HAL
archives-ouvertes.fr

Scene Flow Estimation by Growing Correspondence Seeds

Jan Cech, Jordi Sanchez-Riera, Radu Horaud

▶ To cite this version:

Jan Cech, Jordi Sanchez-Riera, Radu Horaud. Scene Flow Estimation by Growing Correspondence Seeds. CVPR 2011 - IEEE Conference on Computer Vision and Pattern Recognition, Jun 2011, Colorado Springs, United States. pp.3129-3136, 10.1109/CVPR.2011.5995442 . inria-00590274

HAL Id: inria-00590274
https://hal.inria.fr/inria-00590274

Submitted on 14 Jul 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CVPR 2011, (c) IEEE 2011

Scene Flow Estimation by Growing Correspondence Seeds

Jan Čech, Jordi Sanchez-Riera, Radu Horaud
INRIA Rhône-Alpes, 38330 Montbonnot, France
{jan.cech,jordi.sanchez-riera,radu.horaud}@inrialpes.fr

Abstract

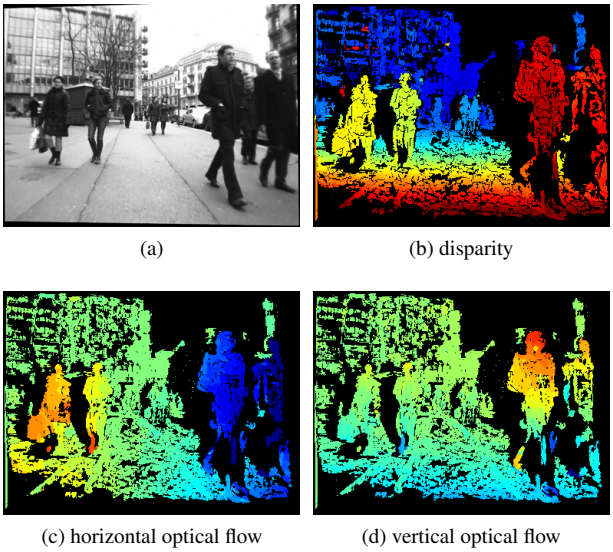
A simple seed growing algorithm for estimating scene flow in a stereo setup is presented. Two calibrated and synchronized cameras observe a scene and output a sequence of image pairs. The algorithm simultaneously computes a disparity map between the image pairs and optical flow maps between consecutive images. This, together with calibration data, is an equivalent representation of the 3D scene flow, i.e. a 3D velocity vector is associated with each reconstructed point. The proposed method starts from correspondence seeds and propagates these correspondences to their neighborhood. It is accurate for complex scenes with large motions and produces temporally-coherent stereo disparity and optical flow results. The algorithm is fast due to inherent search space reduction. An explicit comparison with recent methods of spatiotemporal stereo and variational optical and scene flow is provided.

1. Introduction

A sequence of image pairs gathered with calibrated and synchronized cameras contains more information to estimate depth and 3D motion than a single stereopair or a single image sequence. There are approaches [17, 15, 14] which exploit the extra temporal information 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. *Scene flow* was introduced in [16] as a dense 3D motion field. It can be estimated with: (1) variational methods [1, 6, 13], which are usually well suited for simple scenes with a dominant surface; (2) discrete MRF formulations [10, 7], which involve expensive discrete optimization, and (3) local methods finding the correspondences greedily, which are efficient [5] but not so accurate.

We propose a seed growing algorithm to estimate the scene flow in a binocular-video setup. A basic principle of the seed growing methods is that correspondences are found in a small neighborhood around an initial set of seed correspondences. This idea has been adopted in stereo [3, 4, 9, 8], but to the best of our knowledge, it has not been used for scene flow. The advantage of such approaches is a fast performance compared to global variational and MRF methods, and a good accuracy compared to purely local methods, since neighboring pixel relations are not ignored completely.



(a) (b) disparity (c) horizontal optical flow (d) vertical optical flow

Figure 1: Output of the proposed algorithm on ETH dataset as color coded maps. For disparity, warmer colors are closer to the camera. In optical flow, green color is zero motion, warmer colors is left and up motion, colder colors is right and down motion respectively. Black color denotes unmatched pixels.

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. Small local variations of disparity and flows are captured by the growing process while large displacement are found due to the seeds. Boundaries between objects and different motions are naturally well preserved without smoothing artifacts. Nevertheless, the algorithm produces semi-dense (unambiguous) results only, but

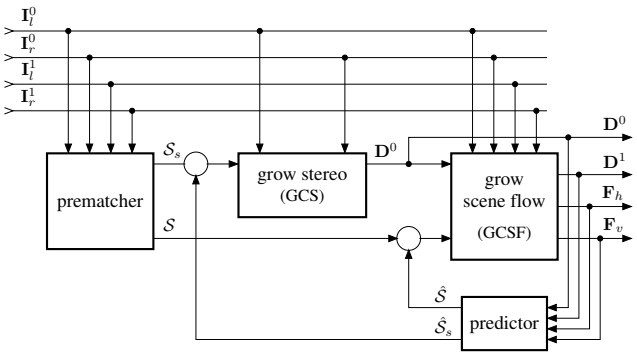


Figure 2: Overview of the proposed algorithm (GCSFs).

