TEMPORALLY COHERENT STEREO MATCHING USING KINEMATIC CONSTRAINTS

Rodrigo Schramm and Claudio R. Jung

Institute of Informatics, Federal University of Rio Grande do Sul, Brazil
Postal Box 15064, Porto Alegre, RS, Brazil, 91501-970
rodrigos@caef.ufrgs.br, crjung@inf.ufrgs.br

ABSTRACT

This paper explores a simple yet effective way to generate temporally coherent disparity maps from binocular video sequences based on kinematic constraints. Given the disparity map at a certain frame, the proposed approach computes the set of possible disparity values for each pixel in the subsequent frame, assuming a maximum displacement constraint (in world coordinates) allowed for each object. These disparity sets are then used to guide the stereo matching procedure in the subsequent frame, generating a temporally coherent disparity map. Experimental results indicate that the proposed approach produces temporally coherent disparity maps comparable to or better than competitive methods.

Index Terms— stereo matching, disparity search range, temporal coherence, view interpolation

1. INTRODUCTION

Estimating the depth of a 3D scene is an active topic in image processing, with several applications such as object tracking, scene recognition and view interpolation. There are several methods for stereo matching, which can be subdivided into local or global approaches [1]: local methods are typically based on an initial cost function for each pixel and disparity, followed by an aggregation step within a neighborhood of each pixel, and the Winner-Takes-All (WTA) procedure is applied to select the disparity with the lowest cost; global methods, on the other hand, usually solve a global optimization problem that takes into account both evidence from the stereo pair (e.g. a cost function) and a smoothness constraint for the disparity map.

When video stereo sequences are used, a frame-by-frame analysis is usually not appropriate, since temporal artifacts in the disparity maps may be generated. Alternatively, temporal consistency can be used to improve the estimate of the disparity map. The work of Leung and colleagues [2] enforces temporal consistency by minimizing the difference between the disparity maps of adjacent frames, which may lead to problems when large motion is present. Davis et al. [3] presented an approach that extended the traditional neighborhood-based stereo matching to include multiple frames across time. However, their approach was focused mostly on active methods (e.g. high-frequency structured light), since the temporal window should not contain strong motion. Richardt and colleagues [4] explored temporal dual-cross-bilateral (TDCB) grids for cost aggregation, along with a fast GPU implementation.

Khoshabeh and collaborators [5] presented a two-stage algorithm for disparity estimation in video sequences. Initially, a frame-by-frame approach is adopted, and then a 3D optimization procedure including temporal information is applied. Their method indeed improves temporal coherence, but the quality of the generated disparity video sequence is highly dependent on the choice of the spatial and temporal penalties used in the optimization stage. Pham [6] presented an approach for spatio-temporal stereo matching based on the information permeability method. The main idea of their approach is to first aggregate the disparity costs in space and then in time (using a few adjacent frames), using color similarity as weights in the aggregation step. Despite the good results shown by the authors, the spatial windows related to moving pixels may not present much overlap in time, compromising the temporal coherence.

There are also approaches that estimate the disparity flow, which contains the spatial displacement of each pixel as well as the corresponding disparity variation, and thus encode information about 3D motion. Gong [7] estimates the disparity flow in an interactive manner, by initially predicting the displacement at a future frame with the current disparity flow, and then computing its disparity map using the predicted value. The computational cost of this approach is high, but the author explored GPU implementations to improve the performance. Wedel and colleagues [8] proposed a variational framework for the estimation of stereoscopic scene flow. They take into account image pairs from two consecutive frames and compute both depth and a 3D motion vector, but decoupling the depth estimation from the motion estimation. Hung et al. [9] also presented a variational approach to jointly estimate depth and flow in stereo binocular sequences, dealing in particular with depth/motion outliers. As in traditional stereo matching for still images, variational methods (such as [8, 9]) tend to present very good result, at the cost of high computational complexity. For the sake of

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

illustration, an average of 1 minute is needed to process a single 640×480 frame using 24 cores of a 2.67 GHz Intel Xeon processor, as reported in [9].

