Using temporal seeding to constrain the disparity search range in stereo matching

Thulani Ndhlovu

Mobile Intelligent Autonomous Systems
CSIR
South Africa
Email: tndhlovu@csir.co.za

Fred Nicolls

Department of Electrical Engineering
University of Cape Town
South Africa
Email: fred.nicolls@uct.ac.za

***Abstract*—In a stereo image sequence, finding feature correspondences is normally done for every frame without taking temporal information into account. Reusing previous computations can add valuable information. A temporal seeding technique is developed for reusing computed disparity estimates on features in a stereo image sequence to constrain the disparity search range. Features are detected on a left image and their disparity estimates are computed using a local-matching algorithm. The features are then tracked to a successive left image of the sequence and by using the previously calculated disparity estimates, the disparity search range is constrained. Errors between the local-matching and the temporal seeding algorithms are analysed. Results show that although temporal seeding suffers from error propagation, a decrease in computational time of approximately 20% is obtained when it is applied on 87 frames of a stereo sequence. Furthermore a confidence measure is used to select start-up points for seeding and experiments are conducted on features with different confidences.**

I. Introduction

In a stereo image sequence, finding feature correspondences is normally done for every pair of frames without taking temporal information into account. Current imaging sensors can acquire images at high frequencies resulting in small movements between consecutive image frames. Since there are small inter-frame movements, the frame-to-frame disparity estimates do not change significantly. We explore using previously computed disparity estimates to seed the matching process of the current stereo image pair.

Most conventional stereo correspondence algorithms contain fixed disparity search ranges by assuming the depth range of the scene. This range stays constant throughout a stereo image sequence. By decreasing the disparity search range the efficiency of the matching process can be improved. This can be seen in [1], where the probability of an incorrect stereo match is given by:

$$P^T \propto P^a + P^b + P^c,$$

where P^a is the probability of mismatching a pair of features when neither feature has its correct match detected in another image, P^b is the probability of mismatch when one feature has had its correct match detected, and P^c is the probability of mismatch when both features have had their correct matches found in the other image. The probabilities, P^a, P^b and P^c are all proportional to the mean number of candidate matches and thus P^T is proportional to the disparity search range of the stereo correspondence algorithm. Therefore by reducing the disparity search range, P^T is reduced assuming that the correct match remains in the reduced range.

A method of using temporal information in stereo image sequences to decrease the disparity search range so as to decrease the probability of a mismatch is developed. A local-matching stereo correspondence algorithm is implemented on KLT (Kanade Lucas Tomasi) features and the disparity estimates obtained are used on the consecutive stereo image frame to seed the matching process. Local-matching stereo algorithms have a uniform structure. This allows the temporal seeding method to be used across different variations of these stereo algorithms. The method is expected to be at the least as accurate as the local-matching algorithm at a lower computational expense when using a small number of frames.

Errors in both the local-matching algorithm and the temporal seeding algorithm are quantified. The algorithms are run on 87 frames of a dataset and temporal seeding is evaluated.

This paper is structured as follows. Section II covers the related literature. Section III defines the problem to be solved. Section IV discusses the feature matching and detection process. The local-matching stereo correspondence algorithm implemented is discussed in Section V. Section VI discusses the temporal seeding process. Section VII discusses the experiments and results and conclusions are made in Section IX.

II. Related work

There have been successful implementations of enforcing temporal constraints for depth estimation in successive stereo image frames. The work in this chapter is directly related to approaches that combine motion and stereo.

Algorithms such as [2]–[4] can be classified as pseudo-temporal stereo vision algorithms. They aim to solve the stereo correspondence problem for a wide baseline. The input to such algorithms is a monocular sequence of images produced by a camera undergoing controlled translating or rotating motion. Stereo image pairs are produced by selecting and pairing different images from the input sequence. The algorithms initially start by finding correspondences in a short baseline stereo pair and use these correspondences to bootstrap the stereo correspondences of a wider baseline stereo image pair. The propagation of disparity values from one set of frames to the next helps to improve computational efficiency and reliability of stereo matching.

