An Energy Minimisation Approach to Stereo-Temporal Dense Reconstruction

Carlos Leung, Ben Appleton, Brian C. Lovell
IRIS Group, ITEE
The University of Queensland
[cleung, appleton, lovell]@itee.uq.edu.au

Changming Sun
Mathematical and Information Sciences
CSIRO Australia
changming.sun@csiro.au

Abstract

We propose a novel energy minimisation framework for the dense reconstruction of stereo image sequences that incorporates data fidelity as well as spatial and temporal regularity. An iterated dynamic programming scheme is proposed to minimise the energy function. We also present an efficient implementation of the minimisation scheme by introducing morphological decomposition techniques to solve the dynamic programming subproblem. Our proposed method is capable of reconstructing dynamic scenes with complex motion. Results are presented demonstrating the strength of our proposed algorithm.

1 Introduction

Dense 3D reconstruction from stereo images has been and continues to be an active area of research in the computer vision community. Two recent review papers in this area highlight the advancements and the large amount of research being undertaken [4, 7]. Although there is strong support that the incorporation of temporal information can achieve better results [5], only a small amount of research has been devoted to the reconstruction of dynamic scenes from stereo image sequences.

A number of researchers have incorporated temporal coherence into the dense stereo reconstruction process by optical flow [2, 12]. A disadvantage of these methods is that the matching analysis is performed only in a local sense, and therefore does not cope well in scenes with multiple moving objects, low texture, and significant occlusions [8]. Similar methods include stereo-temporal reconstruction algorithms that employ Kalman filtering [6] or tracking of lines and edges between images [8]. This class of methods estimates the depth through accurately recovering the complete motion of the whole 3D scene.

However a disparity sequence requires significantly less description than a vector optical flow field. Disparity estimation is concerned solely with the surface geometry and does not rely upon the tangential motion of surface points. Consider for example a scene of a sphere rotating about its axis. While optical flow methods must track the tangential motion of the sphere's surface before determining that the depth does not change, disparity methods are invariant in this case. Therefore disparity methods are only related to the normal component of surface motion. We propose an energy minimisation scheme that provides a simpler model for the dense reconstruction of stereo image sequences.

Zhang *et al.* also described an energy minimisation framework to reconstruct stereo sequences using structured light [13]. Although their energy function only includes a data fidelity term, their formulation allows the reconstruction of dynamic scenes. Strecha and van Gool on the other hand incorporated temporal coherence into their partial differential equation formulation but are restricted to the stereo-temporal reconstruction of static scenes [9].

In this paper, we present an energy minimisation framework for the dense reconstruction of dynamic scenes that incorporates both data fidelity and regularisation terms. Since the global minimisation of this class of energy functions has been demonstrated to be NP-hard [3], we describe the application of an iterated dynamic programming algorithm to obtain a strong minimum. A morphological decomposition method is also proposed for the fast solution of the dynamic programming subproblems. Experimental results are presented verifying the benefits of incorporating temporal regularity into our energy minimisation framework for stereo-temporal reconstruction.

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

2 Stereo-Temporal Reconstruction

2.1 Energy minimisation framework

The goal of stereo-temporal reconstruction is to compute a disparity function $d(x, y, t)$ which describes the changing depth of a scene over time. This can be achieved by casting the scene's motion into an energy minimisation framework, whose minima correspond to likely reconstructions.

In traditional stereo reconstruction, structural information is incorporated by imposing spatial regularity, modelling the expected smoothness of surfaces in the scene. Stereo image sequences display a similar regularity in the time domain. Objects tend to move and deform smoothly, so penalising discontinuities between successive frames has the potential to further improve the accuracy of depth estimation. Here we extend a previous stereo energy minimisation method [1] to incorporate temporal regularity in addition to spatial regularity.

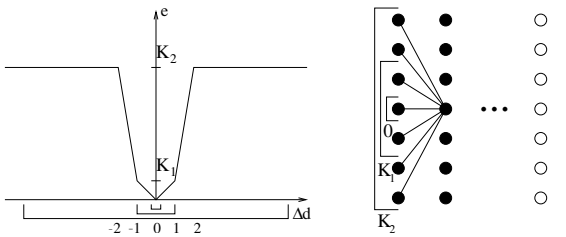
Consider the following first-order energy function for stereo sequence matching,

$$E[d] = \sum_{(\vec{x},t)} c(\vec{x}, t, d) + \frac{1}{2} \sum_{(\vec{x}_1,t_1)\sim(\vec{x}_2,t_2)} e(|d(\vec{x}_1, t_1) - d(\vec{x}_2, t_2)|), \quad (1)$$

where $a \sim b$ implies that a neighbours b and $\vec{x} \equiv (x, y)$. The first term enforces data fidelity, with the cost function c quantifying the matching quality. In stereo reconstruction c is typically chosen as the zero-mean normalised cross-correlation (ZNCC) or the sum of absolute differences (SAD) matching metric. The second term enforces both spatial and temporal regularity, favouring surfaces which deform smoothly over time. The edge function e may be chosen to be discontinuity-preserving, allowing for disparate objects in the reconstructed scene, although this has shown to produce an NP-hard minimisation problem [3].