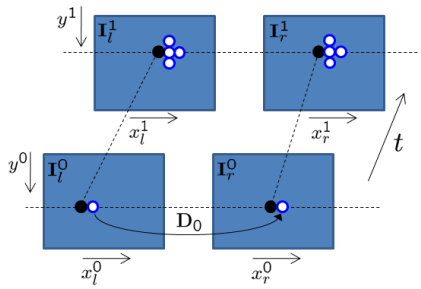


Figure 3: A sequence of consecutive epipolarly rectified stereo images. A seed correspondence $\mathbf{s}$ sketched by filled circles, its right neighborhood $\mathcal{N}_1$ by empty circles.

they are dense enough for many potential applications, see Fig. 1.

The rest of the paper is organized as follows. In Sec. 2, we present the proposed algorithm in details. In Sec. 3, we describe the evaluation and comparison with state-of-the-art methods. Sec. 4 concludes the paper.

2. Algorithm Description

The proposed algorithm for growing correspondences of scene flow in a sequence of stereo images (GCSFs) is summarized in Fig. 2. At each time instance t, it takes as input two epipolarly rectified image pairs, a pair $\mathbf{I}_l^0, \mathbf{I}_r^0$ for time $t-1$ (last frame), and the consecutive pair $\mathbf{I}_l^1, \mathbf{I}_r^1$ for time t (current frame). The output at each time instance is a disparity map $\mathbf{D}^0$ holding the stereo correspondences from the last frame $t-1$, 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 consecutive images $\mathbf{I}_l^0$ and $\mathbf{I}_l^1$.

Notice that having full camera calibration, 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.

First, a prematcher is run to deliver initial correspondences, the seeds. They are used in subsequent growing processes. The prematcher finds sparse correspondences of interest points between left and right images and between consecutive images. 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 $X^0 \in \mathcal{X}^0$ 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. 3. Beside the set of these scene flow seeds, the prematcher also output the stereo seeds $\mathbf{s}_s = (x_l^0, x_r^0, y^0) \in \mathcal{S}_s$ which is a set of two-pixel correspondences between $\mathbf{I}_l^0$ and $\mathbf{I}_r^0$.

Then, the stereo seeds $\mathcal{S}_s$ are grown by a stereo algorithm (GCS), which computes a disparity map $\mathbf{D}^0$ between $\mathbf{I}_l^0$ and $\mathbf{I}_r^0$. Disparity map $\mathbf{D}^0$ together with seeds $\mathcal{S}$ and the input images are an input of the subsequent algorithm (GCSF), which jointly grows disparity map $\mathbf{D}^1$, and the optical flow maps $\mathbf{F}_h, \mathbf{F}_v$.

The solution at time t contains lots of information about the solution at time $t+1$, i.e. when a new frame is available. This information, is exploited in the proposed algorithm by predicting the seeds for the growing processes in the next time instance. Considering the motion of pixels from previous solution, the predictor estimates new correspondence seeds $\hat{\mathcal{S}}$ and $\hat{\mathcal{S}}_s$. These seeds are unified with current seeds given by the prematcher. It means, that starting from the second frame, the growing processes work with larger and richer sets of seeds. The prematcher remains connected for all frames in order to capture the dynamic scene events in which objects suddenly appears. This process is repeated with each subsequent frame.

Details of the algorithm are described below. First, we describe in detail the procedure for growing the scene flow, since it is the essential part. Afterward, we give further details on the rest of the algorithm.

2.1. Growing scene flow (GCSF)

The algorithm is presented in pseudocode as Alg. 1. It takes as input two rectified image pairs $\mathbf{I}_l^0, \mathbf{I}_r^0$ and the consecutive pair $\mathbf{I}_l^1, \mathbf{I}_r^1$, a set of initial correspondence seeds $\mathcal{S}$, a disparity map $\mathbf{D}^0$ for a previous frame $t-1$, and the parameters α (temporal consistency enforcement), β (optical flow regularization), and τ (growing threshold). The output are maps of disparity $\mathbf{D}^1$ and optical flows $\mathbf{F}_h, \mathbf{F}_v$.

