JOINT DISPARITY AND OPTICAL FLOW BY CORRESPONDENCE GROWING

Jan Čech, Radu Horaud

INRIA Rhône-Alpes, 38330 Montbonnot, France

ABSTRACT

The scene flow in binocular stereo setup is estimated using a seed growing algorithm. A pair of calibrated and synchronized cameras observe a scene and output a sequence of images. The algorithm jointly computes a disparity map between the stereo images and optical flow maps between consecutive frames. Having the calibration, this is a representation of the scene flow, i.e. a 3D velocity vector is associated with each reconstructed 3D point.

The proposed algorithm starts from correspondence seeds and propagates the correspondences to the neighborhood. It is accurate for complex scenes with large motion and produces temporally coherent stereo disparity and optical flow results. The algorithm is fast due to inherent search space reduction.

Index Terms— image, stereo, disparity, optical flow.

1. INTRODUCTION

A sequence of images from calibrated and synchronized cameras contains more information to estimate depth than a single set of still images. There are approaches [1, 2, 3] which exploit the extra information from the sequence to estimate disparity maps, but do not estimate the motion explicitly, we call them a spatiotemporal stereo.

Other methods estimate a complete scene flow benefiting from a coupled stereo and optical flow correspondence problem. The notion of scene flow was introduced in [4] as a dense 3D motion field. In the literature, it is estimated by: (1) variational methods [5, 6], which are usually well suitable for simple scenes with a dominant surface; (2) discrete MRF formulation [7, 8], which involves an expensive discrete optimization; (3) local methods finding the correspondences greedily, which can be very fast [9], but not so accurate.

We propose a seed growing algorithm to estimate the scene flow in binocular stereo video setup. A basic principle of the seed growing methods is that correspondences are found in a small neighborhood around given initial seed correspondences. This idea has been adopted in stereo [10, 11, 12, 13]. The advantage of such approaches is a fast performance compared to global MRF methods, and a good accuracy compared to purely local method, since neighboring pixel relations are not ignored completely. Our proposed algorithm can simultaneously estimate accurate temporally coherent disparity and optical flow maps of a scene with a rich 3D structure and large motion between time instances. Large displacements are found due to the seeds, while small local variations of disparity and flows are captured by the growing process. Boundaries between objects and different motions are naturally well preserved without smoothing artifacts. A drawback is that the algorithm produces semi-dense (unambiguous) results only.

The research was supported by EC project FP7-ICT-247525-HUMAVIPS.

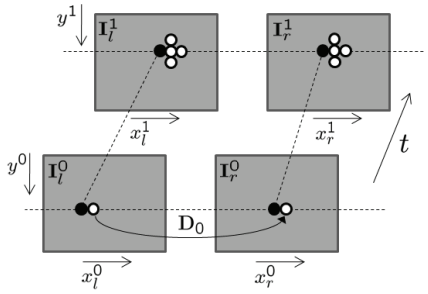

Fig. 1. A sequence of consecutive epipolarly rectified stereo images. A seed correspondence $\mathbf{s}$ sketched by filled circles, its right neighborhood $\mathcal{N}_1$ by empty circles.

2. ALGORITHM DESCRIPTION

The algorithm is presented in pseudocode as Alg. 1. It takes four rectified images, a stereopair $\mathbf{I}_l^0, \mathbf{I}_r^0$ for time $t-1$, and the consecutive stereopair $\mathbf{I}_l^1, \mathbf{I}_r^1$ for time t, a set of initial correspondence seeds $\mathcal{S}$, disparity map $\mathbf{D}^0$ holding the stereo correspondences from a previous frame $t-1$, and parameters α (temporal consistency enforcement), β (optical flow regularization), τ (growing threshold).

The output are disparity map $\mathbf{D}^1$, holding correspondences found between $\mathbf{I}_l^1$ and $\mathbf{I}_r^1$, and horizontal and vertical optical flow maps $\mathbf{F}_h$ and $\mathbf{F}_v$ respectively, encoding the correspondences between $\mathbf{I}_l^0$ and $\mathbf{I}_l^1$. Notice that having full camera calibrations, this representation fully determines the scene flow, since $\mathbf{D}^0$ gives a reconstruction of 3D points $\mathcal{X}^0$, $\mathbf{D}^1$ a reconstruction of 3D points $\mathcal{X}^1$ (after the motion), and $\mathbf{F}_h, \mathbf{F}_v$ gives the mapping between these two sets.

