# Robust Spatiotemporal Stereo for Dynamic Scenes

Jordi Sanchez-Riera, Jan Čech, and Radu Horaud
INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, FRANCE
{jordi.sanchez-riera, jan.cech, radu.horaud}@inria.fr

## Abstract

*Stereo matching is a challenging problem, especially in the presence of noise or of weakly textured objects. Using temporal information in a binocular video sequence to increase the discriminability for matching has been introduced in the recent past, but all the proposed methods assume either constant disparity over time, or small object motions, which is not always true. We introduce a novel stereo algorithm that exploits temporal information by robustly aggregating a similarity statistic over time, in order to improve the matching accuracy for weak data, while preserving regions undergoing large motions without introducing artifacts.*

## 1 Introduction

The use of a stereo video presents advantages over single-frame stereo. The extra information may help the matching process to improve accuracy and to enforce temporal consistency of disparity maps in the case of weakly textured regions or when the matching is ambiguous. We see three main approaches in the literature.

The first group computes the *scene flow*. These methods simultaneously estimate depth and motion. The is a formulation of the coupled estimation problems of disparity (between a stereo pair) and optical flow (between consecutive frames) which mutually constrain each other. This task is traditionally solved by variational [1], MRF [7], or seed-growing [4] methods.

In the second group, there are methods which rely on *independent motion estimates* to improve the stereo matching. In [2], the disparity maps are filtered by a median filter along pixel trajectories obtained by an external optical flow module. Independent optical flow is also used in [13], where the authors propose an MRF framework with an extra term which penalizes discrepancies in photoconsistency of the (optical flow related) neighbourhood in the current-, previous-, and next frames. Similar MRF formulation [6] additionally disconnects the edges to prevent over-smoothing in case of large motion and failure of the optical flow estimates.

The third group is composed of methods of *spatiotemporal stereo* that do not estimate the motion explicitly, but exploit a local spatiotemporal neighbourhood of pixels to increase discriminability of the similarity statistics. Paper [5] projects an artificial pattern varying over time onto the static scene and temporally aggregate the statistic. The similarity statistic (based on bilateral filtering) is temporally aggregated also in [9], such that adjacent frames are weighted by a Gaussian kernel to cope with a small motion. In [12] the authors study how spatiotemoral windows are deformed due to surface slant and motion and propose an optimization framework to find the distortion parameters and invariant similarity statistic. Alternatively, the same insensitivity is achieved in [10, 11] by representing the image using Gabor filter responses and the similarity statistic is computed in a closed form. However, all these methods assume that the disparity between frames changes only slowly. In reality this assumption is not valid near object boundaries or in the presence of rapidly moving objects, which causes serious artifacts.

The main contribution of this paper is a method of spatiotemporal stereo which benefits from aggregating the similarity statistic over a time window. Unlike previous work on spatiotemporal stereo algorithms, the proposed method is robust to abrupt temporal changes in disparity due to large motions. The main idea of our algorithm is to automatically detect image regions corresponding to such changes, such that the aggregation of the similarity statistic over the time window is disconnected for these regions. The algorithm is implemented within the efficient seed growing procedure [3].

## 2 Method description

Let us have two synchronized and epipolar-rectified video streams $\mathbf{I_l}(x, y, t)$ and $\mathbf{I_r}(x, y, t)$. Variables $x$ and $y$ index respective horizontal and vertical image coordinates in pixels, $t$ indexes the time, *i.e.* a

frame of the streams. The streams are related by a disparity function $d(x, y, t)$ which assigns the correspondences between pixels in the left and right image

$$\mathbf{I_l}(x, y, t) \approx \mathbf{I_r}(x + d(x, y, t), y, t). \quad (1)$$

A matching algorithm must measure a certain similarity statistic between potentially corresponding pixels to establish the correspondences. The simplest statistic is a difference of pixel intensities, which is however ambiguous. More discriminable statistics use a small neighbourhood (a window) around the potential correspondence. Then, these algorithms locally approximate the disparity function. For instance, paper [12] uses a linear approximation. In a small spatiotemporal neighbourhood $\mathcal{N}$ around location $(x_0, y_0, t_0)$, *e.g.* a 3D window of $5 \times 5$ pixels over 3 frames, the disparity function is $d(x, y, t) \approx \hat{d}(d_i, d_0, d_1, d_2, d_t) = d_i + d_0 + d_1(x - x_0) + d_2(y - y_0) + d_t(t - t_0)$. Then they use an optimized statistic to measure a photometric consistency of the potential correspondence for candidate disparities $d_i$

$$\text{TSSD}(x_0, y_0, t_0, d_i) = \min_{d_0, d_1, d_2, d_t} \sum_{(x,y,t)\in\mathcal{N}} \Big(\mathbf{I_l}(x, y, t) - \mathbf{I_r}(x + \hat{d}(d_i, d_0, d_1, d_2, d_t), y, t)\Big)^2 \quad (2)$$

to compensate the distortion which occurs due to subpixel displacement $d_0$, surface slant $d_1$, $d_2$, and temporal disparity change $d_t$.