2.2 Iterated dynamic programming

Iterated dynamic programming has been proposed as a fast technique to minimise first-order energy functions in two-view stereo reconstruction [1]. It proceeds by optimally reassigning disparities in both vertical and horizontal directions of a disparity map until convergence is reached. Here we extend this technique to stereo sequence reconstruction by minimising the energy given in Eq. (1).



(a) (b)

Figure 1. (a) The edge function described by Eq. (3), and (b) its three corresponding min-filters for a given node as described by Eq. (4).

We define a disparity line as the set of points of $d(x, y, t)$ defined by holding two of the variables constant. Consider, for example, optimising the line $d(x_0, y_0, t)$ with fixed x_0, y_0. Eq. (1) becomes

$$E[d(x_0, y_0, \cdot)] = \sum_t (c(x_0, y_0, t, d) + e(|d(x_0, y_0, t) - d(x_0 \pm 1, y_0 \pm 1, t)|)) + \sum_t e(|d(x_0, y_0, t) - d(x_0, y_0, t + 1)|) + \kappa, \quad (2)$$

where κ is a constant absorbing terms unaffected by the optimisation.

Since the optimisation is reduced to a single variable in the t direction, we may now apply dynamic programming to optimally update the disparity values along this line. Each update of the disparity function d for each x, y, t direction in this way monotonically reduces the energy until convergence. As the optimisation procedure is able to replace large sets of pixels simultaneously it is robust to local minima. Due to the first-order energy function, only lines directly neighbouring changed pixels must be re-evaluated at each iteration.

3 Implementation

3.1 Fast dynamic programming by morphological decomposition

The basic update in the proposed optimisation procedure uses dynamic programming to optimally reassign disparity values for a disparity line. Without loss of generality we describe the case $d(x, \cdot, \cdot)$ denoted simply as $d(x)$. Dynamic programming proceeds by computing the minimal subpath to each node (x, d), denoted by $v(x, d)$. This is accomplished by sweeping across the x-d plane recursively evaluating each node

in column $(x, \cdot)$ given the values at $(x-1, \cdot)$. The optimal choice of d for each point x may then be computed by backtracking along the minimal path. For a set of D possible disparities along a line of dimension X, a naive update procedure at each point (x, d) considers all predecessors $(x-1, \cdot)$. The overall computation time is then $O(XD^2)$.

The minimisation proposed in Eq. (2) exhibits additional structure which may be utilised to greatly reduce computation time. The edge costs connecting the nodes at neighbouring columns, $(x-1, \cdot)$ and $(x, \cdot)$, is a function of the disparity difference Δd only. As a result each update becomes a minimum filtering with additive weights. Extensive literature has been published on fast morphological filtering, see [11].

Consider the discontinuity-preserving edge function proposed in [1] (Fig. 1(a)):

$$e(\Delta d) = K_1 \cdot \mathbb{T}(|\Delta d| == 1) + K_2 \cdot \mathbb{T}(|\Delta d| > 1), \quad (3)$$

where K_1 and K_2 are regularisation parameters ($K_1 < K_2$) and $\mathbb{T}(\cdot)$ is the indicator function of its argument. The minimal subpath to each point (x, d) may be decomposed into three minimum filters:

$$v(x,d) = \min \begin{cases} v(x-1,d) \\ K_1 + \min\{v(x-1, d+\Delta d)\} \quad |\Delta d| \leq 1 \\ K_2 + \min\{v(x-1,\cdot)\} \end{cases} \quad (4)$$

Observe that the third min-filter is the minimum of the $(x-1, \cdot)$ column. The second min-filter may be implemented in $O(D)$ time using van Herk's algorithm [11], leading to an overall cost of $O(XD)$ for updating the disparity values of a line via dynamic programming. This is a substantial improvement over the original computation time of $O(XD^2)$. Fig. 1(b) illustrates the min-filters for this edge function.

3.2 Multiscale

A multiscale approach is well known to achieve significant speedups. The estimation of $d(x, y, t)$ from a coarse scale allows the fine scale optimisation to be performed in a narrow band. Sun [10] applied a recursive algorithm for computing the ZNCC and SAD cost functions on rectangular subregions of a $c(x, y, d)$ volume. The application of this method requires a minimal set of boxes spanning the narrow band. In this work we apply the optimal quadtree subregioning algorithm proposed in [1].

4 Experimental Results

In this section, we present our stereo-temporal reconstructions using iterated dynamic programming. We minimise the energy function of Eq. (1) using the edge function described in Eq. (3). We present experimental results computed with and without temporal regularity, demonstrating the benefits of incorporating temporal coherence into an energy minimisation framework. All experiments have been performed on a 1.8GHz Pentium IV laptop under the Windows operating system.