Each seed $\mathbf{s} = (x_l^0, x_r^0, y^0, x_l^1, x_r^1, y^1) \in \mathcal{S}$ represents a correspondence of 4 pixels, i.e. projections of a 3D point

Algorithm 1 Growing the scene flow

Require: rectified images $\mathbf{I}_l^0$, $\mathbf{I}_r^0$, $\mathbf{I}_l^1$, $\mathbf{I}_r^1$,
initial correspondence seeds $\mathcal{S}$,
disparity map $\mathbf{D}^0$,
parameters α, β, τ.

1: Compute similarity $\mathbf{s}.c := \mathrm{corr}(\mathbf{s}) + \alpha$ for all seeds $\mathbf{s} \in \mathcal{S}$.
2: **repeat**
3: Draw the seed $\mathbf{s} \in \mathcal{S}$ of the best similarity $\mathbf{s}.c$.
4: **if** $\mathbf{s}.c \geq \tau$ **then** Update output maps. **endif**
5: **for** each of the four best neighbors $i \in \{1,2,3,4\}$
$\mathbf{t}_i^* = (x_l^0, x_r^0, y^0, x_l^1, x_r^1, y^1) = \underset{\mathbf{t} \in \mathcal{N}_i(\mathbf{s}|\mathbf{D}^0)}{\mathrm{argmax}} \mathrm{corr}_{\mathbf{s}}^{\beta}(\mathbf{t})$,
do
6: $\mathbf{t}_i.c := \mathrm{corr}_{\mathbf{s}}^{\beta}(\mathbf{t}_i^*)$,
7: **if** $\mathbf{t}_i.c \geq \tau$ **and** all pixels in $\mathbf{t}$ not matched yet **then**
8: Update output maps.
9: Update the seed queue $\mathcal{S} := \mathcal{S} \cup \{\mathbf{t}_i^*\}$.
10: **end if**
11: **end for**
12: **until** $\mathcal{S}$ is empty.
13: **return** disparity map $\mathbf{D}^1$, flow maps $\mathbf{F}_h$, $\mathbf{F}_v$.

$X^0 \in \mathcal{X}^0$ into $\mathbf{I}_l^0, \mathbf{I}_r^0$ and the same 3D point after the motion $X^1 \in \mathcal{X}^1$ into $\mathbf{I}_l^1, \mathbf{I}_r^1$. The seed encapsulates both stereo and optical flow correspondences, see Fig. 1. First, the algorithm computes a photometric consistency statistic of the 4-correspondence by average correlation

$$\mathrm{corr}(\mathbf{s}) = \frac{\mathrm{c}_{lr}^{11}(x_l^1,y_l^1;x_r^1,y_l^1) + \mathrm{c}_{ll}^{01}(x_l^0,y_l^0;x_l^1,y_l^1) + \mathrm{c}_{rr}^{01}(x_r^0,y_r^0;x_r^1,y_r^1)}{3}. \quad (1)$$

Left-right correlation c_{lr}^{11} is between small windows centered at pixels $\mathbf{I}_l^1(x_l^1, y_l^1)$ and $\mathbf{I}_r^1(x_l^1, y_l^1)$. Similarly the correlations c_{ll}^{01} and c_{rr}^{01} are between consecutive frames in the left and right streams. All the correlations are MNCC statistics [14] on 5×5 px widows. Seed correlation $\mathbf{s}.c$ is enhanced by a small positive α to enforce temporal consistency, Step 1.

The set $\mathcal{S}$ is organized as a correlation priority queue. The seed $\mathbf{s} \in \mathcal{S}$ is removed from the top of the queue, Step 3. If its consistency exceeds threshold τ in Step 4, output maps are updated by

$$\mathbf{D}^1(x_l^1, y^1) := x_l^1 - x_r^1, \quad (2)$$
$$\mathbf{F}_h(x_l^0, y^0) := x_l^0 - x_l^1, \mathbf{F}_v(x_l^0, y^0) := y^0 - y^1.$$

For all 4-neighbors (right, left, up, down) of seed $\mathbf{s}$, the best correlating candidate in $\mathcal{N}_i(\mathbf{s}|\mathbf{D}^0)$ is found, Step 5. For instance

$$\mathcal{N}_1(\mathbf{s}|\mathbf{D}^0) = \Big\{ \bigcup_{\mathbf{k} \in \mathcal{L}} (x_l^0 + 1, x_l^0 + 1 - \mathbf{D}^0(x_l^0 + 1, y^0), y^0, x_l^1 + 1, x_r^1 + 1, y^1) + (0,0,0,\mathbf{k}) \Big\}, \quad (3)$$

where $\mathcal{L} = \{(0,0,0), (\pm 1,0,0), (0,\pm 1,0), (0,0,\pm 1)\}$ is a set of seven local search vectors having the stereo or temporal disparity less or equal to one, see Fig. 1. Notice the candidates depend on the previous disparity $\mathbf{D}^0$. The other neighborhoods $\mathcal{N}_2, \mathcal{N}_3, \mathcal{N}_4$ are analogical.