However, there are several sources of errors in this approach: (i) Tendency to get stuck in a local extrema; (ii) Not a significant gain in discriminability[1] over the case where $d_0 = d_1 = d_2 = d_t = 0$, since the statistic is improved by the optimization for both correct and incorrect matches; (iii) When the assumption on the linearity of the disparity function within the local spatiotemporal neighbourhood is violated (*e.g.* abrupt change in disparity), the method fails dramatically.

Therefore we adopted a different approach in the aggregation. As an elementary similarity statistic, we use Moravec's normalized cross correlation [8],

$$\text{NCC}(x_0, y_0, t_0, d_i) = \frac{2\,\text{cov}\big(\mathbf{W_l}(x_0, y_0, t_0), \mathbf{W_r}(x_0 + d_i, y_0, t_0)\big)}{\text{var}\big(\mathbf{W_l}(x_0, y_0, t_0)\big) + \text{var}\big(\mathbf{W_r}(x_0 + d_i, y_0, t_0)\big) + \epsilon}, \quad (3)$$

where $\mathbf{W_l}(x_0, y_0, t_0)$ is a spatial window (a subimage) of $N \times N$ pixels centered at position $(x_0, y_0)$ of the frame $t_0$ of the left stream. Similarly $\mathbf{W_r}$, and $\epsilon$ is a machine epsilon to prevent instability of the statistic in case of low intensity variance. The statistic has consequently low response in textureless regions [3].

Then the NCC statistic is aggregated over a symmetric time window of $2T + 1$ frames, such that

$$\text{TNCC}(x_0, y_0, t_0, d_i) = \frac{1}{2T+1} \sum_{t=t_0-T}^{t_0+T} \text{NCC}(x_0, y_0, t, d_i). \quad (4)$$

Apparently, the TNCC statistic is decayed when the disparity changes significantly within the temporal window. Notice that the motion in general is not harmful, but the motion changing the disparity is. A typical distribution of NCC$(t)$ statistic over the time $t$ of the window for a correct match $(x_0, y_0, t = 0, d_i)$ is the following: If the disparity is constant over time, all per-frame correlations for $t = \{-T, \ldots, T\}$ are high. If the disparity changes slowly, the correlation is slightly lower more faraway from the central frame. However, when the disparity changes rapidly, the correlation off the central frame drops quickly, since the other correlations measure a photometric consistency of locations which are not corresponding any more.

On the other hand, a potential mismatch (*i.e.* wrong correspondence) has the distribution of per-frame correlations over the time window such that the correlations are low, but due to random fluctuations or texture self-similarity there may be high responses for any frame of the temporal window. The temporal aggregation in (4) averages out these excesses and decreases their correlations and hereby increases the discriminability.

However, it is important to detect phenomena corresponding to large changes in disparity and in these cases to use the central correlation only without any aggregation which would cause artifacts. Therefore, we propose a robust temporal normalized cross correlation

$$\text{RTNCC}(x_0, y_0, t_0, d_i) = \begin{cases} \text{NCC}(x_0, y_0, t_0, d_i) & \text{if } \big(\text{NCC}(t_0) - \text{NCC}(t_0 \pm 1)\big) \geq \alpha, \\ \text{TNCC}(x_0, y_0, t_0, d_i) & \text{otherwise.} \end{cases} \quad (5)$$

This means that RTNCC uses the correlation (3) of the central frame, if it is higher than correlations of adjacent frames by threshold $\alpha$. Otherwise, RTNCC uses the average correlation TNCC over the entire temporal window (4). For simplicity of notation, we omitted all other indexes $x_0, y_0, d_i$ in the condition in (5).

**Matching algorithm.** To establish the matching between stereo images, the proposed RTNCC statistic was integrated in a seed growing procedure [3]. This algorithm uses seed correspondences obtained by match-

[1] The discriminability of the similarity statistic is proportional to a probability that the statistic has better response for the true correspondence than for the incorrect ones.

