Signal Processing: Image Communication 43 (2016) 42–53

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

A data-fusion approach to motion-stereo

Francesco Malapelle [a,*], Andrea Fusiello [a], Beatrice Rossi [b], Pasqualina Fragneto [b]

[a] *Dipartimento di Ingegneria Elettrica, Gestionale e Meccanica, University of Udine, Via Delle Scienze, 208 - 33100 Udine, Italy*
[b] *AST Lab - STMicroelectronics Via Camillo Olivetti, 2 - Agrate Brianza (MB), Italy*

ARTICLE INFO

Article history:
Received 4 February 2015
Received in revised form
31 January 2016
Accepted 29 February 2016
Available online 3 March 2016

Keywords:
Motion-stereo
Temporal-stereo
Dynamic-stereo
Data fusion
Kalman filter
Parallax

ABSTRACT

This paper introduces a novel method for performing motion-stereo, based on dynamic integration of depth (or its proxy) measures obtained by pairwise stereo matching of video frames. The focus is on the data fusion issue raised by the motion-stereo approach, which is solved within a Kalman filtering framework. Integration occurs along the temporal and spatial dimension, so that the final measure for a pixel results from the combination of measures of the same pixel in time and whose of its neighbors. The method has been validated on both synthetic and natural images, using the simplest stereo matching strategy and a range of different confidence measures, and has been compared to baseline and optimal strategies.

1. Introduction

This paper deals with the problem of *motion-stereo*, i.e., depth estimation in a monocular sequence of images taken by a moving camera [31]. Whereas in binocular stereo two cameras separated by a fixed baseline are employed, in motion-stereo a single camera moves through a static scene. As a result, over a period of time, the camera traverses a "baseline" of undetermined length. The grounds for addressing such problem lie in the attempt to solve the *accuracy-precision* trade-off in stereo matching, which can be summarized as follows: due to quantization errors, the estimated disparity is more precise with a larger baseline, but the matching is less accurate, because of the exacerbation of perspective and radiometric nuisances that cause false and missing matches. There is manifestly a conflict between accuracy and precision, which motion-stereo approaches attempt to reconcile.

Early work in motion-stereo [28,16,22], integrates depth maps from different frames into a single map. They require motion and camera parameters to be known, and most of them restricts to lateral motion. A common drawback is that they warp the disparity map from frame to frame, thereby introducing errors and approximations that disrupt the prediction, and make the integration pointless. More recent motion-stereo approaches aggregate measures in a discretized 3D volume [30,18,32], but they need calibrated cameras as well.

The multiple-baseline approach [13,19,11] generalizes binocular stereo by computing an aggregated matching cost which considers all the images simultaneously, and then proceeds as in the binocular case. These methods require camera centers to be collinear (equivalent to lateral motion). Generalizations of these approaches can be found in the multi-view stereo literature, where the aggregated cost is computed along the optical ray in a discretized volume [7,6].

From the geometrical point of view, the problem raised by motion-stereo is how to set a common reference frame where measures from different images can be integrated. The discretized volume seems the natural choice, however computation in 3D space can be avoided by considering image-based quantities such as depth, binocular disparity or *planar parallax*. It will be shown in Section 2 that when camera parameters and its motion are unknown, planar parallax is a suitable depth-proxy that generalizes disparity and depth. This approach based on pixel-based measures – also called "iconic" – is motivated by applications like view synthesis, video interpolation and enhancement (frame rate up-conversion) and free viewpoint 3D TV.

In this work *we concentrate on the data fusion problem posed by the motion-stereo approach*, being agnostic with respect to: (i) the depth-proxy that is being used (ii) the binocular stereo matching algorithm, which is considered as part of the input of our method. As in [16,28], we use a *dynamic* approach, as we apply Kalman

* Corresponding author.
E-mail addresses: francesco.malapelle@uniud.it (F. Malapelle), andrea.fusiello@uniud.it (A. Fusiello), beatrice.rossi@st.com (B. Rossi), pasqualina.fragneto@st.com (P. Fragneto).

http://dx.doi.org/10.1016/j.image.2016.02.003
0923-5965/

filtering for recursive estimation of depth maps by combining measurements along the time line and within a spatial neighborhood. Pixel-wise depth measures are relaxed by considering the information coming from the neighbors within the same *superpixels*, using a spatial Kalman filter. In both temporal and spatial dimensions, the depth measures are trusted using confidence metrics attached to the measures.

An analogous result has been obtained in [16] by smoothing disparity maps with piecewise continuous splines, where a regularization-based smoothing is used to reduce measurement noise and to fill in areas of unknown disparity. Other methods perform adaptive smoothing in a *edge-aware fashion*, e.g. [12] where temporal consistency is enforced among different depth maps using an edge-aware Gaussian filtering extended to the temporal dimension in video volumes, or [25] where the depth map is filled by solving a least square error problem using edge and temporal information as weights. With respect to our approach, the key difference is that these works are post-processing approaches that aim at improving the quality of depth maps whereas our method uses edge information (in the form of superpixels) to be aware of which neighbors are relevant *while* updating depth values on the current reference map.

A preliminary version of this work appeared in [14] without the spatial relaxation.