The optical flow generally suffers from a well known aperture problem. This is not completely avoided in a joint stereo setup. Therefore we regularize assuming the seed has a correct flow, new candidates having a different flow are penalized by lower correlation

$$\mathrm{corr}(\mathbf{t})_{\mathbf{s}}^{\beta} = \mathrm{corr}(\mathbf{t}) - \beta ||\mathbf{s}.f - \mathbf{t}.f||_1, \quad (4)$$

where notation $.f = (x_l^0 - x_l^1, x_r^0 - x_r^1, y^0 - y^1)$ means a vector of optical flows of respective seeds $\mathbf{s}$ and $\mathbf{t}$, β is a small positive constant.

If the highest correlation exceeds a threshold τ and any of the pixels in $\mathbf{t}$ is unmatched so far, then a new match is found, Step 7. Output maps are updated by (2) in Step 8, and the found match becomes a new seed, Step 9. Up to four seeds are created in each growing step. The process continues until there are no seeds in the queue, Step 12.

For processing a long sequence of stereo images, Alg. 1 is used repeatedly, such that when a new stereo frame $\mathbf{I}_l, \mathbf{I}_r$ comes, we assign $\mathbf{I}_l^0 := \mathbf{I}_l^1$, $\mathbf{I}_r^0 := \mathbf{I}_r^1$, $\mathbf{D}^0 := \mathbf{D}^1$, and the new frame to $\mathbf{I}_l^1 := \mathbf{I}_l$, $\mathbf{I}_r^1 := \mathbf{I}_r$.

For the first two frame of the stereo image sequence, the seeds are obtained by matching Harris points and tracking them by LK-tracker [15]. It is important to note that for next frames, the seeds are additionally computed from previous disparity and optical flow results such that we assume the optical flows remains the same in the next time instance. We observed this assumption usually holds and leads to better accuracy and also a speed up of Alg. 1. If the constant motion[1] assumption is violated, the affected seeds become wrong with low correlation and are placed in the unfavorable position in the queue. Such regions are grown from other correct seeds (sparse Harris seeds, or other seeds where the assumption holds).

First disparity map $\mathbf{D}^0$ is obtained by (also seed growing) stereo algorithm [11]. Notice that number of matched pixels in $\mathbf{D}^1$ is always smaller or equal than in $\mathbf{D}^0$, since the candidates in the neighborhood (3) do not exist in locations where $\mathbf{D}^0$ is undefined. Therefore, when next frame comes, disparity map $\mathbf{D}^0 := \mathbf{D}^1$ and this disparity map is further grown again by the left-right stereo algorithm [11], to recover motion occluded pixels. Without this recovery, the algorithm would in fact track the initial points in the first disparity map over time, but it will sooner or later end up with an empty result.

Algorithm parameters were set empirically to $\alpha = \beta = 0.1$, $\tau = 0.6$ and they were fixed in all our experiments.

3. EXPERIMENTS

The experiments show that the proposed algorithm produces accurate semi-dense results and that it benefits from a joint

[1] More sophisticated motion model and Kalman filtering could be used.

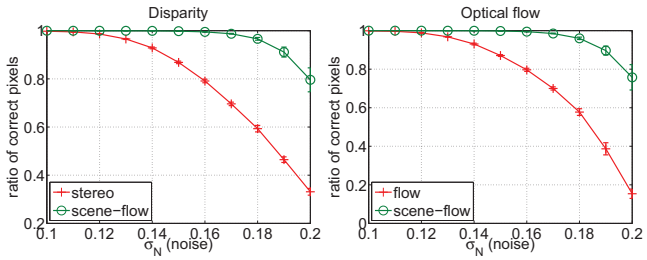

Fig. 2. Algorithm accuracy under contamination with a Gaussian noise. Signal has a range $[0, 1]$.

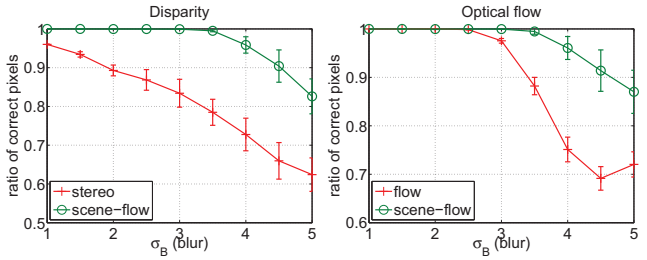

Fig. 3. Algorithm accuracy under a lack of texture simulated by a Gaussian blur.

disparity – optical flow formulation in a sequence of stereo images.