First, the algorithm computes a photometric consistency statistic of the 4-pixel 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 cor-

Algorithm 1 Growing the scene flow (GCSF)

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 $\mathbf{F}_h$, $\mathbf{F}_v$.

relations c_{ll}^{01} and c_{rr}^{01} are between consecutive images in the left and right sequences. All the correlations are MNCC statistics [12] on 5×5 pixel 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

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

For all four 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

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

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. 3. Notice the candidates depend on the previous disparity $\mathbf{D}^0$. The other neighbors $\mathcal{N}_2, \mathcal{N}_3, \mathcal{N}_4$ are defined similarly.

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

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

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

Default values of algorithm parameters were found empirically and set to $\alpha = \beta = 0.05$, $\tau = 0.6$ in all our real-data experiments. The value of temporal consistency parameter α in Step 1 is a trade-off between a temporal coherence of the results and an ability to capture fast changes in the motion. We observed that for $\alpha = 0$, the results are not so temporally coherent, certain matches in the 3D surface were randomly disappearing and reappearing due to noise or various degradations in the image sequence. Small $\alpha > 0$ causes that already matched points have a better position in the priority queue and higher chance to be matched. On the other hand, when α is too high, we observed matching errors in sudden changes of object's motion, since wrong (incorrectly predicted) seeds were accepted in Step 4.

Parameter β in (4) regularizes the growing process to handle the aperture problem. When $\beta = 0$, we observed artifacts of the optical flow estimation in edge-like structures. Growing process finds the matches based on local maxima of correlation, which need not necessarily correspond to the correct solution due to various noise in the images. Very small $\beta > 0$ helps. However, when β is too large, the solution is biased towards seeds and locally flat around them.

The last parameter τ directly controls the trade-off between the density of the solution and mismatch rate.

Note that MNCC statistic in (1) is not invariant to deformation of local image neighborhoods between corresponding pixels related by optical flow, which occurs due to camera or scene motion. A general assumption, which is hardly preserved, is a fronto-parallel surface undergoing a fronto-parallel motion [17]. Nevertheless the statistic is insensitive enough to violations of this assumption. We show in the experiments that the algorithm works well under non-trivial motion and non-planar or slanted surfaces. In cases where this could be a problem, a simple extension would be to associate a set of parameters capturing the local affine transformations with the seed, as in [3, 8] in the context of wide-baseline stereo matching.

2.2. Growing stereo (GCS)

A seed growing algorithm [4] for stereo matching between images $\mathbf{I}_l^0$ and $\mathbf{I}_r^0$ is used. The growing procedure is similar in spirit to Alg. 1, however the neighborhoods $\mathcal{N}_i$ are different. This algorithm is reported being not very sen-

sitive to wrong seeds, which is achieved by a robust matching which selects the final solution among competing correspondence hypotheses from the growing process. In the experiments, we compare this algorithm when run frame-by-frame with the same algorithm integrated in the proposed pipeline shown in Fig. 2.

2.3. Prematcher

The task of the prematcher is to deliver sparse correspondences of interest points. This is achieved in our implementation by matching Harris points and tracking them using multi-level version of LK tracker [11]. The stereo seeds $\mathcal{S}_s$ are simply those Harris points which satisfy the epipolar constraint, and whose 5×5 MNCC correlation exceeds threshold τ. The scene flow seeds $\mathcal{S}$ are obtained by tracking the stereo seeds from $\mathbf{I}_l^0$ to $\mathbf{I}_l^1$ and from $\mathbf{I}_r^0$ to $\mathbf{I}_r^1$. The point matches which violates the epipolar constraint between $\mathbf{I}_l^1$ and $\mathbf{I}_r^1$ are discarded from the set.

The algorithm is not limited to Harris seeds. Any other seeds, e.g. from wide-baseline matching of distinguished regions, or other more sophisticated tracking techniques, could be used.

2.4. Predictor

The predictor estimates seeds for processing of the next frame based on the current solution and other assumptions on the motion of points. In our implementation, we use a simple assumption, that the point moves constantly in the image plane, i.e. its optical flow remains the same in a subsequent frame. For each matched pixel (x_l^1, y^1) in $\mathbf{D}^1$, the predicted seed $\hat{\mathbf{s}} = (\hat{x}_l^0, \hat{x}_r^0, \hat{y}^0, \hat{x}_l^1, \hat{x}_r^1, \hat{y}^1)$ is

$$\begin{aligned} \hat{x}_l^0 &= x_l^1, & \hat{x}_l^1 &= x_l^1 + \mathbf{F}_h(x_l^1, y^1), \\ \hat{x}_r^0 &= x_l^1 - \mathbf{D}^1(x_l^1, y^1), & \hat{x}_r^1 &= \hat{x}_r^0 + (\hat{x}_r^0 - x_r^0), \\ \hat{y}^0 &= y^1, & \hat{y}^1 &= y^1 + \mathbf{F}_v(x_l^1, y^1), \end{aligned} \tag{5}$$