In this work we present a new framework to include temporal coherence into stereo matching algorithms by exploring the expected 3D motion of the scene. The proposed approach explores the projection of 3D objects with a maximum (known) displacement in world coordinates to define a candidate region for each pixel in the image, and also a list of possible disparity values for each pixel. We also present an approach to integrate these sets of temporally coherent disparity values with existing stereo matching algorithms [10, 11] and generate interpolated video sequences. Our experimental results indicate that the inclusion of temporal coherence significantly reduce the number of temporal artifacts in the disparity video sequence. We also applied the proposed approach to generate new synthetic video sequences based on view interpolation algorithms [12], obtaining a PSNR gain around 1dB for each frame in comparison to frame-by-frame analysis.

The remainder of this work is organized as follows. Section 2 presents details of the proposed algorithm, and experimental results are shown in Section 3. Finally, Section 4 draws the conclusions.

2. TEMPORALLY COHERENT DISPARITY MAP

When shooting a stereo movie continuously, we should expect coherence in time, space and disparity. In particular, there are several scenarios for which we can assume a maximum displacement (in world coordinates) in adjacent frames for each object in the scene. For instance, if a static stereo camera is used, the characteristics of the moving objects can be used to define a maximum displacement Δ_{max} between two adjacent frames. If pedestrians are the moving objects, Δ_{max} should be a small value, whereas a larger value for Δ_{max} should be defined if shooting a sports car race.

Given a stereo camera pair C_1, C_2 with the same focal length f and baseline b, and given that each 3D point viewed by the camera presents a maximum displacement in Δ_{max} world coordinates across adjacent frames, the goal of this paper is to define a set of geometrical relationships that relate time, space (image plane) and disparity. We also indicate how these constraints can be explored by existing stereo matching approaches to generate a temporally coherent disparity map, focusing on the problem of view interpolation.

In the pinhole camera model, a point $\boldsymbol{X} = (x, y, z)$ in the 3D scene is projected to the image plane according to

$$u = f\frac{x}{z}, \;\; v = f\frac{y}{z}, \tag{1}$$

where $\boldsymbol{p} = (u, v)$ is the corresponding pixel (in image coordinates) and f is the focal length. Also, if the horizontal disparity of a 3D point viewed by a pair of rectified cameras with baseline b is d, then the object depth $z = f\frac{b}{d}$.

Let $\boldsymbol{\delta X} = (\delta x, \delta y, \delta z)$ be the displacement vector of a 3D point between two frames, with $\|\boldsymbol{\delta X}\| \leq \Delta_{\text{max}}$ according to our maximum displacement hypothesis. In this way, a point $\boldsymbol{X}$ at frame t will be located at position $\boldsymbol{X} + \boldsymbol{\delta X}$ at frame $t+1$, and it will be projected to an image point $\boldsymbol{p} + \boldsymbol{\delta p}$, with $\boldsymbol{\delta p} = (\delta u, \delta v)$, and

$$\delta u = \frac{f\delta x - u\delta z}{z + \delta z}, \; \delta v = \frac{f\delta y - v\delta z}{z + \delta z} \tag{2}$$

are computed using Equation (1).

The maximum displacement on the image plane will happen when the corresponding 3D point presents maximum displacement parallel to the image plane, i.e., when $\delta x^2 + \delta y^2 = \Delta^2_{\text{max}}$. Hence, it is possible to compute a bound $R_{\text{max}}(\boldsymbol{p})$ for the maximum displacement of $\boldsymbol{p}$ (in image coordinates) through

$$R_{\text{max}}(\boldsymbol{p}) = \max \|\boldsymbol{\delta p}\| = \frac{f\Delta_{\text{max}}}{z}, \tag{3}$$

which is obtained by computing $\|\boldsymbol{\delta p}\|$ using Equation (2) with $\delta z = 0$ and $\delta x^2 + \delta y^2 = \Delta^2_{\text{max}}$.

Given a pixel $\boldsymbol{p}$ at frame t and the corresponding point $\boldsymbol{p}' = \boldsymbol{p} + \boldsymbol{\delta p}$ at frame $t+1$, the maximum image displacement $R_{\text{max}}(\boldsymbol{p})$ can be used to define a set of plausible displacement vectors $S(\boldsymbol{p}) = \{\boldsymbol{\delta p} \mid \|\boldsymbol{\delta p}\| \leq R_{\text{max}}(\boldsymbol{p})\}$[1] to relate $\boldsymbol{p}$ with $\boldsymbol{p}'$.