1.1. Contribution

The main contribution of this paper is a data-fusion framework for motion-stereo, integrating both temporal and spatial information. Also, a comprehensive review of the available depth-proxies is presented in a unified framework and it is shown how planar parallax can be applied with general motion and unknown camera parameters.

The paper can be seen as a general-motion, uncalibrated extension of a classical work [16], which constrained motion to be lateral and required camera internal parameters. Moreover, [16] warped the disparity map from frame to frame, thereby introducing errors and approximations that disrupted the prediction (as shown by our experiments), whereas we fix this by keeping the reference frame constant.

We also report an extensive comparison of several confidence measures in the context of our approach.

1.2. Paper structure

This paper is structured as follows: in Section 2 we survey some background knowledge, in particular we present three suitable candidates for the depth-proxy. In Section 3 we present our method: stereo processing is described in Section 3.1, which produces the input data for the subsequent step in the form of depth measures. Then, the actual core of the algorithm (described in Sections 3.2 and 3.3) merges input measurements into the final result. In Section 4 we report experimental results and we draw conclusions in Section 5.

2. Background: depth-proxies

We do not make any hypothesis on whether the camera is calibrated or not, or if motion is constrained/known or not. These assumptions affect the choice of the *depth-proxy*. Several depth-proxies can be computed depending on factors such as the constraints on the motion of the camera and/or the availability of the perspective projection matrices. The depth-proxy must depend only on the reference frame and not on the other frames being considered. In this way each iteration provides a new estimate commensurate with others. In this section we present three suitable candidates.

2.1. Depth

The depth of a point is its distance from the focal plane of the camera (Fig. 1). If the interior camera parameters are available, stereo correspondences can be converted directly into depth values. The depth values for a given pixel obtained from subsequent frames are directly comparable.

Let $\mathbf{M}$ be a 3D point and let ($\mathbf{m}_r$, $\mathbf{m}_i$) be its projections onto the image planes I_r and I_i respectively. Let $P_r = K_r[R_r|\mathbf{t}_r]$ and $P_i = K_i[R_i|\mathbf{t}_i]$ be the perspective projection matrices of the two cameras (that must be known). The equation of the epipolar line of $\mathbf{m}_r$ in I_i is

$$\zeta_i \mathbf{m}_i = \mathbf{e}_i + \zeta_r K_i R_i R_r^\top K_r^{-1} \mathbf{m}_r \tag{1}$$

where $\mathbf{e}_i = K_i(t_i - R_i R_r^\top t_r)$ is the epipole and ζ_r and ζ_i are the unknown depths of $\mathbf{M}$ (with reference to P_r and P_i, respectively). Thus we can write

$$\mathbf{e}_i = \zeta_i \mathbf{m}_i - \zeta_r \mathbf{m}'_r \tag{2}$$

where $\mathbf{m}'_r = K_i R_i R_r^\top K_r^{-1} \mathbf{m}_r$. Since the three points $\mathbf{e}_i$, $\mathbf{m}'_r$ and $\mathbf{m}_i$ are collinear, one can solve for ζ_r using the following closed form expression [10]:

$$\zeta_r = \frac{(\mathbf{e}_i \times \mathbf{m}_i)(\mathbf{m}_i \times \mathbf{m}'_r)}{\|\mathbf{m}_i \times \mathbf{m}'_r\|^2}. \tag{3}$$

Since in real situations camera parameters and image locations are known only approximately, the back-projected rays do not actually intersect in space. However, it can be shown [10] that Formula (3) solves Eq. (2) in a least squares sense.

The actual computation of depth values is performed by applying Eq. (3): $\mathbf{e}_i$ is obtained as the projection of the optical center of the reference camera C_r, through the second camera P_i; the set of dense correspondences $(\mathbf{m}_r^k; \mathbf{m}_i^k)$ with $k = 1, \ldots, K$, where K is the number of correspondences for the current image pair, is known from the stereo matching step; image points $\mathbf{m}'^k_i$ are computed according to Eq. (2).

Please observe how this formulation elegantly avoids the explicit triangulation of $\mathbf{M}$, which would be required in a naive approach.

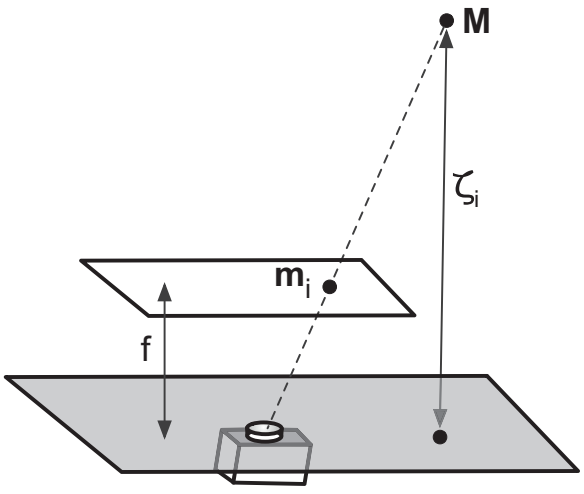

Fig. 1. The depth ζ_i is the distance of the 3D point from the focal plane of the camera (shaded in the picture).