ISPRS Journal of Photogrammetry and Remote Sensing 91 (2014) 29–49

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

On accurate dense stereo-matching using a local adaptive multi-cost approach

C. Stentoumis [a,*], L. Grammatikopoulos [b], I. Kalisperakis [b], G. Karras [a]

[a] Laboratory of Photogrammetry, Department of Surveying, National Technical University of Athens, GR-15780 Athens, Greece
[b] Laboratory of Photogrammetry, Department of Surveying, Technological Educational Institute of Athens, GR-12210 Athens, Greece

ARTICLE INFO

Article history:
Received 29 April 2013
Received in revised form 10 February 2014
Accepted 10 February 2014
Available online 6 March 2014

Keywords:
Photogrammetry
Matching
Stereo
Hierarchical
Cost
Reconstruction

ABSTRACT

Defining pixel correspondences among images is a fundamental process in fully automating image-based 3D reconstruction. In this contribution, we show that an adaptive local stereo-method of high computational efficiency may provide accurate 3D reconstructions under various scenarios, or even outperform global optimizations. We demonstrate that census matching cost on image gradients is more robust, and we exponentially combine it with the absolute difference in colour and in principal image derivatives. An aggregated cost volume is computed by linearly expanded cross skeleton support regions. A novel consideration is the smoothing of the cost volume via a modified 3D Gaussian kernel, which is geometrically constrained; this offers 3D support to cost computation in order to relax the inherent assumption of "fronto-parallelism" in local methods. The above steps are integrated into a hierarchical scheme, which exploits adaptive windows. Hence, failures around surface discontinuities, typical in hierarchical matching, are addressed. Extensive results are presented for datasets from popular benchmarks as well as for aerial and high-resolution close-range images.

1. Introduction

Generation of dense 3D information is a fundamental task in most applications in the fields of photogrammetry and computer vision (3D reconstruction, DSM production, object detection and recognition, automatic navigation, novel view synthesis, augmented reality). Methods for acquiring 3D information can be distinguished as *passive* and *active*. Image-based approaches (*passive*) are lately proven to be competitive to laser and optical scanners (*active*) in terms of accuracy, while exhibiting a clear advantage as regards cost and flexibility. Several theoretical alternatives exist for exploiting images in producing 3D information (shape from X). A core procedure is *image matching*, i.e. essentially the determination of correspondences among pixels. These approaches may be seen as consisting of two processes: establishment of *sparse* correspondences among images for camera calibration/orientation; and *dense* matching for 3D surface reconstruction. *Stereo-matching* algorithms for dense reconstruction, as that described in this paper, mostly exploit the *epipolar constraint*, hence they typically operate on rectified images to produce a *disparity map* (map ping of disparity values for every pixel of the reference image).

A significant number of efficient algorithms have been proposed for creating accurate disparity maps from single stereo-pairs. The effectiveness of such algorithms has been extensively discussed in several surveys (Dhond and Aggarwal, 1989; Banks and Corke, 2001; Scharstein and Szeliski, 2002; Brown et al., 2003). Scharstein and Szeliski (2002) have categorized algorithms by splitting them into four main components: *matching cost computation*, *support aggregation*, *disparity optimization* (*local* and *global*) and *disparity refinement*; publications addressing these components will be referred to below. Gong et al. (2007); Tombari et al. (2008) discuss the question of support region formation, while Hirschmüller and Scharstein (2009) evaluate the cost function itself under different optimization schemes. Enlightening comments are also found in Dhond and Aggarwal (1989); Brown et al. (2003). Wang et al. (2006); Nalpantidis et al. (2008) provide respective surveys focusing on criteria for hardware implementation and for real-time performance, while Tombari et al. (2010) have discussed the capabilities of fast stereo methods with low memory footprint. Evidently, difficulties exist in assessing stereo-methods due to diverging criteria set by different applications, as methods serving one purpose sometimes fail when the scenario changes.

* Corresponding author. Tel.: +30 2107722683; fax: +30 2107722677.
E-mail addresses: cstent@mail.ntua.gr (C. Stentoumis), lazaros@teiath.gr (L. Grammatikopoulos), ikal@teiath.gr (I. Kalisperakis), gkarras@central.ntua.gr (G. Karras).

http://dx.doi.org/10.1016/j.isprsjprs.2014.02.006
0924-2716/

In more detail, in *matching cost computation* a dissimilarity measure is attributed to each pixel for every value in the disparity range. A wide spectrum of such matching measures has been proposed over the years. Most common among them are the absolute difference of pixel intensities, their squared difference, their normalized cross correlation, as well as measures relying on input images transformed by filters such as the median, the mean, the LoG, or more sophisticated tools like bilateral filtering (Tomasi and Manduchi, 1998). Non-parametric image transformations, such as *rank* and *census* (Zabih and Woodfill, 1994), produce robust results based on relationships of pixels with their neighbourhood. Birchfield and Tomasi (1998) have proposed a dissimilarity measure to cope with differences in image sampling. Recently, the *mutual information* approach has been proposed for effectively handling radiometric differences (Hirschmüller, 2008); on the other hand, pixel-wise descriptor measures, like DAISY (Tola et al., 2008) or SIFT variations (e.g. Strecha et al., 2011), have yielded promising results in global formulations for wide-base stereo.