Also, each $\boldsymbol{\delta p} \in S(\boldsymbol{p})$ relates to a set of possible 3D displacement vectors $\boldsymbol{\delta X}$ such that $\boldsymbol{X} + \boldsymbol{\delta X}$ projects to $\boldsymbol{p} + \boldsymbol{\delta p}$. In particular, there is a range $\delta z_{\text{min}} \leq \delta z \leq \delta z_{\text{max}}$ of plausible depth values that relate to a corresponding range $d_{\text{min}} \leq d \leq d_{\text{max}}$ of plausible disparity values. The extreme depth variations δz_{min} and δz_{max} occur when the corresponding 3D points present the maximum allowed displacement. Hence, given $\boldsymbol{\delta p} \in S(\boldsymbol{p})$ and solving Equation (2) for δz with the constraint $\|\boldsymbol{\delta X}\| = \Delta_{\text{max}}$ will generate solutions for δz_{min} and δz_{max}. Computing these solutions explicitly and projecting the 3D points onto the image plane using Equation (1) provides the disparity range $[d_{\text{min}}, d_{\text{max}}]$:

$$d_{\text{min}} = \frac{bf}{(z + \frac{(\sqrt{g_1+g_2+g_3+g_4+g_5}-H)}{J})} \tag{4}$$

$$d_{\text{max}} = \frac{bf}{(z - \frac{(\sqrt{g_1+g_2+g_3+g_4+g_5}+H)}{J})}, \tag{5}$$

where

$$G = -\sqrt{g_1+g_2+g_3+g_4+g_5}, \; Q = (f\Delta_{\text{max}})^2.$$
$$g_1 = ((-u^2-f^2)z^2+Q)\delta v^2, \; g_2 = (2uvz^2\delta u + 2Qv)\delta v$$
$$g_3 = ((-v^2-f^2)z^2+Q)\delta u^2, \; g_4 = 2Qu\delta u + Qv^2 + Qu^2$$
$$g_5 = Q^2, \; H = z\delta v^2 + vz\delta v + z\delta u^2 + uz\delta u$$
$$J = \delta v^2 + 2v\delta v + \delta u^2 + 2u\delta u + v^2 + u^2 + f^2.$$

[1] In fact, the search region is elliptical, since the 3D sphere that represents the set of possible 3D displacements projects to an ellipse on the image plane. However, for the sake of simplicity, we use the superscribing circle.

In summary, given an image point $\boldsymbol{p} = (u, v)$ and the corresponding disparity d at frame t, it is possible to find the depth z of the corresponding 3D point through $z = f\frac{b}{d}$. Assuming a maximum 3D displacement Δ_{max}, a 2D search region $S(\boldsymbol{p})$ can be defined on the image plane, containing the set of possible displacement vectors $\boldsymbol{\delta p} = (\delta u, \delta v)$ for point $\boldsymbol{p}$ between frames t and $t+1$. Furthermore, for each $\boldsymbol{\delta p} \in S(\boldsymbol{p})$, a range of plausible disparity values $R(\boldsymbol{p}, \boldsymbol{\delta p}) = [d_{\text{min}}, d_{\text{max}}]$ can be created, where d_{min} and d_{max} are given in Equations (4) and (5). If we scan each pixel $\boldsymbol{p}$ at frame t, we can build a set of plausible disparity values $\Omega(\boldsymbol{p})$ expected at frame $t+1$. Algorithm 1 describes the steps to compute the disparity search range.

Algorithm 1 Disparity Search Range Algorithm

```
initialize Ω(p) = ∅
for all p do compute R_max(p) and S(p)
    for all δp ∈ S(p) do
        compute R(p, δp)
        Ω(p + δp) = Ω(p + δp) ⋃ R(p, δp)
    end for
end for
```

It is important to note that adjacent pixels with distinct disparities can generate sparse intervals on the disparity search range. In summary, Ω encodes the sparse disparity search range of each (u, v) at frame $t+1$.