The work in [5] and [6] also takes into account temporal information to solve for depth from triangulation. These methods extend support aggregation from 2D to 3D by adding the temporal domain. These algorithms can be viewed as exploiting temporal aggregation to increase matching robustness.

The work in [7] uses previous disparity estimates by analyzing their local neighbourhood to decrease the disparity search range of the current stereo image frame. The computational load and robustness of using temporal information are demonstrated. Although this algorithm performs well, it suffers from start-up problems. Research from [8] addressed the start-up problem and was successfully used on a wide baseline stereo sequence.

The ideas in this paper are similar to [7], but instead of analyzing the local neighbourhood of the previous estimate to decrease the search range, the search range is decreased according to the shape of the correlation curve in the successive image frame given an initial disparity estimate.

III. Problem Statement

The problem to be solved can be defined as follows.

The input is a set of calibrated and rectified stereo image pairs, $\{\mathcal{L}_{\mathrm{t}}, \mathcal{R}_{\mathrm{t}}\}_{t=0}^{N}$, each pair acquired at time $t = 0, ..., N$. $\mathcal{L}_t$ and $\mathcal{R}_t$ denote the left and the right images taken at time t. An image coordinate, $\mathbf{x}_t = (u_t, v_t) \in \mathcal{F}$, represents a pixel location of a detected feature in the set of features $\mathcal{F}$ at row u_t and column v_t in the left image, while $\mathbf{x}'_t = (u'_t, v'_t)$ represents the corresponding pixel on the right image. If $d(\mathbf{a})$ is the disparity estimate of pixel $\mathbf{a}$, then our goal is to determine a disparity estimate $d(\mathbf{x}_{t+1})$ when given $d(\mathbf{x}_t)$ as a prior.

IV. Feature Detection and Matching

A. KLT (Kanade-Lucas-Tomasi)

This section discusses the feature detector and tracker used in determining and tracking the features $\mathcal{F}$. The Kanade-Lucas-Tomasi (KLT) feature tracker is based on the early work by Lucas and Kanade (LK) on optical flow [9]. The LK algorithm attempts to produce dense disparity estimates. The method is easily applied to a subset of points in the image and can be used as a sparse technique. Using the LK algorithm in a sparse context is allowed by the fact that it relies on local information extracted from some small window surrounding the point of interest.

LK works on three assumptions:

- *Brightness consistency*
  The appearance of a pixel does not change with time. This means that the intensity of a pixel denoted as $I(.)$ is assumed to remain constant between image frames: $I_t(\mathbf{x}_t) = I_{t+1}(\mathbf{x}_{t+1})$.
- *Temporal persistence*
  The movement of the image frames is small, so a surface patch changes slowly over time.
- *Spatial coherence*
  Neighbouring pixels that belong to the same surface patch have similar motion and project to nearby image points on the image plane.

The KLT tracker [10] is a frame-by-frame temporal tracker for a single video sequence which is applied at a set of corner points that change throughout the tracking. Good features [11] are selected by examining the minimum eigenvalue of each 2×2 gradient matrix. The features are then tracked using a Newton-Raphson method of minimizing the differences between two image patches. Figure 1 shows an image with 313 detected KLT features.

The features in $\mathcal{F}$ are found by using [11] and these features are tracked on successive frames using the KLT tracker in order to estimate $\mathbf{x}_t \leftrightarrow \mathbf{x}_{t+1}$.

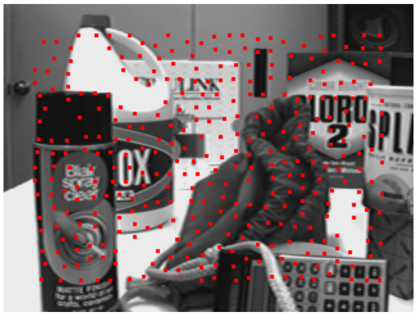


Fig. 1. Features detected in the image. 313 KLT features are detected on this image.