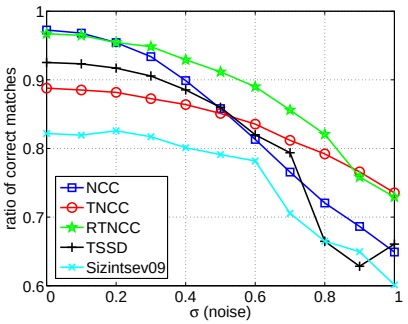



**Figure 1. Quantitative evaluation.**

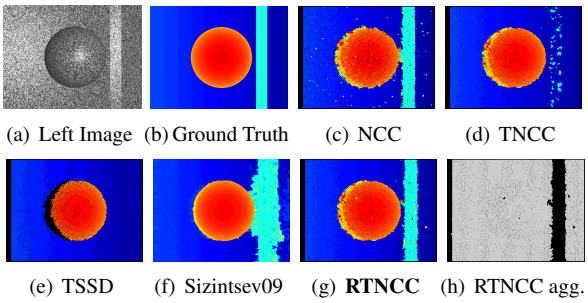



**Figure 2. Synthetic dataset of [4]. Frame 6, noise level $\sigma = 0.5$. Disparity maps.**

ing Harris points. For each seed the algorithm searches other correspondences in their surroundings by maximizing the similarity statistic. If the similarity statistic of a candidate exceeds a threshold, then a new correspondence is found. It becomes a new seed and the process repeats until there are no more seeds to be grown.

Besides low computational complexity, the advantage of the algorithm in our context is the input of seeds. Namely, we observed the condition in (5) of RTNCC is reliable in textured regions only. Nevertheless, the seed correspondences are points with the Harris property and for them the decision works well. Therefore, we propose to take this decision for the initial seeds only. Each seed then propagates a flag indicating whether the aggregation in RTNCC is used or not and this flag is inherited by its 'offspring' seeds in the growing process.

## 3 Experiments

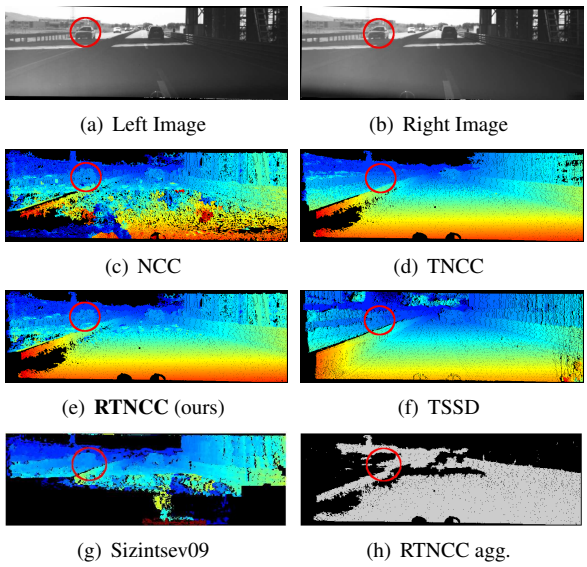
We performed a set of experiments to demonstrate that the proposed algorithm can cope with weak or ambiguous data and images corrupted by noise, without introducing artifacts of smoothing boundaries of rapidly moving objects. We compare the proposed method (RTNCC) with two baseline instances of the growing algorithm: (i) the algorithm which uses the spatial neighbourhood only for matching (NCC), and (ii) the algorithm which trivially uses the spatio-temporal neighbourhood, such that all per-frame correlations are averaged (TNCC). The other comparisons are with two state-of-the art spatio-temporal methods: (i) temporal SSD optimization [12] integrated in the growing algorithm (TSSD), and (ii) the stequel matching algorithm [10] (Sizintsev09).

For all experiments, we used $5 \times 5$ pixel windows as the spatial neighbourhood of all statistics, parameter $\alpha$ in RTNCC (5) was empirically set to $0.8$. For the short synthetic sequence, we set temporal window half-size to $T = 2$, while for the real one it was set to $T = 7$.

**Ground-truth experiment.** To quantitatively compare the different algorithms, we tested on a stereo sequence with ground-truth disparity maps used in [4], Fig. 2(a). It consists of three objects: a background plane, a sphere, and a thin bar. The plane and sphere move slowly, while the thin bar moves rapidly (about 30 pixels per frame) from right to left crossing the entire scene. It is textured randomly with a white noise.

We measured ratio of correctly matched pixels in non-occluded regions, *i.e.* number of all pixels without mismatches (error $\geq 1$ pixel) and unmatched pixels divided by the total number of pixels. The input images were perturbed with zero mean additive Gaussian noise with successively increasing standard deviation $\sigma$. The noise has equal variance as the signal for $\sigma = 1$.

