A SIMPLE AND EFFICIENT WAY TO COMPUTE DEPTH MAPS FOR MULTI-VIEW VIDEOS

Gaël Sourimant

INRIA Rennes - Bretagne Atlantique, Campus de Beaulieu, 35042, Rennes, France

ABSTRACT

This paper deals with depth maps extraction from multi-view video. Contrary to standard stereo matching-based approaches, depth maps are computed here using optical flow estimations between consecutive views. We compare our approach with the one proposed in the Depth Estimation Reference Software (DERS) for normalization purposes in the ISO-MPEG 3DV group. Experiments conducted on sequences provided to the normalization community show that the presented method provides high quality depth maps in terms of depth fidelity and virtual views synthesis. Moreover, being implemented on the GPU, it is far faster than the DERS.

***Index Terms* —** Depth maps, disparity maps, Multi-view videos, 3D Videos, 3DTV, Normalization, GPGPU, Optical flow

1. INTRODUCTION

In the future 3DTV broadcasting scheme, many display types will be available, leading to many required input 3DTV content. However, at acquisition time, it is inconceivable to capture specific content for all possible targetted displays. That is why an intermediate three-dimensional representation of the viewed scene is necessary. The fundamental 3D information attached to multi-view videos remains the depth map, leading to multi-view plus depth (MVD) videos. These depth maps can be further processed and transformed into other 3D representation, for instance for coding efficiency purposes. Depth maps extraction from multi-view content is thus a mandatory and challenging step, since the quality of the computed maps impacts all the remaining broadcasting chain, from coding to rendering quality.

In this paper, we present a different point of view of depth maps generation from the standard stereo matching-based approach, using high quality optical flow estimation algorithms. The context of normalization works to which our method can be compared to is presented in Section 2. The optical flow-based approach to depth maps estimation is briefly reviewed in Section 3, and results are presented in Section 4.

2. MULTI-VIEW NORMALIZATION CONTEXT

Ongoing works for future 3DTV complete framework normalization deal with several aspects: acquisition of multi-view content, 3D representation, intermediate views generation, coding / compression and transmission, etc. The main 3D representation in use is here the depth map. As such, extraction of this information has been addressed by the MPEG community under the form of a Depth Estimation Reference Software (DERS). This software transforms multi-view videos into multi-view plus depth videos (MVD). Evaluation of such generated depth maps is performed through virtual views synthesis using another reference software: VSRS, the View Synthesis Reference Software.

We assume here the same acquisition context than the one proposed by the MPEG normalization group. We consider that input videos are recorded by an aligned camera bank, with non converging cameras. Their image planes are thus aligned to the same virtual 3D plane. We notice that such recording process is very difficult to set up, and thus input images are generally corrected to be perfectly aligned not only geometrically speaking, but also chromatically.

Depth estimation. Depth maps estimation within the DERS is mainly inspired by the works belonging to the stereo matching community of the past few years. To simplify, disparities are estimated in three different steps [1]:

1. Local search of pixel matches along image lines
2. Global optimization over the whole image
3. Post-processing

The DERS has been adapted here to fit the requirements of the multi-view context. Instead of using only two views (left and right), three input view (left, central and right) are considered. The disparity map is computed for the central view using motion estimation from both the left and right views, to tackle efficiently with occluded zones for instance. Implicitly, this framework imposes that motions from left or right views be equivalent and thus that the three cameras are perfectly aligned, the central one being at an equal distance from the two others. This depth estimation framework for the DERS is illustrated on Figure 1 (here, for illustration purposes, we show depth maps for the sequence *Cafe* computed with our method, described in Section 3). For instance, the disparity map for view 3 is computed using pixels motion between view 3 and views 2 and 4.

