Comments on ”Consistent Depth Maps Recovery from a Video Sequence”

N.P. van der Aa

D.S. Grootendorst

B.F. Böggemann

R.T. Tan

Technical Report UU-CS-2011-014
May 2011

Department of Information and Computing Sciences
Utrecht University, Utrecht, The Netherlands
www.cs.uu.nl

ISSN: 0924-3275

Department of Information and Computing Sciences
Utrecht University
P.O. Box 80.089
3508 TB Utrecht
The Netherlands

Comments on "Consistent Depth Maps Recovery from a Video Sequence"

Nico P. van der Aa, Dennis S. Grootendorst, Bob F. Böggemann, and Robby T. Tan

Abstract—The paper by Zhang *et al.* [1] proposes a novel method to automatically construct a view-dependent depth map for each frame of an input video sequence taken by a freely moving camera. The paper introduces the concept of bundle optimization to make the corresponding depth values in multiple frames consistent and to assign distinctive depth values for pixels that fall in different depth layers. To determine the data term of the energy function used in this optimization process, epipolar geometry is used in the form of Eq. (6) of their paper. This crucial relation directly uses the disparity value to describe the epipolar line. However, the variables in this formula are not explained in detail, the derivation is left out and it needs some adjustments to use it in the general wide-baseline case. To support other readers to implement the ideas of this paper, we present the proper version of this relation, with notations which are commenly used in the community, and its geometrical derivation.

Index Terms—Consistent depth maps recovery, multiview stereo, epipolar geometry.

1 Introduction

Epipolar geometry [2] is commonly used in depth estimation and 3D reconstruction techniques. The idea that a point in one camera view corresponds to an epipolar line in the other camera view is implemented by Zhang *et al.* [1] such that this epipolar line is given in terms of the camera parameters of both cameras and a disparity value, which is inversely related to the depth. Given two camera views t and t', Eq. (6) of [1] states

$$\mathbf{x}'^h \sim \mathbf{K}_{t'}\mathbf{R}_{t'}^\top \mathbf{R}_t \mathbf{K}_t^{-1}\mathbf{x}^h + d_{\mathbf{x}}\mathbf{K}_{t'}\mathbf{R}_{t'}^\top (\mathbf{T}_t - \mathbf{T}_{t'}), \quad (1)$$

where $\{\mathbf{K}_t, \mathbf{R}_t, \mathbf{T}_t\}$ is the set of camera parameters for view t with $\mathbf{K}_t$ the calibration matrix describing the central projection with a correction for the principal point offset, $\mathbf{R}_t$ the rotation matrix representing the orientation of the camera coordinate frame, and $\mathbf{T}_t$ the so-called "translation vector". For view t' a similar set of camera parameters is defined. The homogeneous vectors $\mathbf{x}^h$ and $\mathbf{x}'^h$ are the projections of a world point $\mathbf{X}$ on the image planes t and t', respectively.

• *N.P. van der Aa, D.S. Grootendorst, B.F. Böggemann and R.T. Tan are with the Department of Information and Computing Sciences, Utrecht University, The Netherlands. E-mail: nico@cs.uu.nl*

Eq. (1) has several shortcomings. First, the parameters $\mathbf{T}_t$ and $\mathbf{T}_{t'}$ are declared as "translation vectors" instead of the world coordinates of the camera centers of view t and t', respectively. If $\mathbf{C}_t = \begin{bmatrix}\tilde{\mathbf{C}}_t^\top & 1\end{bmatrix}^\top$, where $\tilde{\mathbf{C}}_t$ and $\tilde{\mathbf{C}}_{t'}$ represent the coordinates of the camera centers in 3D Euclidean space, then $\mathbf{T}_t \equiv \tilde{\mathbf{C}}_t$ and $\mathbf{T}_{t'} \equiv \tilde{\mathbf{C}}_{t'}$. This is confusing since it is common practice to define the translation vector $\mathbf{T}_t$ as $\mathbf{T}_t = -\mathbf{R}\tilde{\mathbf{C}}_t$. Second, if the notation is consistent with [2], the transpose should not be taken from $\mathbf{R}_{t'}$, but from $\mathbf{R}_t$. In other words, the reverse angle definition should be used. Finally, Eq. (1) assumes that the disparity can take any value between 0 and ∞, but if one camera center is positioned behind another, e.g. when the camera takes another image after zooming in, there is an upper bound on the disparity value μ (or a lower bound on the observable depth value) that both cameras can register. The corrected version of Eq. (6) of [1] is given by

$$\mathbf{x}'^h \sim \mathbf{K}_{t'}\mathbf{R}_{t'}\mathbf{R}_t^\top \mathbf{K}_t^{-1}\mathbf{x}^h + d_{\mathbf{x}}\mathbf{K}_{t'}\mathbf{R}_{t'}(\tilde{\mathbf{C}}_t - \tilde{\mathbf{C}}_{t'}), \quad (2)$$

for $0 \leq d_x \leq \mu < \infty$. To motivate these corrections, we give a geometric derivation of Eq. (2) in Section 2 and some illustrative examples in Section 3.