The dynamic disparity range defined in the previous section can be explored in several ways by existing stereo matching techniques. In global methods, for instance, the evidence (cost) related to valid disparities can have an increased weight in the global cost function, to prioritize temporally coherent disparities. In local methods, the dynamic disparity range can be used in the initial cost computation (to guide the disparity search space), in the aggregation step (to prioritize the aggregation of neighbors with valid disparities), or in the WTA step (also prioritizing disparities that lie within temporally coherent values).

In this paper, we present a possible solution for temporally coherent stereo matching based on an existing local method for computing the disparity map of still stereo pairs [13, 10]. The key point of [10] is the use of a local adaptive aggregation window that can be implemented in a computationally efficient manner using integral images. In [10], the first step is to compute a cost matrix $C(\boldsymbol{p}, d)$, where $\boldsymbol{p} = (u, v)$ represents the position of the pixel in the reference image (without loss of generality, we assume that the left image I_l is the reference) and d represents a horizontal disparity hypothesis between the rectified stereo image pair, given by

$$C(\boldsymbol{p}, d) = \sum_{(m,n) \in \mathcal{N}_c(\boldsymbol{p})} \rho(|I_l(u, v) - I_r(m - d, n)|), \quad (6)$$

where $\mathcal{N}_c(\boldsymbol{p})$ is a neighborhood around $\boldsymbol{p}$ where the costs are computed, and $\rho(\cdot)$ is a robust matching function given by $\rho(x) = 1 - e^{-x/\lambda}$, so that large errors x are bounded by the exponential, and λ is a parameter (set to 15 experimentally).

In the second stage, the values of $C(\boldsymbol{p}, d)$ are aggregated locally by simply computing the sum of costs within an adaptive window, defined based on the color similarity of the central pixel $\boldsymbol{p}$ and its neighbors. This neighborhood, called $\mathcal{N}_g(\boldsymbol{p})$, presents a cross-like structure that allows a fast implementation based on integral images (for more details please refer to [13, 10]). In this paper, we also explore the same neighborhood $\mathcal{N}_g(\boldsymbol{p})$, but also include the dynamic disparity range in the aggregation step. More precisely, aggregation is performed through

$$E(\boldsymbol{p}, d) = \sum_{q \in \mathcal{N}_g(\boldsymbol{p})} \frac{1}{N_p} \omega(\boldsymbol{q}, d) C(\boldsymbol{q}, d), \quad (7)$$

where N_p is the number of pixels in $\mathcal{N}_g(\boldsymbol{p})$, and $\omega(\boldsymbol{p}, d)$ is a weighting function that prioritizes neighbors of $\boldsymbol{p}$ for which d is a plausible disparity. Regions related to static objects should have temporal coherence enforced strongly, opposed to regions related to occlusions/disocclusions (where temporal discontinuities in the disparity map may arise). In this work, we use the pixel-wise color (or intensity) difference in adjacent frame to estimate how static a pixel is, and propose

$$\omega(\boldsymbol{q}, d) = \begin{cases} 1 & \text{if } d \in \Omega(\boldsymbol{q}) \\ 1 + e^{-\gamma(|I_l^t(\boldsymbol{q}) - I_l^{t-1}(\boldsymbol{q})|)} & \text{otherwise} \end{cases}, \quad (8)$$

where $|\cdot|$ denotes the L^2 difference of RGB values when color images are used, or the absolute value if intensity images are used (please recall that $\Omega(\boldsymbol{q})$ is the set of plausible disparity values for pixel q). The γ parameter, with $\gamma \in [0, 1]$, controls the penalty assigned to neighbors that are not coherent with the range of plausible disparities: a small value leads to the same of standard aggregation result, whereas a larger value tends to increase the aggregation cost function when several neighbors are not within the range of plausible disparities. Our experiments indicated that $\gamma = 0.1$ is a good choice. Finally, the disparity for each pixel $\boldsymbol{p}$ is based on the WTA approach, i.e. the disparity value that minimizes $E(\boldsymbol{p}, d)$.