In Fig. 1, we can see the algorithm using NCC performs very well without noise. The texture is optimal and hence the correlation is very high and unambiguous. However, it degrades with noise, Fig. 2(c), producing mismatches as small spatial only image windows do not correlate well. The TNCC degradates slowly with increasing level of noise, however the ratio of correct matches is lower since it tends to completely miss the rapidly moving bar, Fig. 2(d). The temporal aggregation helps to filter out the noise in slowly moving regions, but the aggregation is harmful for the bar where the disparity changes abruptly over time, since for this region TNCC of the false background wins over TNCC of the true bar. Similarly, the other two methods Sizintsev09 and TSSD perform well filtering the noise but both of them have serious problems with the rapidly moving bar where the disparity changes abruptly over time, Fig. 2(e), 2(f). The proposed RTNCC performs the best. It is as good as the NCC for low noise, and it is always superior to other methods with increasing levels of noise. It has fewer mismatches in slowly moving regions, but at the same time it preserves the rapidly moving bar without artifacts, Fig. 2(g). It correctly aggregates over time using TNCC in regions where it helps, and for other regions it uses the spatial statistic NCC.



(a) Left Image (b) Right Image

(c) NCC (d) TNCC

(e) **RTNCC** (ours) (f) TSSD

(g) Sizintsev09 (h) RTNCC agg.

**Figure 3. DAGM Exposure Challenge dataset. Disparity maps.**

The map in Fig. 2(h) shows which case in (5) was used in results of RTNCC. Pixels matched using the temporal aggregation are indicated by gray colour, pixels matched by spatial statistic by black colour. We can see, it correctly used the spatial statistic NCC for the region of the bar, while for other pixels, it correctly used the temporal aggregation TNCC.

**Real outdoor scene.** To show the validity of the proposed algorithm on a real outdoor scene we tested under the DAGM Exposure Changes dataset[2] (DAGM). The stereo camera is in a car driving in a highway in difficult lighting conditions, sudden exposure changes. Cars going in the opposite lane moves very fast, see Fig. 3(a).

We show results for the frame, where the car is passing under the bridge. The texture of the road almost disappears. It causes the spatial statistic NCC to fail, producing many mismatches, as in Fig. 3(c). The spatiotemporal version TNCC works better, Fig. 3(d). Much information is retained due to the temporal aggregation. Notice that the disparity of the road remains constant over time and this is also the case of the car going in the same direction, since the distance to it is more or less constant. However, the problem is, that the car going in the opposite direction (in red circle), whose relative velocity is very high, is missed by the TNCC. This is the same effect as the case of the rapidly moving bar in Fig. 2(d). TSSD has similar difficulties there, Fig. 3(f). Surprisingly, algorithm Sizintsev09 has severe problems with all rapidly moving pixels in the scene, including those where the disparity remains constant. It produces large artifacts in regions near the camera. The proposed RTNCC works well, Fig. 3(e). It is significantly superior to NCC and all objects, including the car in the opposite direction, are preserved.

[2] www.dagm2011.org/adverse-vision-conditions-challenge.html

## 4 Conclusions

We presented a spatiotemporal correlation statistic that increases the discriminability by aggregating over time and hereby produces higher quality matching results. We showed the proposed method is robust to a rapid motion in the scene, which is a situation where the state-of-the-art algorithms are prone to artifacts.

**Acknowledgements.** This research was supported by EC project FP7-ICT-247525-HUMAVIPS.

## References

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

[2] M. Bleyer and M. Gelautz. Temporally consistent disparity maps from uncalibrated stereo videos. In *ISPA*, 2009.

[3] J. Cech, J. Matas, and J. Perdoch. Efficient sequential correspondence selection by cosegmentation. *IEEE Trans. PAMI*, 32(9), 2010.

[4] J. Cech, J. Sanchez-Riera, and R. P. Horaud. Scene flow estimation by growing correspondence seeds. In *CVPR*, 2011.

[5] J. Davis, D. Nehab, R. Ramamoorthi, and S. Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. *IEEE Trans. PAMI*, 27(2), 2005.

[6] E. S. Larsen, P. Mordohai, M. Pollefeys, and H. Fuchs. Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In *ICCV*, 2007.

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

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

[9] 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.

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

[11] M. Sizintsev and R. P. Wildes. Spatiotemporal oriented energies for spacetime stereo. In *ICCV*, 2011.

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

[13] J. Zhu, L. Wang, J. Gao, and R. Yang. Spatial-temporal fusion for high accuracy depth maps using dynamic MRFs. *IEEE Trans. PAMI*, 32(5), 2010.