where $x_r^0 = x_l^0 - \mathbf{D}^0(x_l^0, y^0)$ and $x_l^0 = x_l^1 - \mathbf{F}_h(x_l^1, y^1)$, $y^0 = y^1 - \mathbf{F}_h(x_l^1, y^1)$. It follows from the output maps in (2). Notice that for stereo seed $\hat{\mathbf{s}}_s = (\hat{x}_l^0, \hat{x}_r^0, \hat{y}^0)$, the disparity map $\mathbf{D}^1$ is only 'translated' into the seed representation and subsequently grown again by stereo [4] to provide new disparity map $\mathbf{D}^0$. This is important since certain pixels may not be matched in $\mathbf{D}^1$ due to motion occlusions, and they are hereby recovered.

The constant motion assumption is rather naïve. More correct would be to use more sophisticated dynamic motion models and Kalman filtering. Nevertheless, despite the simplicity, the predictor usually helps producing enough correct seeds. When the assumption of the constant motion is violated, the affected seeds become wrong with low correlation and they are placed in an unfavorable position in the priority queue. Such regions are grown from other correct seeds (sparse Harris seeds from prematcher, or other seeds where the assumption holds).

2.5. Complexity of the algorithm

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 [5], which 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.

3. Experiments

The experiments demonstrate that the proposed algorithm produces accurate semi-dense results and that it benefits from a joint disparity – optical flow formulation in a sequence of stereo images. The proposed method is compared with a recent spatiotemporal stereo algorithm by Sizintsev et al. [15], with a variational scene flow algorithm by Huguet and Devernay [6], and with a recent optical flow by Brox and Malik [2]. The experiments show that our algorithm is more precise in disparity than [15] and [6], and in optical flow comparable to [6], and slightly inferior to [2].

3.1. Synthetic Data

To quantitatively evaluate and compare the methods, we carried out an experiment with simulated data. The synthetic scene consists of three moving objects: a sphere performing a complicated rotation while moving slowly to the right and away from the cameras, a small vertical bar moving very fast to the left (30 pixels/frame), and a slanted background plane moving towards the cameras. The scene was textured randomly with a white noise, see Fig. 4. The scene was synthesized using Blender. The resulting sequence has 25 frames of stereopair images and each frame has associated ground-truth disparity, optical flow maps, and maps of stereo and motion occlusions.

The algorithms were tested under noise perturbation of data. An independent Gaussian noise was added into each image of the stereo sequence. The experiment was performed with several noise levels, starting from $\sigma = 0$ (no noise) up to $\sigma = 1$ where the variation of the noise is the same as of the image signal.

For all the experiments, we measured an average ratio of correctly matched pixels in non-occluded regions, i.e. number of all pixels without mismatches (error ≥ 1 pixel) and non-matches divided by total number of pixels, over all frames in the sequence. Notice, this evaluation is very strict for algorithms which do not give fully dense results, like

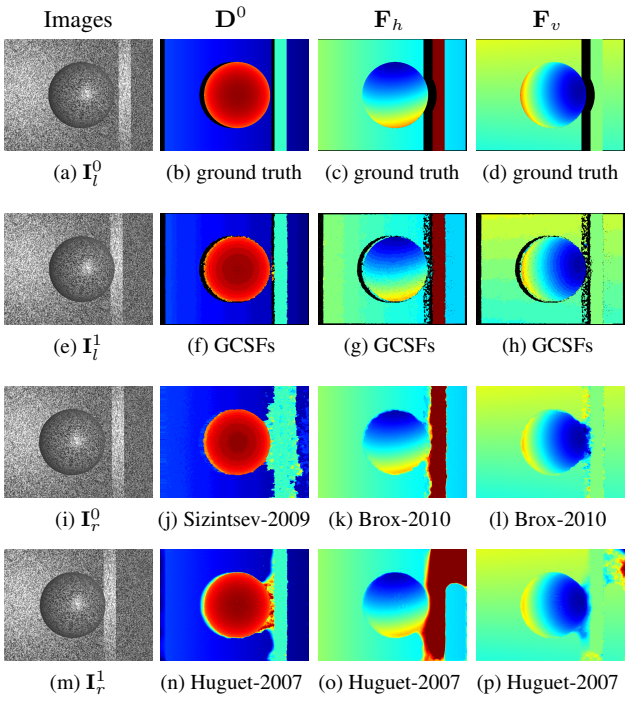
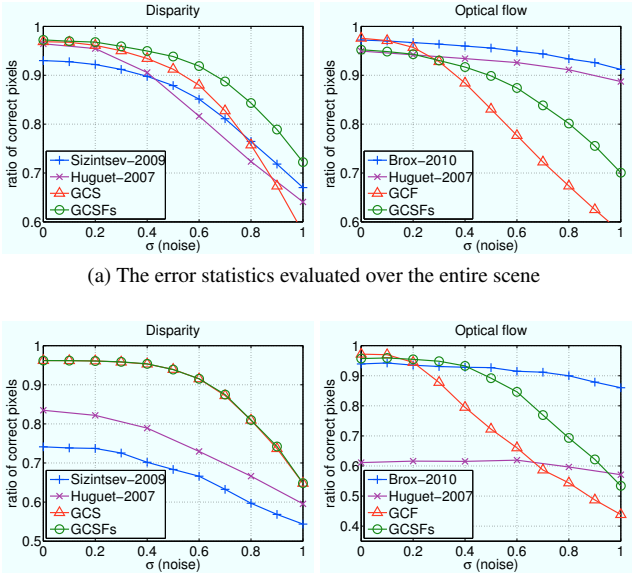