2 GEOMETRIC DERIVATION

The projection of a world point $\mathbf{X}$ on an image plane t is defined by the projection matrix $\mathbf{P}_t$, such that

$$\mathbf{x}^h = \mathbf{P}_t\mathbf{X} = \mathbf{K}_t\mathbf{R}_t\begin{bmatrix}\mathbf{I} & -\tilde{\mathbf{C}}_t\end{bmatrix}\mathbf{X}. \quad (3)$$

Inversely, all points $\mathbf{X}$ on the line spanned by the camera center $\mathbf{C}_t$ and the point $\mathbf{x}^h$ would project to $\mathbf{x}^h$. The projection of this line in view t', the so-called epipolar line, contains all possible candidate positions of the image point $\mathbf{x}'^h$ that would correspond to the candidate world points $\mathbf{X}$ (see Figure 1). The idea of Zhang *et al.* is to parametrize this line dependent on the disparity. Therefore, we search for pairs of image points in both camera views corresponding to zero disparity and infinite disparity. Let us assume that point $\mathbf{X}_\infty \equiv \begin{bmatrix}\tilde{\mathbf{X}}_\infty^\top & 0\end{bmatrix}^\top$ is a world point at infinity (and thus has an infinite distance to the camera center $\mathbf{C}_t$ and zero disparity), that projects to the image point $\mathbf{x}^h$ in view t, so

$$\mathbf{x}^h = \mathbf{P}_t\mathbf{X}_\infty = \mathbf{K}_t\mathbf{R}_t\tilde{\mathbf{X}}_\infty. \quad (4)$$

Since this is an one-to-one relation and the inverse of $\mathbf{K}_t$ and $\mathbf{R}_t$ exist, $\mathbf{X}_\infty$ is uniquely related to $\mathbf{x}^h$. Note that the inverse of $\mathbf{R}_t$ is its transpose. Thus,

$$\tilde{\mathbf{X}}_\infty = \mathbf{R}_t^\top\mathbf{K}_t^{-1}\mathbf{x}^h. \quad (5)$$

By projecting $\mathbf{X}_\infty$ to the image plane t', we obtain the point on the epipolar line which represents infinite depth or zero disparity, namely

$$\mathbf{x}'^h = \mathbf{K}_{t'}\mathbf{R}_{t'}\begin{bmatrix}\tilde{\mathbf{X}}_\infty \\ 0\end{bmatrix} = \mathbf{K}_{t'}\mathbf{R}_{t'}\mathbf{R}_t^\top\mathbf{K}_t^{-1}\mathbf{x}^h. \quad (6)$$

The point with zero depth with respect to the camera center $\mathbf{C}_t$ is the camera center itself. Since the epipole is defined as the image of the camera center of one view in the other view, the projection of this point in view t' is the epipole $\mathbf{e}'$, viz.

$$\mathbf{e}' = \mathbf{P}_{t'}\mathbf{C}_t. \quad (7)$$

Let $\mathbf{C}_t \equiv (\tilde{\mathbf{C}}_t^\top, 1)^\top$, then Eq. (7) can be written as

$$\mathbf{e}_{t'} = \mathbf{K}_{t'}\mathbf{R}_{t'}\left(\tilde{\mathbf{C}}_t - \tilde{\mathbf{C}}_{t'}\right). \quad (8)$$

If the angle between the vector $\tilde{\mathbf{C}}_t - \tilde{\mathbf{C}}_{t'}$ and the principal axis of camera t' is smaller than 90 degrees, the situation of Figure 1 occurs and the epipolar point is defined on the image plane. But, if the angle is larger than 90 degrees, the camera center $\mathbf{C}_t$ lies behind camera center $\mathbf{C}_{t'}$ and cannot be "seen" by that camera and the situation of Figure 2 occurs. There is no intersection of this line segment with image plane t', so it leads to a virtual epipole. However, the vector computed in Eq. (8) still gives a direction for the epipolar line.

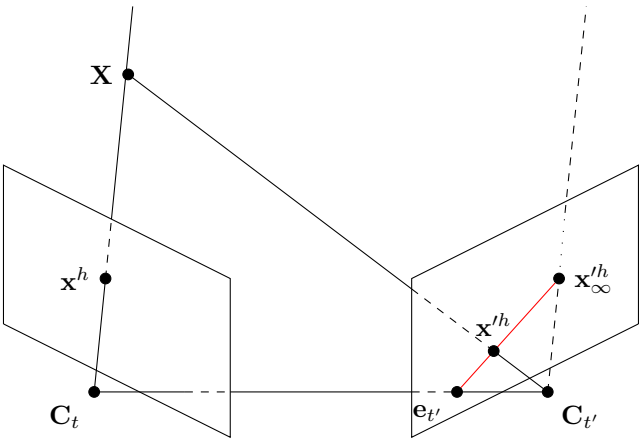

Fig. 1. Illustration of the parametric construction of the epipolar line (red line) in view t' by Zhang *et al.*

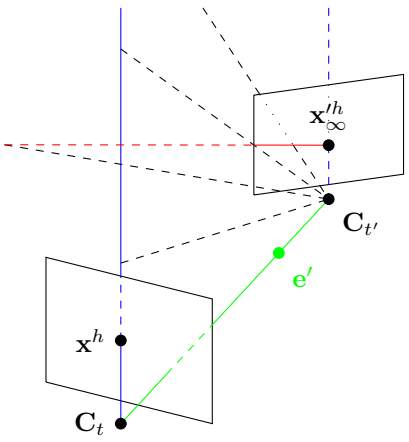

Fig. 2. Illustration of the construction of the epipolar line (red) when the epipole in view t' becomes virtual.