3. EXPERIMENTAL RESULTS

We have tested our method using several natural and synthetic stereo/multiview sequences, using $\Delta_{\text{max}} = 0.5$m in all examples. In the first set of experiments, five synthetic video sequences with ground truth data proposed in [4][2] were used to evaluate the quality of the estimated disparity maps in a quantitative manner, based on the average (in time) percent of bad pixels [1] with a threshold of 1 pixel. Results of the frame-by-frame approach (CROSS) inspired by [13, 10], our temporal coherence (CROSS-TC), HBP-TV [5] and TDCB [4] for the

[2]Available at `http://www.cl.cam.ac.uk/research/rainbow/projects/dcbgrid/datasets/`

five sequences corrupted with Gaussian noise $\mathcal{N}(0, 20)$ are summarized in Table 1. As can be observed, our approach presented better results than TDCB in all cases, and in three of the sequences it was also better than HBP-TV.

	Video Sequence				
Technique	Book	Street	Tanks	Temple	Tunnel
CROSS	35.85	19.16	30.93	19.83	29.42
CROSS-TC	29.70	18.68	**24.73**	**15.20**	**25.57**
HBP-TV	**26.97**	**17.69**	26.50	18.01	29.50
TDCB	38.95	24.17	29.34	29.89	33.01

Table 1. Quantitative evaluation of disparity maps (average percent of bad pixels), best results shown in bold. Following [5], Gaussian noise $\mathcal{N}(0, 20)$ was added to all frames.

We have also used two publicly available multiview datasets, *Book Arrival* and *Door Flowers* [14], and our own multiview sequence called *Herodion*, to test the proposed approach. Since ground truth disparity maps are not available for these sequences, and the average percent of bad pixels may have different impact on the final 3D reconstruction [15], we have performed another quantitative evaluation in an indirect manner by exploring an important application of temporal disparity maps: video view interpolation. For that purpose, we use sequences acquired with three or more cameras, select a subset of three adjacent cameras, and compute the disparity maps (LR and RL) with the external cameras. Then, we use standardized view interpolation algorithm (VSRS 3.5 [12]) to produce a synthetic video sequence with in-between views, and compare it with the feed from the actual camera using the PSNR metric.

Fig. 1 shows the PSNR values along all frames of Flowers video sequence without (CROSS) and with (CROSS-TC) temporal coherence. It is possible to notice a significant improvement (over 1 dB in average) when including temporal coherence. Table 2 shows the temporal average, median and standard deviation of the PSNR values computed for each video sequence, without and with temporal coherence. Some examples of estimated disparity maps and synthesized views for the *Door Flowers* sequence are shown in Fig. 2. As can be observed, the temporally coherent disparity map is at the same time smooth and able to retain fine details. Also, the synthesized views using our method present less artifacts (please see the right side of the desk and the posters on the wall). The full video sequences (reference, disparity maps and interpolated views) can be found at `http://inf.ufrgs.br/~rschramm/projects/stereo/temporal/videos/`.

Video Dataset	**PSNR** - without temporal coherence		
	Average	Median	St. Dev.
Book Arrival	33.19	33.32	0.76
Door Flowers	35.73	35.73	0.65
Herodion	21.68	21.34	1.27
Video Dataset	**PSNR** - with temporal coherence		
	Average	Median	St. Dev.
Book Arrival	34.42	34.37	0.92
Door Flowers	36.86	36.85	0.76
Herodion	22.64	22.61	1.68

Table 2. Quantitative evaluation of interpolated video sequences without (top) and with (bottom) temporal coherence.

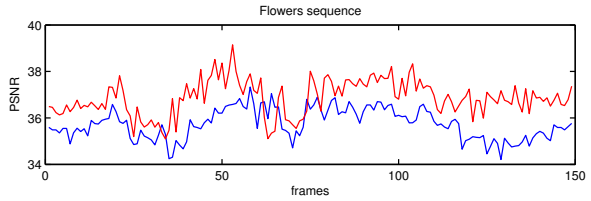

Fig. 1. PSNR measures from *Door Flowers* sequence (red with and blue without temporal coherence).

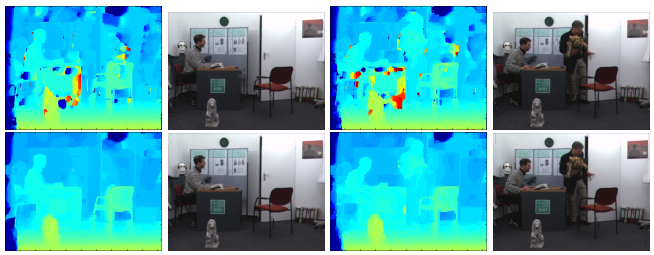

Fig. 2. Disparity maps and interpolated views for the *Door Flowers* sequence related to frames 28 and 71 without (top) and with (bottom) temporal coherence.