Figure 4: Synthetic experiment. Disparity and optical flow maps of the 6th frame of the sequence: Ground-truth maps with marked occlusions, results of tested algorithms.



(a) The error statistics evaluated over the entire scene

(b) The error statistics evaluated only in the area of the thin vertical bar.

Figure 5: Algorithm accuracy under contamination with a Gaussian Noise. The signal has equal variance as the noise for $\sigma = 1$.

ours. However this is an easy way to simply compare semi-dense and fully-dense results. On the other hand, since the mismatches are counted the same as unmatched pixels, we relax the correlation threshold $\tau = 0$ for all synthetic experiments, other parameters remained of the default values ($\alpha = \beta = 0.05$). This is the only exception in all the experiments in this paper.

This statistic was measured for both disparity and optical flow errors. Optical flow is usually evaluated by average angular error, however the proposed algorithm is of the pixel level accuracy and therefore this usual evaluation would not be suitable. We understand the optical flow as pixel matching problem, similar to stereo without epipolar constraint. It is important to capture gross errors of the optical flow estimates, i.e. mismatches by more than 1 pixel error. This evaluation is again fair for classical sub-pixel optical flow methods, since the ground-truth is provided with a sub-pixel precision.

Results of the experiment are shown in Fig. 5a. In case of stereo, we compared the proposed algorithm (GCSFs) which jointly estimates disparity and optical flow with: a seed growing algorithm which computes disparity maps frame-by-frame independently [4] (GCS), scene flow algorithm [6] (Huguet-2007), and the spatiotemporal stereo [15] (Sizintsev-2009). We can see, there is not much difference for GCSFs and GCS for low level of noise, however the GCSFs is more stable for higher level of noise. Algorithm [15], while performing well in slow moving regions, has severe difficulties with the quickly moving bar even without noise, see Fig. 4j, which causes its inferior performance compared to the proposed method. Algorithm [6] has also severe difficulties with this scene. Corresponding disparity map of GCSFs is shown in Fig. 4f. We can see no significant mismatches in either part of the scene, object boundaries are well preserved except for small phenomena due to fluctuations of the window similarity statistic. There are also small mismatches in occluded regions, since the threshold τ is relaxed, but they are not included in the evaluation.

In case of optical flow, we compared the flow provided by proposed GCSFs algorithm with another seed growing algorithm which frame-by-frame independently searches the stereo-correspondences without epipolar constraint (GCF). This growing mechanism was used in [3]. Additionally we compare this with a recent variational method which can handle large displacement [2] (Brox-2010) and with the scene flow [6] (Huguet-2007). We can see, the results are even slightly better without noise for GCF then for GCSFs. This is because GCF allows non-bijective matching, while GCSFs insists on uniqueness which may cause small 1-pixel gaps of unmatched pixels between different motion layers. However, with increasing level of noise GCSFs outperforms its frame-by-frame seed growing counterpart. Results of [2] an [6] are compara-

ble with GCSFs for low level of noise. For stronger noise these methods are significantly better than GCSFs. This is natural, since these global methods have reported excellent properties under perturbation by this kind of noise. Optical flow maps of GCSFs are shown in Fig. 4g–4h. Object and motion boundaries are well preserved, there are no clear mismatches, there are a few 1-pixel gaps as mentioned above. Notice that, the motion occlusion on the bar, which is due to its motion behind the sphere in the next frame, has a 'correct' motion estimate, despite there is no evidence in data. This is a side effect of the prediction. Optical flow maps of [2] are shown in Fig. 4k–4l. They are very precise inside the objects, however visually, there are some imperfections in motion boundaries of the objects.

Although the plot of [6] suggests its good overall performance, there are strong artifacts around the quickly moving bar, see Fig. 4o–4p. Since the bar is relatively small with respect to the rest of the image, where the algorithm performs excellently, the error statistics do not reflect visually disturbing artifacts. Therefore, we evaluated the error statistics additionally in the area of the vertical bar only, see Fig. 5b. Then, we can see the low performance of [6] compared to other algorithms.