Figure 2 illustrates depth maps extraction results for two test sequences: *Newspaper* and *Book Arrival*. Disparity maps are encoded in greyscale images: dark values indicate pixels far away from the cameras while bright values depict near objects. Depths maps for *Newspaper* are computed in an automatic way while manual disparity data have been integrated in the estimation for *Book Arrival*. Some disparities are badly estimated (*Newspaper*: pixels on the background above the left-most character are noted much nearer than they should do). Some inconsistencies between the maps can also be noticed (*Newspaper*: on the top-right part of the background ; *Book Arrival*: on the ground, to the bottom-right side).

This disparity estimation phase is crucial, since it impacts on all the remaining steps of the chain. To our point of view, the DERS

These works have been supported by the French national project *Futurim@ges*

978-1-4244-6379-4/10/$26.00 ©2010 IEEE

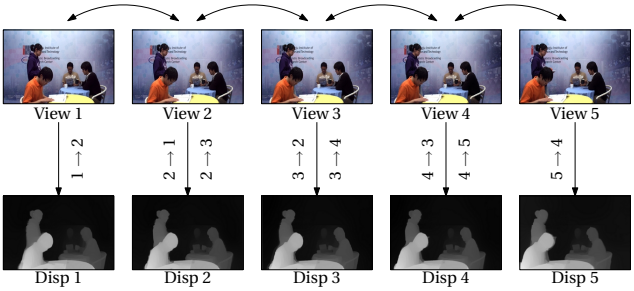
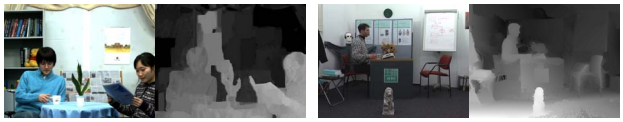


Figure 1: Depth estimation framework for the DERS



(a) *Newspaper* (b) *Book Arrival*

Figure 2: Example of disparity maps extraction for two MPEG test sequences with our method

did not behave well enough as is. We thus present in Section 3 another software similar to the spirit of DERS. It is based on Werlberger's optical flow algorithm [2].

View Synthesis. Evaluation of disparity or depth maps can be performed using the following protocol. Two non neighboring views and their associated maps are considered. The VSRS is used to generate an intermediate view, corresponding to one of the acquired views. These two views — the generated one and the acquired one – are then compared using an objective metric. This metric can be for instance the PSNR, or the more recent PSPNR measure, which aims to consider perceptual differences between the images.

The VSRS uses as input data two videos and their associated disparity maps (which are converted internally to depth maps), corresponding to two different acquired views. It also needs the camera parameters of these two views (both intrinsic and extrinsic). A central view is generated, using the camera parameters which are desired for this virtual point of view and the other input data. No 3D information is generated with the VSRS, only 2D images.

3. OPTICAL FLOW-BASED APPROACH

Principle and Werlberger's method. Based on the observation that a disparity field estimation is nothing else but a dense motion estimation between two images, we explored a new path towards recent optical flow estimation methods. These optical flow algorithms are a part of the 2D motion estimation algorithms. Their particularity comes from the fact that they seek to find the projected relative motion of scene points with regards to the cameras, which should be the closest possible to the "true" projected motion. With such definition, they can be opposed to motion estimation methods used in the video coding community, which aim at finding motion vectors that minimize an motion-compensated image difference in a local search window. For instance, in large textureless regions, if all motion vectors are equivalent in terms of image difference, they will favor null vectors since they are easier to encode, which will not represent the true 2D motion.