Synthetic Data. To quantitatively evaluate these claims, we preformed a simulated experiment. A moving planar scene textured randomly with a white noise was synthesized. It was projected into a sequence of 20 frames of stereopair images. Each frame has associated ground-truth disparity and optical flow maps. For simplicity, the disparity and optical flows were constant throughout the sequence.

For all the experiments, we measured an average ratio of correctly matched pixels, i.e. number of all pixels without mismatches (error $\geq$ 1 px) and non-matches divided by number of all pixels, over all frames in the sequence. We measured it for both disparity and optical flow errors. The simulation was carried out for 10 random trials over texture (and noise) generation and the plots have error bars of standard deviation.

First, we tested the algorithm under noise contamination, Fig 2. An independent Gaussian noise was added into each image of the stereo sequence. We compared the proposed algorithm which jointly estimates disparity and optical flow (scene-flow, green circles) with other seed growing algorithm which separately computes the disparity and optical flow frame-by-frame independently [11, 10] (stereo resp. flow, red crosses). We can see the scene-flow algorithm is more robust to noise than independent estimates. For instance, in case of disparity, having 20% of noise in the signal of the image texture, the proposed algorithm is still 80% accurate, while the independent algorithm about 30% only.

Second, we tested influence of a lack of texture which was simulated by a convolution of the image with a 2D Gaussian filter, Fig. 3. The lack of texture is a challenging phenomenon for all stereo algorithms. Again, the behavior of the proposed algorithm is superior to the independent estimates.

The favorable results of the proposed algorithm are a consequence of: (1) joint disparity and optical flow estimates which constraint each other, and (2) good temporal consistency and coherence. When data is weak due to noise or insufficient texture, there is a lack of correctly matched seeds and also the growing process is stopped early (by the condition in Step 7 of Alg. 1) in a small distance from the seeds. However, if we feed partially grown disparity and optical flow maps as the seeds to the scene flow algorithm, it grows them further if they were correct. This effect is repeated, and after certain number of frames, high quality seeds are accumulated.

Real data. The proposed algorithm was tested on real data as well. We show results on CAVA dataset of INRIA[2], where the stereo camera is static, and on the dataset of ETH Zürich[3] acquired by a mobile stereo platform.

The results of the proposed algorithm are shown in Fig. 4 as disparity $\mathbf{D}^1$ and optical flow $\mathbf{F}_h, \mathbf{F}_v$ maps. For CAVA, the results are sufficiently dense even for weakly textured office environment. Important scene structures are matched. Notice sharply preserved boundaries between objects in both disparity and optical flow. We can see a left-down motion of the man coming through the door, which are closing afterward performing a slower left motion. One of the ladies is walking to the right to reach the chair, while moving her arm down. We can also recognize a hand gesticulation of the sitting man. ETH dataset represents a complex scene with both camera forward motion and motion of pedestrians. There are up to 30 px displacements between consecutive frames. We can see a motion of the planar sidewalk close to cameras and well captured depth and motion boundaries of the people walking. There are only few small mismatches which are visible in both depth and flow maps. This is in the region of the leftmost building which effects complicated non-Lambertian mirror like reflections. The results of the sequence are temporally coherent without flickering.

4. CONCLUSIONS

We presented an algorithm which simultaneously estimates semi-dense disparity and optical flows of a stereo sequence by growing correspondence seeds. We experimentally proved that results are more accurate and temporally coherent than frame-by-frame independent algorithms.

The algorithm has low complexity. Assuming $n \times n$ images, any algorithm searching the correspondences exhaustively has the complexity at least $\mathcal{O}(n^5)$ per frame [9], which

[2] http://perception.inrialpes.fr/CAVA_Dataset/
[3] http://www.vision.ee.ethz.ch/~aess/dataset/