The favorable results of the proposed GCSFs algorithm compared to the frame-by-frame independent seed growing methods are a consequence of: (1) joint disparity and optical flow estimates which constrain each other, and (2) good temporal consistency and coherence. The mechanism is the following. When data are weak due to noise, there is a lack of correctly matched seeds and the growing process is either stopped early (by the condition in Step 7 of Alg. 1) for conservative choice of threshold τ, or produces mismatches if τ is relaxed. However, if we feed partially grown disparity and optical flow maps as the seeds to GCSF algorithm (using the predictor), it grows them further if they were correct. This effect is repeated, and after certain number of frames, high quality seeds are accumulated.

3.2. Real data

The proposed algorithm was tested on real data as well. For all these experiments, we used default values of parameters of the proposed GCSFs algorithm, $\alpha = \beta = 0.05$, $\tau = 0.6$. We show results on CAVA dataset of INRIA[1], where the stereo camera is static, and on the dataset of ETH Zürich[2] acquired by a mobile stereo platform.The results of tested algorithms are shown in Fig. 6 and 7 as disparity $\mathbf{D}^1$ and optical flow $\mathbf{F}_h, \mathbf{F}_v$ maps.

For INRIA dataset, the results of the proposed GCSFs algorithm, Fig. 6b–6d, 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 women 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.

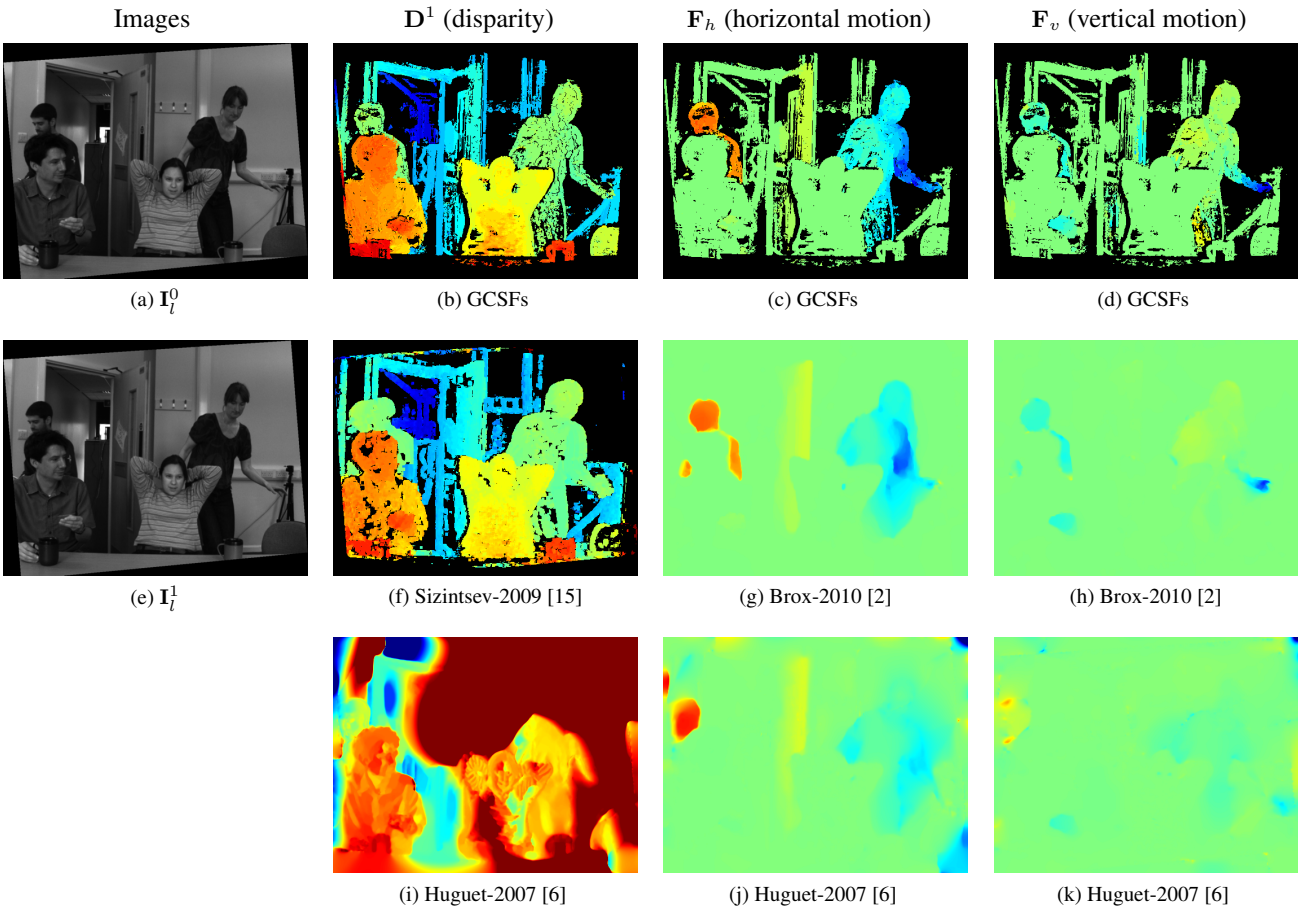
ETH dataset represents a complex scene with both camera forward motion and motion of pedestrians. There are up to 30 pixel displacements between consecutive frames. In our results, Fig. 7b–7d, we can see a motion of the planar sidewalk close to cameras and well captured depth and motion boundaries of the people walking. There are only few small mismatches which are visible in disparity map. This is in the region of the leftmost building which effects complicated non-Lambertian mirror like reflections. Some small mismatches can be found in optical flow in edge-like structures, which are consequence of improperly handled aperture problem.