V. Stereo correspondence

Given an image pair that is calibrated and rectified, $\{\mathcal{L}_t, \mathcal{R}_t\}$, a set of detected features $\mathcal{F}$ in the left image, and corresponding pixels $\mathbf{x}_t \leftrightarrow \mathbf{x}'_t$, the objective is to determine $d(\mathbf{x}_t)$.

A. Stereo algorithm

The stereo algorithm implemented uses Birchfield and Tomasi's (BT's) sampling insensitive matching cost [12]. This matching cost $e(\mathbf{x_t}, \mathbf{x'_t})$ is formulated as follows:

$$I_l(\mathbf{x}'_t) = \frac{1}{2}(I_R(\mathbf{x}'_t) + I_R(u'_t, v'_t - 1))$$

is the linearly interpolated intensity to the left of pixel $\mathbf{x}'_t$ and, analogously,

$$I_r(\mathbf{x}'_t) = \frac{1}{2}(I_R(\mathbf{x}'_t) + I_R(u'_t, v'_t + 1))$$

is to the right of pixel $\mathbf{x}'_t$, then $I_{min}(\mathbf{x}'_t)$ and $I_{max}(\mathbf{x}'_t)$ are computed as follows:

$$I_{min}(\mathbf{x}'_t) = \min(I_l(\mathbf{x}'_t), I_r(\mathbf{x}'_t), I_R(\mathbf{x}'_t)),$$

$$I_{max}(\mathbf{x}'_t) = \max(I_l(\mathbf{x}'_t), I_r(\mathbf{x}'_t), I_R(\mathbf{x}'_t)).$$

The matching cost is then computed as

$$e(\mathbf{x}_t, \mathbf{x}'_t) = \max(0, I_L(\mathbf{x}_t) - I_{max}(\mathbf{x}'_t), I_{min}(\mathbf{x}'_t) - I_L(\mathbf{x}_t)). \tag{1}$$

A square window is used to aggregate the cost [13]. Furthermore, sub-pixel interpolation is performed for disparity refinement [14]. From this point, the described stereo algorithm will be called BT's stereo algorithm.

VI. Temporal seeding

The main contribution of this work lies in the way the computed $d(\mathbf{x}_t)$ is reused as a prior to determine $d(\mathbf{x}_{t+1})$. In order to achieve the objective, some important assumptions have to be made and justified.

A. Assumptions

The local structure of successive stereo disparity maps does not change significantly. This means that one can assume that the shape of the correlation curve, which is the correlation of the error over the disparities, of a tracked feature point at time $t+1$ does not change by much compared with the feature point's correlation curve at time t. This assumption is made feasible because of the LK algorithm's three assumptions, as stated in Section IV-A. Figure 2 shows an example where the assumption holds. The shapes and positions of the two correlation curves are almost identical despite the frame-to-frame movement.

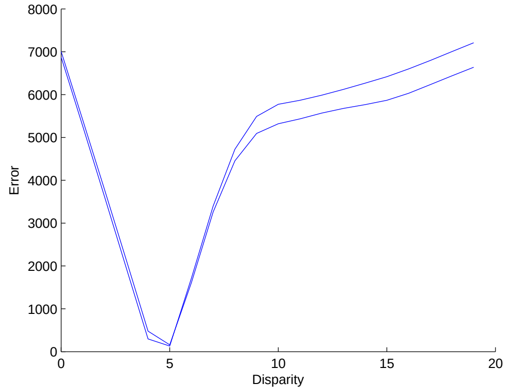


Fig. 2. Correlation curves for a feature point in $\mathcal{F}$ at time t and $t+1$.

B. Seeding

The disparity estimate $d(\mathbf{x}_t)$ is used as an initial disparity estimate for $d(\mathbf{x}_{t+1})$. The new disparity estimate is found by the local minimum around the initial estimate. This local minimum is then assumed to be the global minimum of the correlation curve. A technique similar to gradient descent is used to determine the local minimum. The gradients to the left and to the right of $d(\mathbf{x}_t)$ are determined. If the gradients on both sides of $d(\mathbf{x}_t)$ are negative, then we move a single step towards the descending slope until the gradient changes the sign from negative to positive. This signals the local minimum around the initial point. A similar process is carried out if the gradients on both sides of $d(\mathbf{x}_t)$ are positive.