We now derive the Optical Flow Constraint (OFC) in a more formal way, going to Werlberger's formulation. The essential principles are explained. Much more details can be found in [3]. The OFC comes from the brightness consistency between two images I_1 and I_2. The brightness of a pixel $\mathbf{x}$ in I_1 should be equal to the brightness of the matching pixel displaced by a motion vector $\mathbf{u}(\mathbf{x})$ in I_2:

$$I_1(\mathbf{x}) = I_2(\mathbf{x} + \mathbf{u}(\mathbf{x})) \tag{1}$$

By linearizing this brightness consistency constraint with a Taylor expansion, and dropping the negligible second- and higher-order terms, one gets the OFC:

$$\mathbf{u}(\mathbf{x})^\top \nabla I_2(\mathbf{x}) + I_2(\mathbf{x}) - I_1(\mathbf{x}) = 0 \tag{2}$$

Horn & Schunk showed that solving for $\mathbf{u}$ can be performed in an energy minimization framework [4]:

$$\bar{\mathbf{u}} = \underset{\mathbf{u}}{\operatorname{argmin}}\, E(\mathbf{u}) = \underset{\mathbf{u}}{\operatorname{argmin}}\, \left(E_{\text{data}}(I_1, I_2) + E_{\text{prior}}(\mathbf{u})\right) \tag{3}$$

Starting from this model, one can setup the energy formalization as a disparity preserving and spatially continuous formulation of the optical flow problem, based on a L^1 data term and an isotropic Total-Variation regularization term [5, 6]:

$$E_{\text{data}}(I_1, I_2) = \lambda \int_\Omega |I_2(\mathbf{x} + \mathbf{u}(\mathbf{x})) - I_1(\mathbf{x})|\, dx\, dy \tag{4}$$

$$E_{\text{prior}}(\mathbf{u}) = \int_\Omega |\nabla \mathbf{u}_x| + |\nabla \mathbf{u}_y|\, dx\, dy \tag{5}$$

Here λ balances data and prior terms, Ω represents the image domain, and $\nabla \mathbf{u}$ is the spatial gradient of the motion field. By linearizing the data term, one gets a convex optimization problem:

$$E_{\text{data}}(I_1, I_2) = \lambda \int_\Omega |\rho(\mathbf{u}(\mathbf{x}))|\, dx\, dy, \tag{6}$$

with $\rho(\mathbf{u}(\mathbf{x}))$ being the Optical Flow Constraint from Equation 2. Such convex formulation ensures the minimizer to find the global minimum of the energy functional. Finally, to make their algorithm even more robust, [2] introduce an anisotropic (*i.e.* image-driven) regularization based on the robust Huber norm [7].

Another extension of their approach consists in integration in the flow estimation not only the current image and the following, but also the previous image. The goal, in the original publication, is to cope with single degraded images within a video, for instance with historical video material.

Using optical flow in a MVD context. We present here a software based on optical flow estimation explained in the previous paragraph, designed to directly convert multi-view videos to multi-view videos plus depth, called *mv2mvd*. Contrary to the DERS, it computes the disparity and / or depth maps for all views in one single pass, instead of one view after the other. It's core uses the CUDA-based library[1] developed in parallel with [2]. The core framework of *mv2mvd* is depicted on Figure 3. On one hand, disparities are computed from left to right views (Figure 3, second row). On the other hand, they are estimated from right to left (Figure 3, third row). The interest of such two-side computation is to be able to locate occlusion zones where the motion field would be incorrect (a pixel in one view would not be visible in the other view). In fact, cross check is performed [1] to detect outlier pixels in each computed disparity map, which are finally combined (Figure 3, fourth row) by taking the minimal value of each disparity pair, to avoid the foreground fattening effect exhibited by window-based algorithms [1].