Results of the spatiotemporal stereo [15] can be seen in Fig. 6f and Fig. 7f. Disparity maps were thresholded according to a stequel significance map to remove spurious matches. The threshold was set to 0.4 according to author's recommendation. After the thresholding, the disparity map on INRIA has roughly the same density as our result. However, the results are not so precise. It seems that all objects are fattened and especially those which moves in front of the weakly textured regions, see the walking woman and the man coming through the door in Fig. 6f. These artifacts are probably caused by the large spatiotemporal extent of the matching elements (stequels). The method has severe difficulties with the ETH sequence. The part of the scene which is close to cameras and hereby undergoes a fast motion is not captured by this algorithm, Fig. 7f. Matching of stequels probably does not work well for large displacement between frames.

Results of the large displacement optical flow [2] are shown in Fig. 6g–6h and Fig. 7g–7h. They are more or less consistent with our results, but they are fully dense. The motion boundaries seem to be a little bit fuzzy, but this could be only in the motion occluded regions, where there is no evidence in data. There are a few small patchy mismatches in ETH.

Results of the variational scene flow algorithm are shown [6] in Fig. 6i–6k and Fig. 7i–7k. The disparity maps are erratic, the algorithm fails dramatically in stereo for these scenes. This failure is probably due to a complexity of the scene (many occlusions, complicated motions, and varying strength of the texture), and perhaps also due to improper initialization and consequent problems with convergence. The optical flow given by this method is surprisingly much better than the stereo disparity. Nevertheless, we can see typical artifacts of smoothed motion boundaries, which is a consequence of the prior term winning over the data.

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



(a) $\mathbf{I}_l^0$ (b) GCSFs (c) GCSFs (d) GCSFs

(e) $\mathbf{I}_l^1$ (f) Sizintsev-2009 [15] (g) Brox-2010 [2] (h) Brox-2010 [2]

(i) Huguet-2007 [6] (j) Huguet-2007 [6] (k) Huguet-2007 [6]

Figure 6: Real experiments: Results on INRIA dataset. This figure is better seen in the electronic version of the paper.

GCSFs	1.5 seconds
Sizintsev-2009 [15]	35 seconds
Brox-2010 [2]	3 minutes
Huguet-2007 [6]	3 hours

Table 1: Average running time per frame of VGA images.

For both sequences, our results are temporally coherent without flickering artifacts, which is not the case of results using [15] and [6]. Results of [2] are fairly stable temporally, despite computed frame-by-frame.

3.3. Running time of tested algorithms

An average running time per frame of the tested algorithms is shown in Tab. 1. These times were measured on our synthetic sequence of 640×480 images, using a standard PC (Intel Core 2 2.6 GHz, 6 GB memory, Linux). Our GCSFs algorithm is faster by order of magnitudes than the other tested methods. Our implementation is not optimized and partially in Matlab. For the other algorithms we had binaries.

4. Conclusions

We presented an algorithm which jointly estimates semi-dense disparity and optical flow 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.We tested with two different publicly available datasets and performed a quantitative ground-truth experiment. We made a fair comparison with state-of-the art methods spanning over spatiotemporal stereo, and variational methods for optical and scene flow.

The proposed algorithm is a practically well working trade-off between simple local methods and theoretically sound global MRF algorithms, since local relations between adjacent pixels are considered. It can be also viewed as a 'semi-supervised' matching algorithm, where a few initial seeds are propagated. We plan to investigate properties of this propagation (growing) as a diffusion process on the correspondence manifold.