VII. Experiments and results

A. Ground truth

Since features are tracked along the left images of a stereo sequence, one needs ground truth disparity estimates to evaluate the implemented stereo correspondence algorithm and the potential of temporal seeding. Ground truth is determined by first using BT's stereo algorithm on each feature point of a stereo pair. The disparity estimates are then refined to subpixel accuracy by using the LK optical flow method.

B. Experiments

In this section we evaluate temporal seeding. The dataset used is the Microsoft i2i chairs dataset[1] which consists of a stereo video of an indoor scene with chairs and tables. Figure 3 shows a stereo image pair of the dataset.

The experiments carried out in this section are on 87 successive stereo image frames. In the first stereo pair, feature points are detected and the disparity estimates of the features are calculated with BT's stereo algorithm. The disparity search range used is $[0, 20]$. Features are then tracked on the left images of the sequence and temporal seeding is applied to determine the disparity estimates in the range of $[0, 20]$ on

[1] http://research.microsoft.com/en-us/projects/i2i

(a)

(b)

Fig. 3. First stereo image pair of the Microsoft i2i chairs dataset.

the successive stereo image pairs. To quantitatively evaluate the results, the $RMSE$ is computed as follows:

$$RMSE = 100 \times \sqrt{\frac{1}{N_f} \sum_{f \in \mathcal{F}} (d_f - d_g(f))^2}, \tag{2}$$

where f is a detected feature point in the set $\mathcal{F}$, N_f is the number of detected feature points, d_f is the estimated disparity of a feature, and d_g is the ground truth disparity of the detected feature.

The temporal seeding algorithm is initialised by the disparity estimates of BT's stereo algorithm on the first frame of the image sequence. Temporal seeding is then performed on the successive frames. Seed values for frame number k are taken from frame number $k-1$.

Figure 4 shows the number of features which were successfully tracked along the stereo sequence. Features which were not successfully tracked are replaced by new features. Although some features are replaced, features which lie close to the border of the image are removed. Further, features with a low reliability of being tracked are also removed. This causes the number of features to drop as the number of frames increase.

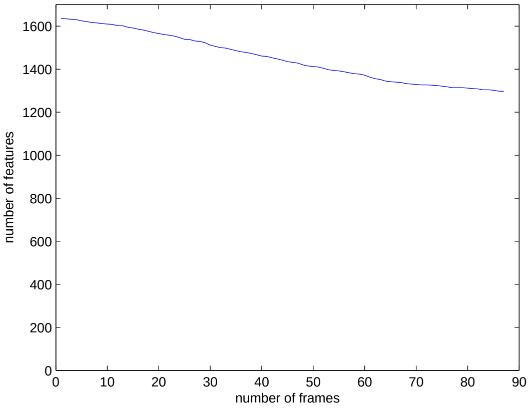


Fig. 4. Number of frames versus the number of features in temporal seeding.

Figure 5 shows the $RMSE$ for every stereo frame in the sequence. The results show that the error increases as the number of frames increase. This means that temporal seeding suffers from error propagation. The results also show that the error propagation rate decreases as the number of frames increase. This is caused by the fact that features which cannot be tracked are replaced by new features. As the sequence progresses, more features fall away because parts of the scene fall out of the field of view of the camera. Furthermore, the lighting of the scene changes because of the change in viewpoint. Some of the features which fall away are erroneous and are replaced with new features which have not suffered from error propagation. The computational time when using temporal seeding on the stereo sequence is approximately 20% less than that of BT's stereo algorithm. The improvement in speed is achieved because the developed temporal seeding approach does not always do the full disparity search. It only searches for a local minimum around the initial estimate. This is demonstrated in Figure 6. The figure shows the average number of disparity values visited by the temporal seeding algorithm for every image frame. The average number of disparities visited are less than the disparity range hence the speed improvement.

