Review: Markerless Tracking using Planar Structures in the Scene

Wei Zhang
Department of Computer Science
University of California, Santa Barbara
wei@cs.ucsb.edu
http://www.cs.ucsb.edu/~wei

Abstract

In this paper, the author proposed a taxonomy for dense two-frame stereo correspondence algorithms. This taxonomy can be used to highlight the most important features of existing stereo algorithms, and to study important algorithmic components in isolation. A software framework of stereo matching algorithm components are implemented to test the performance of these algorithms. And they also performed an exhaustive experimental investigation in order to assess the impact of the different algorithmic components.

Since this paper is proposing a survey rather than a new algorithm, and because I'm still a beginner in stereo vision, in this review I will mainly focused on summarizing the taxonomy proposed by the author. By doing this I can describe an overview of the essential idea and structure of most dense stereo algorithms that being covered in this paper.

1. Introduction

In stereo correspondence, it is difficult to measure the progress of this field because a lack of a common quantitative method. There are two previous comparative papers that focused on the performance of sparse feature matchers (Hsieh et al. 1992, Bolles et al. 1993), and two recent papers (Szeliski 1999, Mulligan et al. 2001) that compared the performance of several popular algorithms, but did not provide a detailed taxonomy or a complete coverage of algorithms. This paper continued the investigations begun by Szeliski and Zabih (1999).

However, this paper is not an simple attempt to provide a taxonomy of stereo algorithms (otherwise it might not get published on IJCV). The taxonomy they proposed is designed to identify the individual components and design decisions within published algorithms. The goal of such effort is to finally create a standard test-bed for the quantitative evaluation of dense stereo algorithms and thus help researchers to develop new and better algorithms.

2. The Taxonomy

Since the goal of this work is to compare a large number of methods within one common framework, several general assumptions about the physical world and the image formation process are needed. The first common assumption is that algorithms will mostly deal with Lambertian surfaces whose appearance does not vary with viewpoint. Another important assumption is physical world consists of piecewise-smooth surfaces. Furthermore, algorithms in this framework are assumed to be given a pair of rectified images as input, and the expected output is the disparity space image, which is an univalued function in disparity space $d(x, y)$ that best describes the shape of the surfaces in the scene.

Then the taxonomy can be proposed based on the observation that many stereo algorithms generally perform some or all of the following four steps (actual sequence of steps taken depends on the specific algorithm).

2.1. Matching cost computation

In this step, the problem of matching cost computation is standardized as: Calculate the matching cost values over all pixels and all disparities form the initial disparity space image $C_0(x, y, d)$.

In its implementation, this paper evaluated many algorithms and additional techniques, such as *squared intensity differences* (most commonly used), *cross-correlation*, *robust matching score*, *fractional disparity evaluation*, *sampling insensitive interval-based matching criterion*.

2.2. Aggregation of cost

There are many local and window-based methods aggregate the matching cost by summing or averaging over a support region in the DSI $C(x, y, d)$. A support region can be either two-dimensional at a fixed disparity (favoring fronto-parallel surfaces), or three-dimensional in $x - y - d$

space (supporting slanted surfaces). There's also different method of aggregation, such as iterative diffusion, in which an aggregation (or averaging) operation is implemented by repeatedly adding to each pixel's cost the weighted values of its neighboring pixels' costs.

This paper implemented two commonly used aggregation methods, *box filter* and *binomial filter*. *Separable square min-filter*, *cascaded effect of a box-filter* and an *equal-sized min-filter* can be added afterwards if the algorithm need the effect of shiftable windows.

2.3. Disparity computation and optimization

For this component, the optimization methods are summarized as 5 algorithms, all of which minimize the same objective function, enabling a more meaningful comparison of their performance.

- *Winner-take-all* (WTA): Simply picks the lowest (aggregated) matching cost as the selected disparity at each pixel.

- *Dynamic programming* (DP): A global optimization technique, works by computing the minimum-cost path through each $x - d$ slice in the DSI.

- *Scanline optimization* (SO): A simple approach designed to assess different smoothness terms. It operates on individual $x - d$ DSI slices and optimizes one scanline at a time, but it is asymmetric and does not utilize visibility or ordering constraints.

- *Simulated annealing* (SA): A classic optimization method, in this implementation it supports both the Metropolis variant (where downhill steps are always taken, and uphill steps are sometimes taken), and the Gibbs Sampler, which chooses among several possible states according to the full marginal distribution).

- *Graph cut* (GC): Implements the a-b swap move algorithm described in (Boykov et al. 1999, Veksler 1999).

2.4. Refinement of disparities

They only discussed the sub-pixel refinement of disparities.

3. Experiment

In the experiment section they made comprehensive evaluation for individual building blocks of stereo algorithms, and there are much more detailed result available on their website. The experiments discussed in this paper demonstrated the limitations of local methods, and have assessed the value of different global techniques and their sensitivity to key parameters.

3.1. Test data

The evaluation requires data sets that either have a ground truth disparity map, or a set of additional views that can be used for prediction error test (or preferably both. Each image sequence consists of 9 images, taken at regular intervals with a camera mounted on a horizontal translation stage, with the camera pointing perpendicularly to the direction of motion. First all of the sequences are made up of piecewise planar objects, then use a direct alignment technique on each planar region (Baker et al. 1998) to estimate the affine motion of each patch. The horizontal component of these motions is then used to compute the ground truth disparity.

3.2. Quality metrics

They actually measured three quality measures based on known ground truth data:

- RMS (root-mean-squared) error (measured in disparity units) between the computed depth map $d_C(x, y)$ and the ground truth map $d_T(x, y)$.

- Percentage of bad matching pixels.

- Use the color images and disparity maps to predict the appearance of other views (Szeliski 1999), then compare with the real image.

4. Contribution and limitations

I think the major contribution of this paper are:

- It has an comprehensive overview on the state of art in dense stereo algorithms, this could be very helpful for people who wants to enter this area.

- It provides a taxonomy of existing stereo algorithms, with standard data sets for testing. This is valuable since the ground truth of image is not easy to get.

- Under the taxonomy and dataset, the author implemented a lot of dense stereo vision algorithms and evaulated them.

- Finally it proposes a framework for researchers to modularize their algorithms, thus makes the results to be more fairly comparable, and future algorithm evaulation becomes easier.

However, from my point of view, the author's methodology also has some limitations (but since I'm not an expert in this area, these might be wrong):

- First, many published methods include special features and post-processing steps to improve the results, is it fair to ignore these features and only compare the basic version? Maybe these features can affect the results of some algorithms seriously, especially when I found that the results of some algorithms on the website are too bad to get published.

- Second, only algorithms in certain kind can be evaluated in this framework. So I have question in how well can later algorithms be fitted into it? But since there are many algorithms implemented, this is already a great work.

5. Summary

All in all, this is a very good paper. It is very well-written, covers major dense two-frame stereo correspondence algorithms at the state of art, and still keep updating on the website. The analysis, comparison, and evaluation of these algorithms are comprehensive and impressive. This paper might be a little difficult for beginners since it requires some prior experiences , but it is an important tool and perfect reference for researchers who are working in this area.