[1] See `http://www.gpu4vision.org`

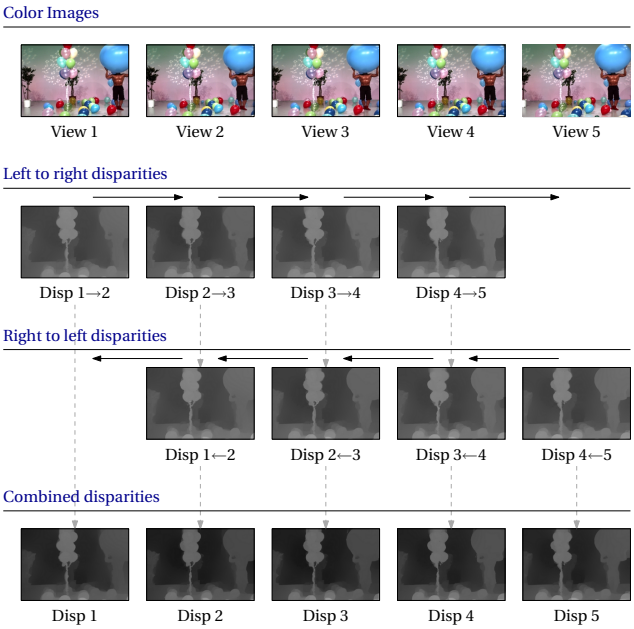


Figure 3: Global framework of disparity estimation with *mv2mvd* (sequence *Balloons*)

4. RESULTS

We present in Figure 4 disparity maps extracted with our method for the sequences *Newspaper* and *Book Arrival*, together with original images and disparity maps computed with the DERS and provided to the MPEG community. Each line in the figure is related to one of the original views. The original images are in the first column, disparity maps computed with *mv2mvd* are in the second one, while DERS-based maps are in the third column.
We can notice that globally, estimated disparities seem perceptually more relevant with regards to the scene with our approach. For instance, for the sequence *Newspaper*, the background of the scene is correct. With the DERS, numerous zones with different depths appear, while the depth of the background is physically the same with regards to acquisition cameras. As for the sequence *Book Arrival*, we can notice a greater spatial stability in the estimated depths, which appear more noisy in the DERS case. This comes for the latter from the application of plane fitting on mean-shift segments, which break the local spatial depth consistency applied to each segment.

On Figure 5, we show visual differences between our optical flow-based disparities, and disparities deduced from a depth camera (z-cam) acquisition of the scene. These results are presented for the central of the five input views of the *Cafe* sequence. How such z-cam acquisition has been performed is described in [8]. Keeping in mind that these z-cam-based disparities are not raw and have been interpolated to fit the full video resolution, it is worth notice that our method competes very well with the depth sensors. For instance, depth contours seem sharper with our method (all sub-images). We are even able to retrieve small depth details with much less noise (bottom right sub-image, for the chair part). However, for uniform zones with depths gradients, disparities are better estimated with the z-cam (see the furthest table for instance), where our method discretizes too heavily the resulting signal, while however better retrieving depth contours. Notice that

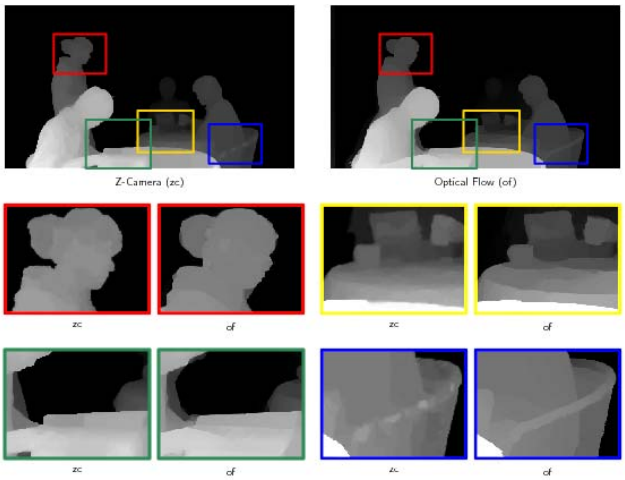


Figure 5: Comparison between Z-Camera- and Optical Flow-based disparities

image gamma has been modified on sub-images for visualization purposes.