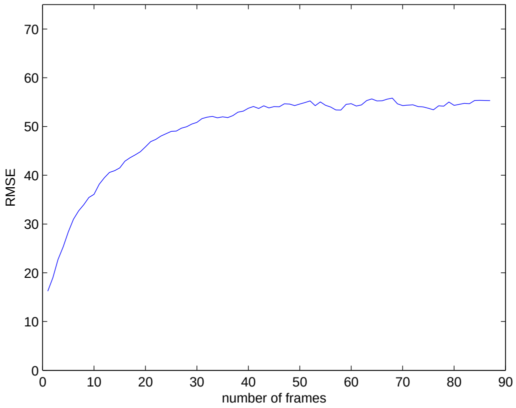


Fig. 5. Number of frames versus $RMSE$ in temporal seeding.

VIII. Confidence measure in temporal seeding

In the previous section, temporal seeding is applied on a stereo image sequence. The startup disparities for temporal seeding are calculated using a full disparity search on the KLT features of the left image in the first stereo image frame. In this section, the confidences of the detected features in the first stereo image frame are determined. The features are then selected according to their confidences and used as the startup features. Temporal seeding is then applied and the results are evaluated.

The confidence measure is calculated as a function of

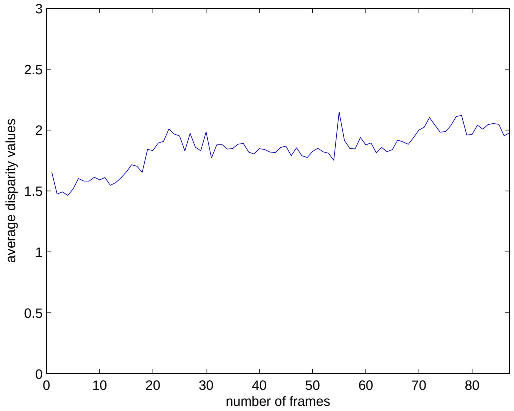


Fig. 6. Number of frames versus average disparities visited by the temporal seeding algorithm.

(u, v, d), where (u, v) are the image coordinates and d is the disparity. A typical correlation curve is shown in Figure 7. Local-matching algorithms aim to find the disparity that minimizes the error represented by this curve. Given a disparity d, we propose computing the confidence of a disparity estimate as follows:

$$C_d = \frac{B(d)}{d_{max} - d_{min}}. \tag{3}$$

Here C_d is the confidence for a given disparity, $B(d)$ is the basin of convergence (refer to Figure 7) of the disparity estimate d, and $d_{max} - d_{min}$ is the disparity range. It is expected that in textureless regions the correlation curve will have multiple local minima with small $B(d)$ values, and since C_d is proportional to $B(d)$ we expect low confidences. A high confidence value would have few local minima in the correlation curve and a fully confident disparity estimate would arise where the local minimum is the global minimum of the correlation curve.

The value of $B(d)$ is determined by using an approach similar to gradient ascent. Given a disparity estimate d, the gradient to the right of d is expected to be positive and the gradient to the left of d is expected to be negative. The algorithm takes single steps on both sides of d until the sign of the gradient changes. This represents the local maxima on the left and on the right of d. The disparity range covered by the two local maxima is defined as the basin of convergence $B(d)$.

The results of the experiments are shown below. Figure 8 shows the number of features which satisfy a particular confidence. For the same reason as in Section VII-B, the number of tracked features decreases as the number of frames progresses. Also, the number of startup features are high for low confidence values and low for high confidence values.