Fig. 2(a) depicts every third frame from a stereo image sequence of a man walking towards the camera. Fig. 2(b) is the result of performing stereo matching on independent frames. This is computed by omitting temporal terms in the minimisation of the energy function of Eq. (1). Fig. 2(c) is the matching result of the same sequence whilst enforcing both spatial and temporal regularity. Notice that the addition of temporal smoothness increases the robustness of matching scene components. This can be observed by comparing the reconstruction of the background between Fig. 2(b) and (c). Greater consistency can be observed in the reconstructions with temporal coherence. While it is acknowledged that background detection methods can be applied with similar results, our energy minimisation approach also improves the reconstruction of moving objects.

A moving object can be assumed to travel smoothly throughout a scene with respect to a sufficiently high frame-rate. Rather than computing the depth of moving objects solely from a single frame, we may integrate the path of an object over several frames to produce significantly more reliable estimates of its depth. In a dynamic scene, regions of matching uncertainties due to moving objects can be more reliably reconstructed given the additional knowledge from its neighbouring frames. This can be observed in Fig. 3 where a second dynamic scene of two people walking is depicted. Fig. 3(a) and (b) are the left and right images of a frame in the image sequence and Fig. 3(c) and (d) are the dense reconstructions without and with temporal regularity considered. It can be seen that the stereo-temporal reconstruction correctly tracks the two people, whereas matching without temporal coherence fails.

The results depicted in Fig. 2 and 3 are computed from 320×225 stereo images with 15 and 31 disparity levels respectively. Using three scales with quadtree subregioning, as well as morphological decomposition for fast dynamic programming, our proposed method takes 0.9 seconds per frame for the first image sequence. The second sequence requires 1.6 seconds per frame.

Figure 2. (a) A set of left input images from a stereo image sequence. (b) The disparity map computed without temporal coherence. (c) The disparity map computed using both spatial and temporal regularity.

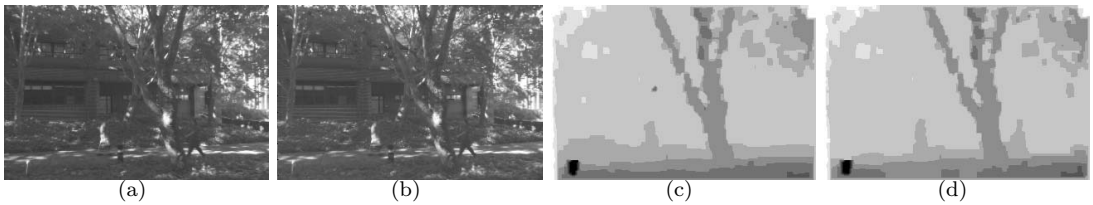


Figure 3. The (a) left and (b) right images of a frame from a stereo image sequence. (c) The disparity map computed without temporal coherence. (d) The disparity map computed using both spatial and temporal regularity.

5 Conclusion

We have presented an energy minimisation framework for the efficient reconstruction of stereo image sequences. An iterated dynamic programming scheme has been proposed to minimise an energy function for matching stereo image sequences incorporating temporal coherence. We also propose a fast solution to the dynamic programming subproblem inspired by efficient morphological filtering algorithms. Timings and results have been presented demonstrating the strength of the proposed method for stereo-temporal reconstruction.

References

[1] Fast stereo matching using quadtree subregioning and energy minimisation. In *British Machine Vision Conference*, 2004. submitted.

[2] J. Barron and R. Eagleson. Binocular estimation of motion and structure from long sequences using optical flow without correspondence. In *Proc. ICIP*, volume 2, pages 193–196, 1995.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE PAMI*, 23(11):1222–1239, November 2001.

[4] M. Brown, D. Burschka, and G. Hager. Advances in computational stereo. *IEEE PAMI*, 25(8):993–1008, January 2003.

[5] J. Davis, R. Ramamoorthi, and S. Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. In *Proc. CVPR*, volume 2, pages 359–366, June 2003.

[6] L. Falkenhagen. 3D object-based depth estimation from stereoscopic image sequences. In *International Workshop on Stereoscopic and Three Dimensional Imaging*, Greece, September 1995.

[7] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dence two-frame stereo correspondence algorithms. *IJCV*, 47(1/2/3):7–42, April–June 2002.

[8] J. Shao. Generation of temporally consistent multiple virtual camera views from stereoscopic image sequences. *IJCV*, 47(1/2/3):171–180, April–June 2002.

[9] C. Strecha and L. J. van Gool. Motion-stereo integration for depth estimation. In *Proc. ECCV*, volume 2, pages 170–185, 2002.

[10] C. Sun. Fast stereo matching using rectangular subregioning and 3D maximum-surface techniques. *IJCV*, 47(1/2/3):99–117, April–June 2002.

[11] M. van Herk. A fast algorithm for local minimum and maximum filters on rectangular and octagonal kernels. *Pattern Recognition Letters*, 13(7):517–521, July 1992.

[12] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. In *Proc. ICCV*, volume 2, pages 722–729, 1999.

[13] L. Zhang, B. Curless, and S. Seitz. Spacetime stereo: Shape recovery for dynamic scenes. In *Proc. CVPR*, volume 2, pages 367–374, June 2003.