Disparity Estimation in Stereo Sequences using Scene Flow

Fang Liu
Fang.F.Liu@Philips.com
Philips Germany

Vasanth Philomin
Vasanth.Philomin@Philips.com
Philips Germany

Estimating the disparity images for a stereo image sequence has been a challenging problem for many years. Some existing stereo match algorithms can achieve good results on a single pair of stereo frames [2, 3, 4]. However, they are not sufficient for depth estimation on a stereo image sequence, without considering the temporal consistency between adjacent frames. In this paper, we propose a method to estimate the disparity maps for a stereo image sequence. We model the temporal consistency of disparity maps using scene flow [1]. The scene flow is built into the stereo model for the next frame as a soft constraint, which helps to resolve stereo ambiguity from the temporal domain while at the same time reducing the error propagation in the disparity maps.

Algorithm Overview In the workflow for a stereo sequence, assuming that the disparity maps (D_t^L, D_t^R) (in left and right views) at time t are estimated, we compute the scene flow between time t and $t+1$, and use them to predict the disparity maps at time $t+1$ (P_{t+1}^L, P_{t+1}^R). The scene flow is also used to compute the confidence maps that model the strength of *temporal links*. These are further employed in the iterative stereo estimation to generate the disparity maps at time $t+1$.

We adopt a hybrid approach to generate the initial disparity maps for the first frame: we first use the over-segmentation based stereo algorithm similar to [5] to estimate the preliminary disparity maps. Then we incorporate the estimated disparity of each segment into the pixelwise stereo model as a soft constraint. The reasons are twofold: 1. The over-segmentation based approaches prove to be very robust and insensitive to image noise and color bias between the left and right views. 2. Over-segmentation may not be correct in cluttered scenes since it is only based on color and segmentation artifacts will directly affect the accuracy of the disparity maps. At the pixel level, the disparity estimation can maintain the robustness of segment-based approach and at the same time correct the segmentation errors using stereo match measurements.

Iterative Stereo Estimation The iterative stereo algorithm comprises of two steps: the first step is the over-segmentation based stereo algorithm similar to [5], and then in the second step, we apply the pixelwise iterative algorithm similar to [3], where the occlusions are explicitly handled in a symmetric stereo match model with the segment disparity maps encoded as a soft constraint.

The correlation volume is computed using the locally adaptive support weight method. Then we merge the correlation volume and segment cost into the final data term, using the adaptive weights based on stable pixel and occluded pixel mask images. The additional temporal cost, when available, is also coalesced into the data term with the adaptive weights. The smoothness cost is defined as a truncated linear model and is tuned based on color edges. Hierarchical loopy belief propagation is employed to realize the optimization of total energy. Once the optimization is done, the current solutions are used to update the occlusion masks and the data cost. We then optimize again with the new data cost and this process is stopped once the solutions converge.

Scene Flow Estimation The optic flow and disparity maps are jointly estimated from the rectified image sequences. Let $I_l(x,y,t)$ and $I_r(x,y,t)$ be the two stereo images at time t. Assuming that the disparity maps at time t have been estimated, to predict the disparity maps at $t+1$, the scene flow field is defined as (u,v,w), where (u,v) is the optic flow in left view, and w is the change in disparity. Each element of the scene flow is the scalar function defined in the reference view. The global energy functional is defined as follows:

$$E(u,v,w) = E_{Data} + \alpha E_{Smooth} = E_{fl} + E_{fr} + E_{st} + \alpha E_{Smooth} \quad (1)$$

where α is regularization parameter, and the data cost E_{Data} consists of three data terms: E_{fl} and E_{fr} for left and right optic flows, and E_{st} for stereo at time $t+1$. The smoothness term is defined to penalize the local variations in the flow fields. Their definitions are described in the paper. According to calculus of variations, an extreme of the energy functional E can be achieved by solving its Euler-Lagrange equations, which are nonlinear Partial Differential Equations (PDEs). To solve these nonlinear PDEs, we adopt the strategy of two nested fixed point iteration loops as in [1]. These fixed point iterations are combined with a coarse-to-fine strategy to better approximate the global optimum of the energy.

Disparity maps prediction From the scene flow definitions, the predicted disparity maps in the left view can be computed as:

$$P_{t+1}^L(x+u, y+v) = D_t^L(x,y) + w(x,y) \quad (2)$$

P_{t+1}^R is computed similarly. To prevent errors from propagating into the predicted disparity maps, we compute the confidence image together with the disparity predictions. The confidence of a pixel is measured by the color similarity between that pixel and its correspondences according to the scene flow vector:

$$\omega = m_{fl} m_{fr} m_{st} exp\left(-\frac{\|I_{lz}\|^2 + \|I_{rz}\|^2 + \|I_{dz}\|^2}{\sigma_c^2}\right) \quad (3)$$

where ω denotes the confidence of the correspondences specified by the scene flow field, and σ_c is the normalization parameter for the color difference. The confidence image is then warped to the left and right frames at time $t+1$ with the estimated scene flow vectors. The temporal cost is defined to penalize the differences between the predicted disparity and the disparity to be estimated based on the confidence:

$$E_{temporal}(D) = \sum_i \omega_i min(|d_i - P_i|, T_{temporal}) \quad (4)$$

where for each pixel i, d_i is the disparity to be estimated. P_i and ω_i are the predicted disparity and confidence respectively. $T_{temporal}$ is a truncation threshold for temporal cost. Basically the confidence controls the contribution of the temporal consistency term into the overall energy for stereo estimation. The temporal cost is fed into the iterative stereo estimation algorithm as an additional cost and merged into the final data term as explained before.

Experiments We first test the hybrid stereo algorithm on the Middlebury stereo datasets [2] and the results rank 8th overall in the evaluation as of April 2009. To evaluate the scene flow algorithm, we use the same datasets with the initial disparity map estimated from our single-frame hybrid stereo method. We achieve much lower error rates for the initial disparity map compared with Huguet's method [1], and this helps to improve the accuracy of the scene flow fields. The algorithm is also tested on a real world stereo image sequence. We compare the disparity maps obtained by our stereo approach with and without the temporal constraint.

Conclusion Our temporal modeling using scene flow can indeed improve the temporal consistency in the disparity maps. There could be some error propagation in the sequential estimation workflow, which requires a more extensive MRF framework that includes temporal and stereo constraints to approximate the optimal solution better. This will be investigated in future work.

[1] F. Huguet and F. Devernay. A variational method for scene flow estimation from stereo sequences. In *Proc. Intl. Conf. on Computer Vision*, Rio de Janeiro, Brasil, October 2007. IEEE.

[2] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, 47(1-3):7–42, 2002. ISSN 0920-5691.

[3] J. Sun, Y. Li, S. B. Kang, and H.-Y. Shum. Symmetric stereo matching for occlusion handling. In *CVPR '05*, pages 399–406, 2005.

[4] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister. Stereo matching with color-weighted correlation, hierachical belief propagation and occlusion handling. In *CVPR '06*, pages 2347–2354, 2006.

[5] C. L. Zitnick and S. B. Kang. Stereo for image-based rendering using image over-segmentation. *Int. J. Comput. Vision*, 75(1):49–65, 2007.