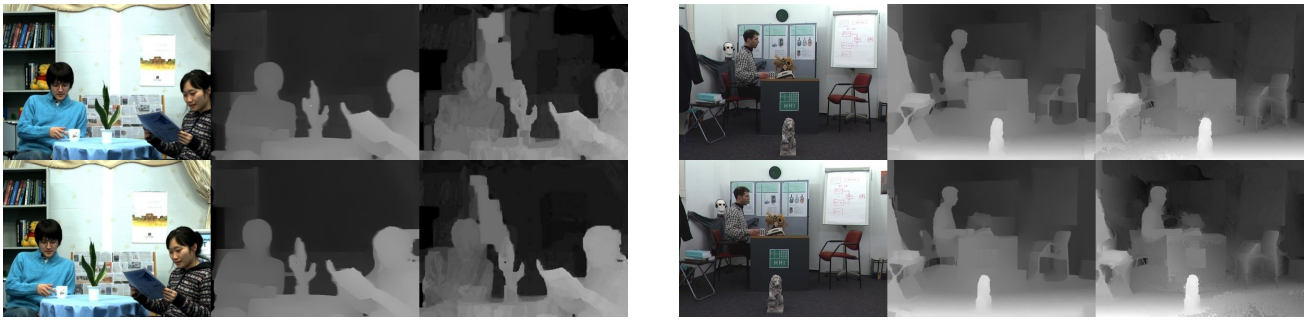
On Figure 6, we present evaluation results of our disparity maps in terms of virtual views synthesis quality. The evaluation protocol used is the one used by the MPEG community.
Disparity maps are computed for views $N-1$ and $N+1$. These maps are used as input to the VSRS in order to synthesize the view N. This virtual view is then compared to the original view N in terms of spatial PSPNR [9] and temporal PSPNR, with the *Pspnr* tool provided to the MPEG community. We present for each sequence and each of these two measures three different plots (quality measure against video frame number). The red curve is associated to disparity maps generated by the DERS. The blue and black curves are respectively associated to our method without (MV2MVD F2) or with (MV2MVD F3) the integration of the symmetry constraint (see Section 3). The dashed horizontal lines in the figures correspond to the mean values over the whole considered sequence.
We notice that from now on, it is difficult to bring a clear conclusion to this evaluation procedure. Indeed, the quality of synthesized views seem to depend mainly on the input data. An ordering of the different methods per sequence is provided in Table 1. It appears however that our method seem to better behave in most cases than the DERS. We must also notice that compared to the DERS, there is absolutely no temporal consistency enforced in our algorithm, since it seems to provide stable enough results from one instant to the other, which is not the case of the reference software.

Sequence	**1st**	**2nd**	**3rd**
Newspaper	mv2mvd F3	DERS	mv2mvd F2
Book Arrival	mv2mvd F2	mv2mvd F3	DERS
Lovebird 1	mv2mvd F3	mv2mvd F2	DERS

Table 1: Ordering of methods by synthesized views quality

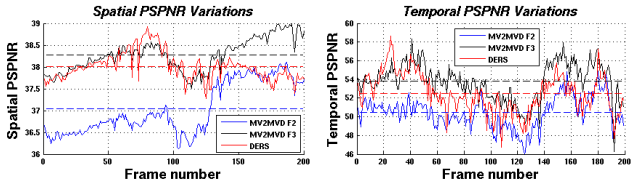
5. CONCLUSION

This paper presents a framework for depth maps computation for multi-view videos, based on a high-quality optical flow estimation algorithm. The generated maps have been evaluated in terms of synthesized views quality using the VSRS reference software.