Cost computed per pixel is supported by a neighbourhood around pixels in the *cost aggregation* step. There exist three main approaches through which this question may be addressed: use of support weights, support regions of arbitrary shapes and variations of rectangular windows. Methods based on *support weights* make use of a window fixed in size and shape, and adjust the weights attributed to each neighbouring pixel. Weights can be calculated according to colour similarity and geo metric proximity (Yoon and Kweon, 2006) or additional criteria (Xu et al., 2002). Support regions of *arbitrary shape* re present an attempt to establish an optimal window shape and size. Theory from the field of image filtering has contributed the idea of shape-adaptive windows based on separate circular sectors across multiple directions around a pixel (Foi et al., 2007; Lu et al., 2008). Cross-based windows have been proposed by Zhang et al. (2009). *Rectangular windows*, and their variations for improving efficiency, are the most obvious choice thanks to their simplicity of implementation. Shiftable windows or windows anchored at pixels other than the central one (Kang et al., 1995; Fusiello et al., 1997; Bobick and Intille, 1999), as well as multiple windows relying on local variation of intensity and disparity (Kanade and Okutomi, 1994; Veksler, 2003), have also been proposed.

Disparity estimation in *local* (region-based) methods is usually performed in the *winner-takes-all* (WTA) mode, i.e. the disparity with the lowest aggregated cost is chosen. *Global* methods, on the other hand, perform *disparity optimization* on an energy function defined over all image pixels by simultaneously imposing a smoothness constraint. Regarding the latter, various approaches have been implemented based on partial differential equations (Faugeras and Keriven, 1998; Strecha et al., 2004; Ranftl et al., 2012), dynamic programming (Veksler, 2005), simulated annealing (Barnard, 1986), belief propagation (Sun et al., 2003; Felzenszwalb and Huttenlocher, 2004) and graph-cuts (Kolmogorov and Zabih, 2001; Boykov et al., 2001).

The last step, *disparity refinement*, aims at elaborating the disparity map. This can include correcting inaccurate disparity values and handling occlusion areas (Bobick and Intille, 1999). A common approach is to enforce constraints which are not explicitly implemented in the disparity optimization phase (Marr and Poggio, 1976; Yuille and Poggio, 1984; Brown et al., 2003). Typically, a sub-pixel estimation step is taken to increase the resolution of the disparity map, as most algorithms search in a discrete disparity space. This includes adjusting a curve to pixel cost for each disparity (Hirschmüller, 2008) and sub-pixel interpolation to disparity values (e.g. Yang et al., 2009 use a mean filter). In the past, Tian and Huhns (1986) had proposed intensity interpolation, a differential method and phase correlation. For more details on the variety of stereo-matching algorithms one may refer to dedicated on-line evaluation platforms, such as the Middlebury evaluation platform (Section 8.1) and the KITTI benchmark site (Section 8.5), where developments and new trends in the field are being continuously reported.

In this publication we address a number of open questions which, in principle, regard the performance of both local and global stereo methods. The successful estimation of disparities around discontinuities, which denote surface boundaries in 3D space, and within poorly textured areas is always a challenge. Regarding the latter case, sparse matching methods tend to fill untextured areas through interpolation, but it is of course preferable to obtain actual disparities. Furthermore, typical deficiencies to be dealt with are the "fronto-parallel effect" (flat surfaces in 3D space parallel to the image plane are favoured) in highly inclined surfaces and the quantization of disparities in discrete methods (e.g. MRF models, most local and semi-global approaches). At the same time, we wish to retain the computational efficiency required for real-time (and lately on-line) processes or high-resolution images. Our effort also aims at defining parameter-stable models, which is not a trivial task; most global methods involve handling a variety of parameters (i.e. the smoothness term is highly dependent on parameters defining the function, and hence on the image scenario), but also many local methods are sensitive to their empirically determined parameters.

Local methods are, typically, considered to be more straightforward and simple, and hence adequately fast for real-time applications, but of lower accuracy. On the other hand, methods based on a global optimization framework use elaborate models to describe the matching process; this often results in disparity maps of high quality, but at the cost of a computational load which may be restrictive for real-time tasks or large data manipulation. Notwithstanding this general remark, research work demonstrated in major evaluation platforms indicates that the limitations of each category are partially losing in significance: theoretical improvements and hardware upgrade make fast global implementations possible; at the same time, local methods achieve disparity maps of high accuracy, superior to several global methods, mainly via elaborate cost aggregation. Obviously, the increase of computational resources in personal computers, and even small dedicated processors, have made the exploitation of highly elaborate and demanding algorithms feasible, but at the same time image acquisition hardware keeps evolving and offering images with high spatial and radiometric resolution. Thus, algorithms offering lower complexity are still required in certain applications, while local approaches will probably continue to be easier to implement regardless of hardware.

Furthermore, manipulating large data, real-time and recent on-line applications, all require speed. This objective is partially served through hardware implementation. A wide range of such implementations can be found in surveys mentioned later in the text. Local methods have an inherent ability to bundle with commercial (e.g. in FPGAs) and special purpose hardware, as well as with their existing supportive software, e.g. CUDA for Nvidia GPUs. On the other hand, stereo matching methods based on global optimization algorithms are in general difficult to implement on hardware. Their complex and usually iterative nature makes defining and running of parallel processes unappealing, i.e. parallelism is difficult in MRF because of variable connectivity. This, of course, is not to suggest that fast and accurate global algorithms do not exist, but constructing them needs more time and effort.

Bearing in mind the above considerations, this work presents a hierarchical matching scheme based on local patch-based matching in a way that the previously discussed requirements are met, while keeping the complexity of the algorithm substantially low. We have chosen to extend the cross-based support regions