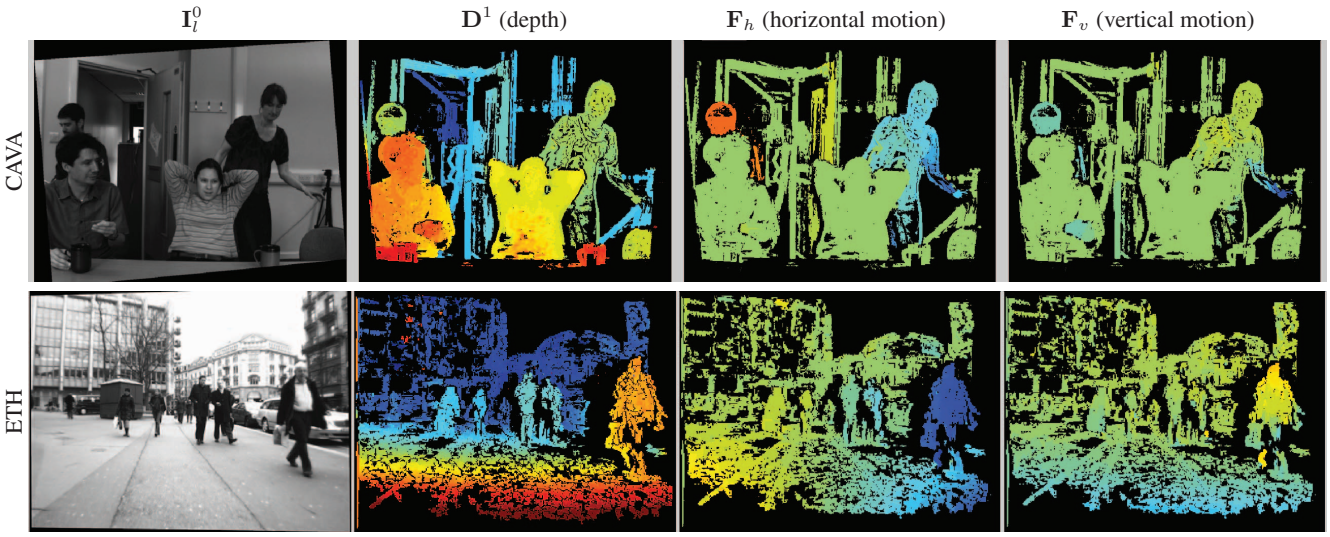

Fig. 4. Results on real data. Color-coded disparity and optical flow maps. For disparity $\mathbf{D}^1$, warmer colors are closer to the camera. For horizontal $\mathbf{F}_h$ and vertical $\mathbf{F}_v$ flows, green color is zero motion, warmer colors is left and up motion, colder colors is right and down motion respectively. Black color denotes unmatched pixel. The figure is best seen in the electronic version.

is the size of the search space without limiting the ranges for disparity and horizontal and vertical flow. However, the proposed algorithm has the complexity $\mathcal{O}(n^2)$ per frame, since it searches the correspondences in a neighborhood of the seeds tracing discrete manifolds of a high correlation defined above the pixels of the reference image. With our non-optimized partially Matlab implementation, it runs about 1.5 seconds per frame for VGA images on a common PC.

5. REFERENCES

[1] L. Zhang, B. Curless, and S. M. Seitz, "Spacetime stereo: Shape recovery for dynamic scenes," in *CVPR*, 2003.

[2] M. Sizintsev and R. P. Wildes, "Spatiotemporal stereo via spatiotemporal quadratic element (stequel) matching," in *CVPR*, 2009.

[3] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N. A. Dodgson, "Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid," in *ECCV*, 2010.

[4] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade, "Three-dimensional scene flow," in *ICCV*, 1999.

[5] T. Basha, Y. Moses, and N. Kiryati, "Multi-view scene flow estimation: A view centered variational approach," in *CVPR*, 2010.

[6] F. Huguet and F. Devernay, "A variational method for scene flow estimation from stereo sequences," in *ICCV*, 2007.

[7] F. Liu and V. Philomin, "Disparity estimation in stereo sequences using scene flow," in *BMVC*, 2009.

[8] M. Isard and J. MacCormick, "Dense motion and disparity estimation via loopy belief propagation," in *ACCV*, 2006.

[9] M. Gong, "Real-time joint disparity and disparity flow estimation on programmable graphics hardware," *CVIU*, vol. 113, no. 1, 2009.

[10] J. Čech, J. Matas, and M. Perďoch, "Efficient sequential correspondence selection by cosegmentation," *IEEE Trans. on PAMI*, vol. 32, no. 9, 2010.

[11] J. Čech and R. Šára, "Efficient sampling of disparity space for fast and accurate matching," in *BenCOS Workshop, CVPR*, 2007, `http://cmp.felk.cvut.cz/~cechj/gcs`.

[12] M. Lhuillier and L. Quan, "Match propagation for image-based modeling and rendering," *IEEE Trans. on PAMI*, vol. 24, no. 8, 2002.

[13] J. Kannala and S. S. Brandt, "Quasi-dense wide baseline matching using match propagation," in *CVPR*, 2007.

[14] H. P. Moravec, "Towards automatic visual obstacle avoidance," in *IJCAI*, 1977, p. 584.

[15] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI*, 1981.