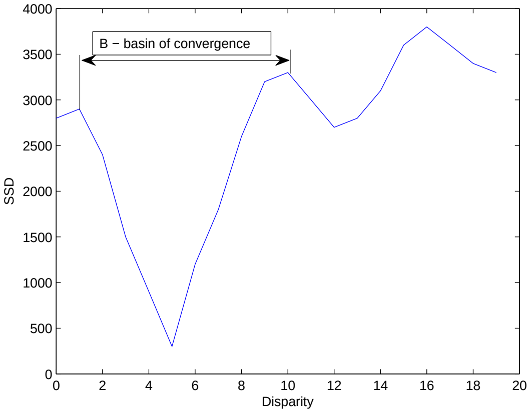
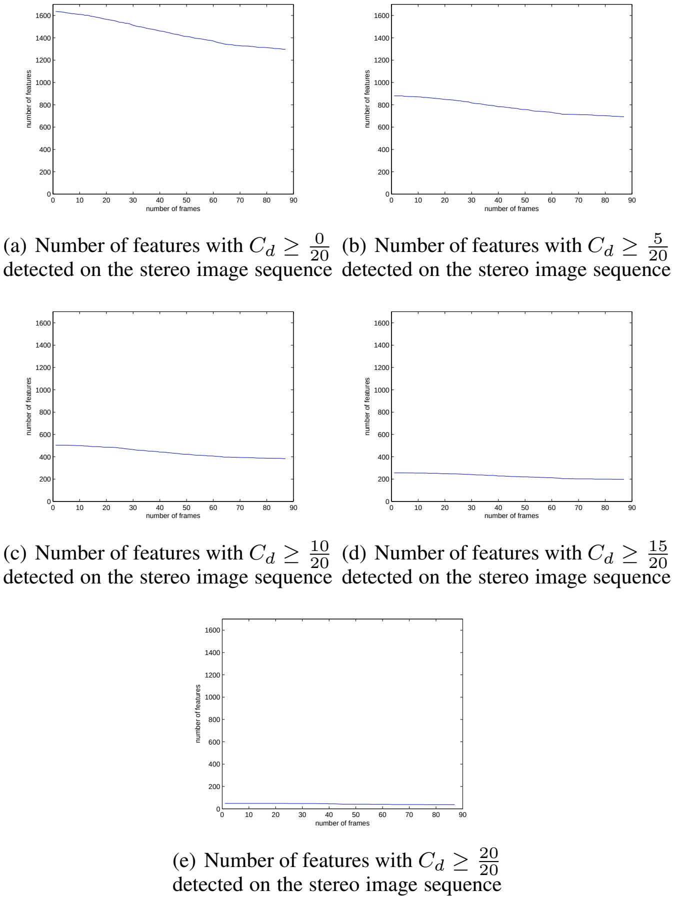


Fig. 7. Correlation curve and the basin of convergence.

Figure 9 shows the $RMSE$ for every frame when using startup disparities with a particular confidence. The results show that the shapes of the Figures 9(a)-9(e) are similar meaning that the confidence measure is not removing enough erroneous points in temporal seeding. The peak error becomes



(a) Number of features with $C_d \geq \frac{0}{20}$ detected on the stereo image sequence

(b) Number of features with $C_d \geq \frac{5}{20}$ detected on the stereo image sequence

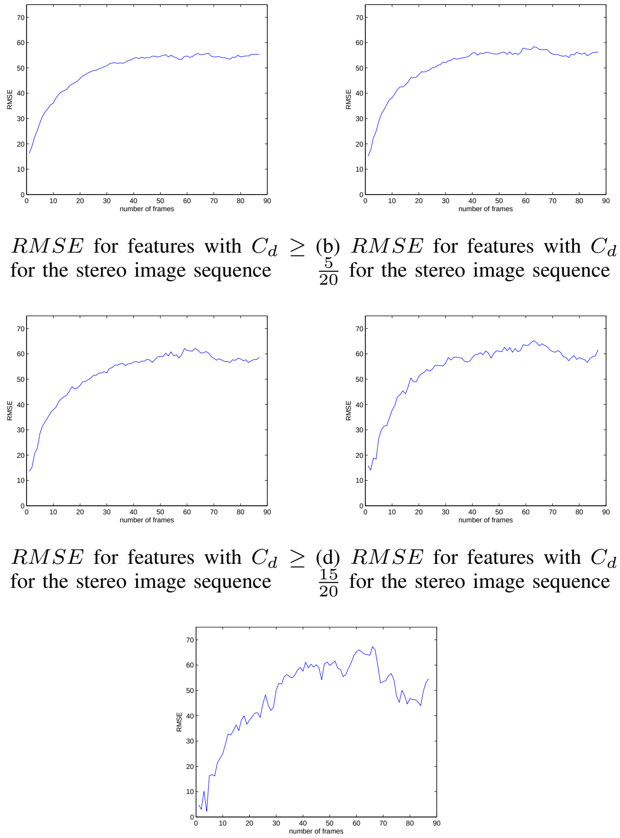
(c) Number of features with $C_d \geq \frac{10}{20}$ detected on the stereo image sequence