4. CONCLUSIONS

In this paper we proposed a geometric approach to generate temporally coherent disparity maps based on rectified stereo video sequences. Our method produces a set of plausible disparity values for each pixel of a given frame based on the disparity values of the previous frame, and can be incorporated into different stereo matching frameworks. We have also presented an extension of an existing algorithm [13] by including temporal coherence.

Our experimental results indicate that the inclusion of temporal coherence indeed improves the quality of the estimated disparity maps, being also as good as or better than competitive approaches [4, 5]. We have also explored the disparity maps to generate new synthetic views in the context of view interpolation, and showed that the new views obtained with temporally coherent disparity maps present higher PSNR values than the underlying frame-by-frame method. As future work, we plan to integrate the proposed algorithm with other state-of-the-art stereo matching techniques focused on still stereo pairs. We also intend to implement our method in GPU, following a trend of stero matching methods [4, 10].

5. REFERENCES

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, pp. 7–42, April 2002.

[2] Carlos Leung and Brian C. Lovell, "An energy minimisation approach to stereo-temporal dense reconstruction," in *In International Conference on Pattern Recognition*, 2004, pp. 72–75.

[3] James Davis, Diego Nehab, Ravi Ramamoorthi, and Szymon Rusinkiewicz, "Spacetime stereo: A unifying framework for depth from triangulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 296–302, 2005.

[4] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N. A. Dodgson, "Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid," in *European conference on computer vision conference on Computer vision*, 2010, pp. 510–523.

[5] Ramsin Khoshabeh, Stanley H. Chan, and Truong Q. Nguyen, "Spatio-temporal consistency in video disparity estimation.," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011*. May 2011, pp. 885–888, IEEE.

[6] C. C. Pham, V. D. Nguyen, and J. W. Jeon, "Efficient spatio-temporal local stereo matching using information permeability filtering," in *Proceedings of the 16th IEEE International Conference on Image Processing*, 2012, pp. 2965–2968.

[7] Minglun Gong, "Real-time joint disparity and disparity flow estimation on programmable graphics hardware," *Computer Vision and Image Understanding*, vol. 113, no. 1, pp. 90 – 100, 2009.

[8] A. Wedel, T. Brox, T. Vaudrey, C. Rabe, U. Franke, and D. Cremers, "Stereoscopic scene flow computation for 3d motion understanding," *International Journal of Computer Vision*, vol. 95, no. 1, pp. 29–51, 2011.

[9] Chun Ho Hung, Li Xu, and Jiaya Jia, "Consistent binocular depth and scene flow with chained temporal profiles," *International Journal of Computer Vision*, vol. 102, pp. 271–292, 2013.

[10] Xing Mei, Xun Sun, Mingcai Zhou, Shaohui Jiao, Haitao Wang, and Xiaopeng Zhang, "On building an accurate stereo matching system on graphics hardware," in *2011 IEEE International Conference on Computer Vision (ICCV Workshops)*, nov. 2011, pp. 467 –474.

[11] GwnagYul Song, SeongIk Cho, DongYong Kwak, and JoonWoong Lee, "Accurate dense stereo matching of slanted surfaces using 2d integral images," in *Computer Vision Systems*, vol. 7963 of *Lecture Notes in Computer Science*, pp. 284–293. Springer Berlin Heidelberg, 2013.

[12] Y. Zhao, L. Yu, and D. Tian, "Improved 1D mode in VSRS 3.1," 2009, MPEG document M16582.

[13] Ke Zhang, Jiangbo Lu, and G. Lafruit, "Cross-based local stereo matching using orthogonal integral images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 7, pp. 1073 –1079, july 2009.

[14] FhG-HHI, "Mobile 3DTV content delivery optimization over DVB-H system. Mobile 3DTV, Solideyesight.," 2011.

[15] Ivan Cabezas, Victor Padilla, Maria Trujillo, and Margaret Florian, "On the impact of the error measure selection in evaluating disparity maps," in *World Automation Congress (WAC), 2012*, June 2012, pp. 1–6.