference*, June 1998, pp. 417-421.

[9] Ho A.Y.K. and Pong T.C. Cooperative Fusion of Stereo and Motion. *Pattern Recognition*, January 1996, Vol. 29, No. 1, pp. 121-130.

[10] Hung Y.P. Tang C.Y. Shih S.W. Chen Z. and Lin W.S. A 3D Predictive Visual Tracker for Tracking Multiple Moving Objects with a Stereo Vision System. *Lecture Notes in Computer Science*, 1995, Vol. 1024, pp. 25-32.

[11] Lane R.A. Thacker N.A. and Seed N.L. Stretch Correlation as a Real Time Alternative to Feature Based Stereo Matching Algorithms. *Image and Vision Computing*, 1994, Vol. 12, No. 4, pp. 203-212.

[12] Levine M.D. O'Handley D.A. and Yagi G.M. Computer Determination of Depth Maps. *Computer Graphics and Image Processing*, 1973, Vol. 2, No. 2, pp. 131-150.

[13] Marik R. Kittler J. and Petrou M. Error Sensitivity Assessment of Vision Algorithms Based on Direct Error Propagation. *1996 EPSRC Computer Vision Summer School*, University of Surrey, Guildford.

[14] Matthies L. and Okutomi M. Bootstrap Algorithms for Dynamic Stereo Vision. *Proc. of the 6th Multidimensional Signal Processing Workshop*, 1989, p. 12.

[15] O'Neill M. and Denos M. Automated System for Coarse to Fine Pyramidal Area Correlation Stereo Matching. *Image and Vision Computing*, 1996, Vol. 14, pp. 225-236.

[16] Pollard S.B. Mayhew J.E.W. and Frisby J.P. PMF: A Stereo Correspondence Algorithm using a Disparity Gradient Limit. *Perception*, 1985, Vol. 14, pp. 449-470.

[17] Persistence of Vision (tm) Ray-Tracer (POV-Ray (tm)), *http://www.povray.org*.

[18] Thacker N.A. and Courtney P. Statistical Analysis of a Stereo Matching Algorithm. *Proc. of the British Machine Vision Conference*, 1992, pp. 316-326.

[19] Wang W. and Duncan J.H. Recovering the Three Dimensional Motion and Structure of Multiple Moving Objects from Binocular Image Flows. *Computer Vision and Image Understanding*, May 1996, Vol. 63, No. 3, pp. 430-446.

[20] Waxman A.M. and Duncan J.H. Binocular Image Flows: Steps Towards Stereo-Motion Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, November 1986, Vol. 8, No. 6, pp. 715-729.

[21] Xu G. Tsuji S. and Asada M. A Motion Stereo Method Based on Coarse to Fine Control Strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1987, Vol. 9, No. 2, pp. 332-336.