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

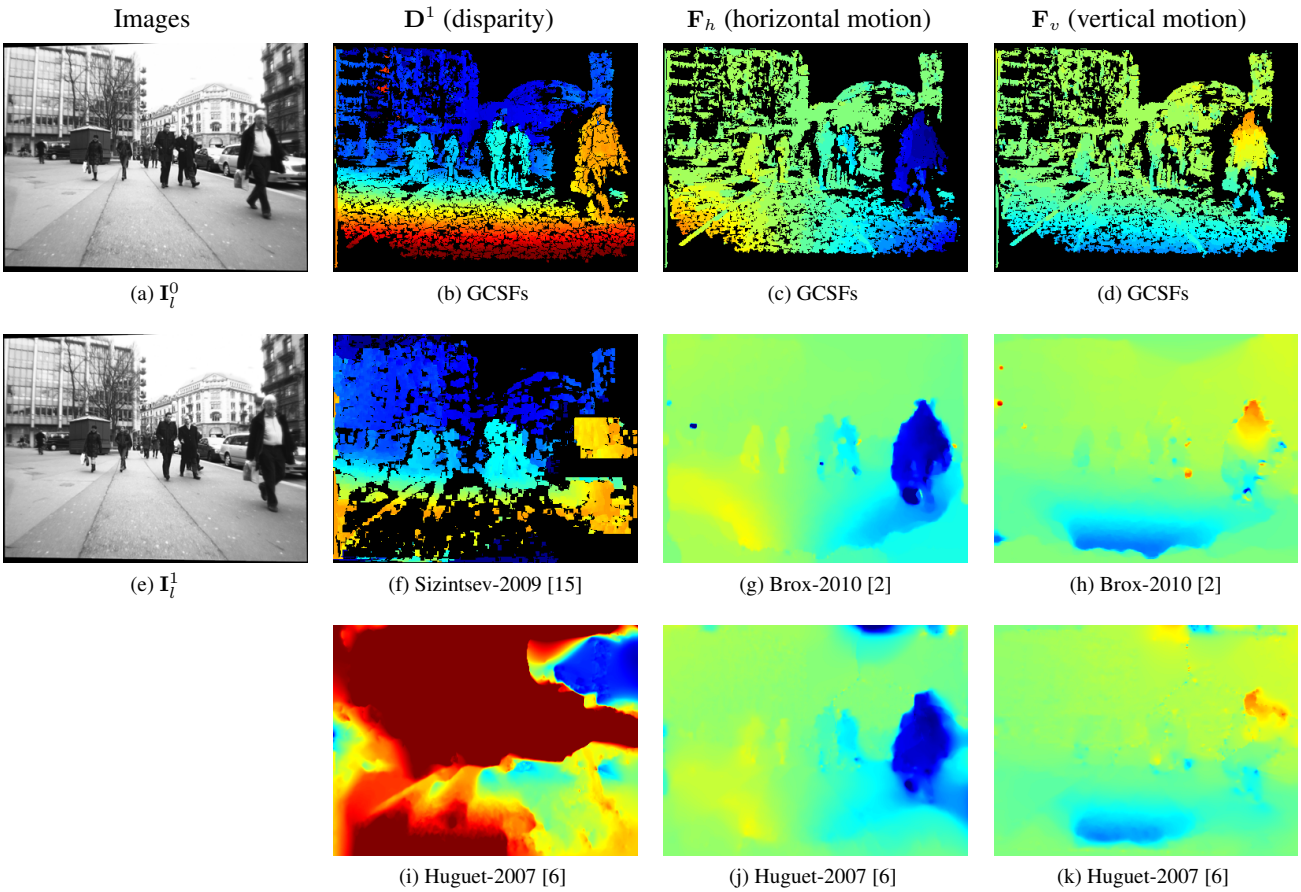


Figure 7: Real experiments: Results on ETH dataset. This figure is better seen in the electronic version of the paper.

References

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

[2] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Trans. on PAMI*, 2010. In press.

[3] J. Čech, J. Matas, and M. Perďoch. Efficient sequential correspondence selection by cosegmentation. *IEEE Trans. on PAMI*, 32(9), 2010.

[4] J. Čech and R. Šára. Efficient sampling of disparity space for fast and accurate matching. In *BenCOS Workshop, CVPR*, 2007.

[5] M. Gong. Real-time joint disparity and disparity flow estimation on programmable graphics hardware. *CVIU*, 113(1), 2009.

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

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

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

[9] M. Lhuillier and L. Quan. Match propagation for image-based modeling and rendering. *IEEE Trans. on PAMI*, 24(8), 2002.

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

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

[12] H. P. Moravec. Towards automatic visual obstacle avoidance. In *IJCAI*, page 584, 1977.

[13] J.-P. Pons, R. Kerive, O. Faugeras, and G. Hermosillo. Variational stereovision and 3D scene flow estimation with statistical similarity measures. In *ICCV*, 2003.

[14] 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.

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

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

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