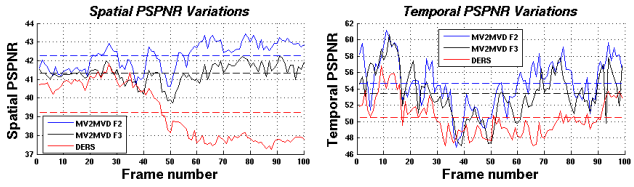
(a) Sequence *Newspaper* (b) Sequence *Book Arrival*

Figure 4: Comparison of extracted disparity maps between DERS (right column) and our method (middle column)



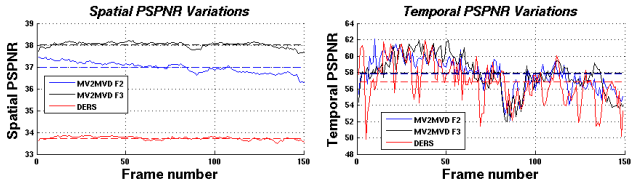
(a) *Newspaper* . . . . . Spatial PSPNR

(b) *Newspaper* . . . Temporal PSPNR



(c) *Book Arrival* . . . Spatial PSPNR

(d) *Book Arrival* . Temporal PSPNR



(e) *Lovebird 1* . . . . . Spatial PSPNR

(f) *Lovebird 1* . . . Temporal PSPNR

Figure 6: Virtual view evaluation for *Newspaper*, *Book Arrival* and *Lovebird 1*

It appears that our method gives promising results compared to maps computed by the associated reference software for depths extraction (the DERS). However, these results are subject to many interpretations and both methods are hardly comparable for the following reasons:

- No temporal consistency enforcement is integrated in out method, contrary to the DERS. This is due to the fact that such temporal consistency in the DERS can be interpreted as based on the assumption that the cameras are fixed, which we believe is way too restrictive.
- We do not propose to integrate depth masks during the estimation to provide manual data enhancing the final results, contrary to the DERS.
- At writing time, we were not able to constrain the optical flow estimation to be performed along image lines, as it should be with correctly rectified input stereo or multi-view images. We only select the horizontal component of the computed motion flow.
- Our method is only based on the luminance component of the input images, not on the RGB space, contrary, again, to the DERS.

Despite all these limitations with regards to the DERS, our method is able to compute depth maps totally relevant in terms of virtual views synthesis. Moreover, being implemented on the GPU, it is far faster that the DERS. The computational time can be reduced from 15 times to 150 times depending on the method used to compute the disparities with the reference software. And lastly, on an ease of use vision, our method computes maps for all views when a DERS execution is only valid for a single view, and has to be run independently for all of them.

6. REFERENCES

[1] Daniel Scharstein and Richard Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 7–42, April 2002.

[2] Manuel Werlberger, Werner Trobin, Thomas Pock, Andreas Wedel, Daniel Cremers, and Horst Bischof, "Anisotropic huber-L1 optical flow," in *Proceedings of the British Machine Vision Conference (BMVC'09)*, London, UK, 2009.

[3] Werner Trobin, *Local, semi-global and global optimization for motion estimation*, Ph.D. thesis, Graz University of Technology, December 2009.

[4] Berthold K.P. Horn and Brian G. Schunk, "Determing optical flow," *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.

[5] Nils Papenberg, Andrés Bruhn, Thomas Brox, Stephan Didas, and Joachim Weickert, "Highly accurate optic flow computation with theoretically justified warping," *International Journal of Computer Vision (IJCV)*, vol. 67, no. 2, pp. 141–158, April 2006.

[6] Christopher Zach, Thomas Pock, and Horst Bischof, "A duality based approach for realtime TV-L1 optical flow," in *Proceedings of the DAGM-Symposium*, 2007, pp. 214–223.

[7] Peter J. Huber, *Robust Statistics*, John Wiley and sons, 1981.

[8] Eun-Kyung Lee, Yun-Suk Kang, Jae-Il Jung, and Yo-Shun Ho, "3-d video generation using multi-depth camera system," ISO-MPEG Input document m17225, Gwangju Institute of Science and Technology (GIST), Kyoto, Japan, 2010.

[9] E. Gershikov, E. Lavi Burlak, and M. Porat, "Correlation-based approach to color image compression," *SP:IC*, vol. 22, no. 9, pp. 719–733, Oct. 2007.

Acknowledgments. The authors would like to thank the authors of the presented optical flow method for providing their library to the scientific community.