(d) Number of features with $C_d \geq \frac{15}{20}$ detected on the stereo image sequence

(e) Number of features with $C_d \geq \frac{20}{20}$ detected on the stereo image sequence

Fig. 8. Number of features detected on the stereo image sequence which have a chosen confidence.

higher as C_d increases. This is caused by the fact that features with a higher C_d value are tracked for longer in the image sequence. This means the error is propagated longer in the sequence leading to a higher peak error. Figure 9(e) appears to have more noise compared to the other plots. This is due to the low number of features which have a confidence value of 1.



(a) $RMSE$ for features with $C_d \geq \frac{0}{20}$ for the stereo image sequence

(b) $RMSE$ for features with $C_d \geq \frac{5}{20}$ for the stereo image sequence

(c) $RMSE$ for features with $C_d \geq \frac{10}{20}$ for the stereo image sequence

(d) $RMSE$ for features with $C_d \geq \frac{15}{20}$ for the stereo image sequence

(e) $RMSE$ for features with $C_d \geq \frac{20}{20}$ for the stereo image sequence

Fig. 9. $RMSE$ versus the number of frames for different confidence values.

IX. Conclusion

A method of reusing computed disparity estimates of KLT features in a stereo image sequence is presented. We have shown that using temporal seeding on 87 frames of a stereo sequence produces a computational improvement of approximately 20%. However, temporal seeding suffers from error propagation.

Further experiments involve using a confidence measure in temporal seeding. The confidence measure is then used to select different start-up features for temporal seeding in the attempt of limiting error propagation. The results show that the confidence measure does not succeed in removing features which produce high errors.

Acknowledgment

The author thanks the Council for Scientific and Industrial Research (CSIR) and the Mobile Intelligent Autonomous Systems (MIAS) group for their support on this work.

References

[1] N. Thacker and P. Courtney, "Statistical analysis of a stereo matching algorithm," in *Proceedings of the British Machine Vision Conference*, 1992, pp. 316–326.

[2] L. Matthies and M. Okutomi, "Bootstrap algorithms for dynamic stereo vision," in *Proceedings of the 6th Multidimensional Signal Processing Workshop*, 1989, p. 12.

[3] A. Dalmia and M. Trivedi, "High speed extraction of 3d structure of selectable quality using a translating camera," *Computer Vision and Image Understanding*, vol. 64, no. 1, pp. 97–110, 1996.

[4] S. T. G. Xu and M. Asada, "A motion stereo method based on coarse to fine control strategy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 2, pp. 332–336, 1987.

[5] B. C. L. Zhang and S. Seitz, "Spacetime stereo: Shape recovery for dynamic scenes," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003, pp. 367–374.

[6] R. R. S. R. J. Davis, D. Nehab, "Spacetime stereo : A unifying framework for depth from triangulation," *IEEE Transaction On Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, 2005.

[7] N. T. S. Crossley and N. Seed, "Benchmarking of bootstrap temporal stereo using statistical and physical scene modelling," in *Proceedings of the British Machine Vision Conference*, 1998, pp. 346–355.

[8] N. T. S. Crossley, A.J. Lacey and N. Seed, "Robust stereo via temporal consistency," in *Proceedings of the British Machine Vision Conference*, 1997, pp. 659–668.

[9] D. B. L. Kanade and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *International Joint Conference on Artificial Intelligence*, 1981, pp. 674–679.

[10] C. Tomasi and T. Kanade, "Detection and tracking of point features," Tech. Rep., 1991.

[11] J. Shi and C. Tomasi, "Good features to track," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 593–600.

[12] S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 4, pp. 401–406, 1998.

[13] L. Wang, M. Gong, M. Gong, and R. Yang, "How far can we go with local optimization in real-time stereo matching," in *Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 129–136.

[14] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.