Thus, at least one point on and a direction of the epipolar line in frame t' are found. The line through this point with the direction found is given by

$$\mathbf{x}'^h(\lambda) \sim \mathbf{K}_{t'}\mathbf{R}_{t'}\mathbf{R}_t^\top\mathbf{K}_t^{-1}\mathbf{x}^h + \lambda\mathbf{K}_{t'}\mathbf{R}_{t'}\left(\tilde{\mathbf{C}}_t - \tilde{\mathbf{C}}_{t'}\right). \quad (9)$$

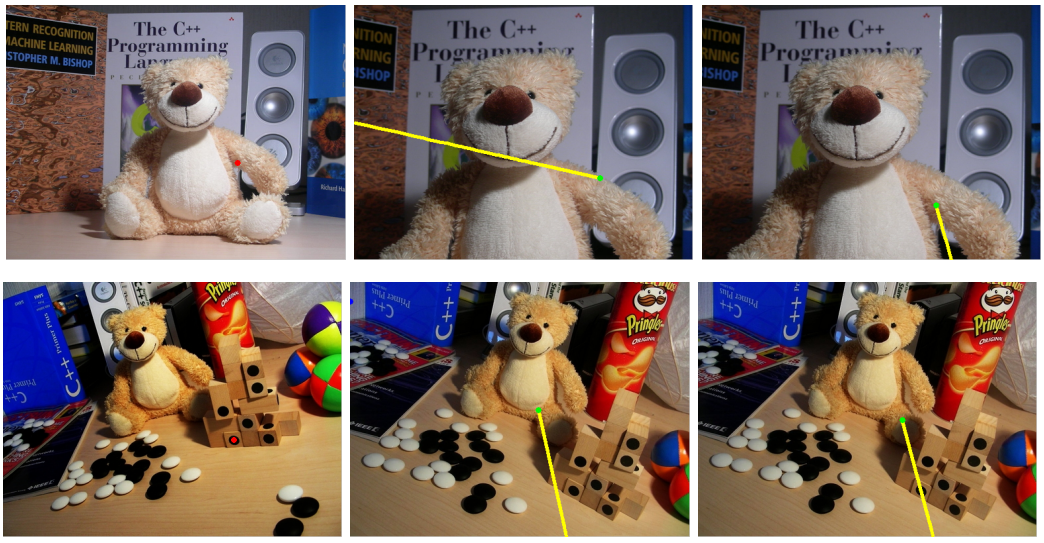

Fig. 3. Illustration of improvements on epipolar line construction on serie 1 (top row) and 2 (bottom row) of the data set. *Left*: Point selection illustrated by a red dot in camera view 1. *Middle*: The constructed epipolar line (yellow line) using the original version of Zhang's equation (1) using the conventions of [2], where the green dot is the image of the world point at infinity. *Right*: The constructed epipolar line using the corrected version of Zhang's equation (2).

The image point $\mathbf{x}'^h$ on the epipolar line is determined by dividing the result of Eq. (9) by its third coordinate. This image point only corresponds to a valid disparity value as long as this third coordinate is positive. This gives a way to compute the maximum disparity value μ. This concludes the geometrical derivation of the corrected version of Zhang's equation Eq. (2).

3 Illustrative Examples

To illustrate the effect of the corrections to Zhang's equation, we use Tola's data set [3], [4]. This data set consists of 6 image pairs of a static scene. In Figure 3 the result of constructing the epipolar line is shown for series 1 and 2. In both series, each camera is placed within a different distance to the object, causing one camera to be behind the other and the special case as illustrated in Figure 2 occurs. Straightforward computation of Zhang's equation would lead to an epipolar line that goes into the wrong direction. The maximum disparity value μ for the selected pixel for serie 1 is 71.6, while for serie 2 it is 32.9. Figure 3 shows the results of the epipolar line construction with the corrections applied to the original equation presented by Zhang *et al.*, leading to the correct epipolar line.

References

[1] G. Zhang, J. Jia, T.-T. Wong and H. Bao, "Consistent Depth Maps Recovery from a Video Sequence", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 31, no. 6, pp. 974-988, June 2009.

[2] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, second ed., Cambridge University Press, 2003.

[3] E. Tola, V. Lepetit, P. Fua, "A fast local descriptor for dense matching", CVPR'08, 2008.

[4] E. Tola, V. Lepetit, P. Fua, "DAISY: An efficient dense descriptor applied to wide baseline stereo", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 32, no. 5, May 2010, pp. 815-830.