J. Vis. Commun. Image R. 26 (2015) 115–131

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

An iterative adaptive multi-modal stereo-vision method using mutual information ☆

Mustafa Yaman *, Sinan Kalkan

Dept. of Computer Engineering, Middle East Technical University, Ankara, Turkey

ARTICLE INFO

Article history:
Received 28 May 2014
Accepted 17 November 2014
Available online 24 November 2014

Keywords:
Multi-modal stereo-vision
Mutual information
Adaptive windowing
Adaptive cost aggregation
Iterative stereo
Plane fitting
RGB-D
Middleburry dataset

ABSTRACT

We propose a method for computing disparity maps from a multi-modal stereo-vision system composed of an infrared–visible camera pair. The method uses mutual information (MI) as the basic similarity measure where a segment-based adaptive windowing mechanism is proposed along with a novel MI computation surface with joint prior probabilities incorporated. The computed cost confidences are aggregated using a novel adaptive cost aggregation method, and the resultant minimum cost disparities in segments are plane-fitted in their respective segments which are iteratively refined by merging and splitting segments reducing dependency to initial segmentation. Finally, the estimated disparities are iteratively refined by repeating all the steps. On an artificially-modified version of the Middlebury dataset and a Kinect dataset that we created in this study, we show that (i) our proposal improves the quality of existing MI formulation, and (ii) our method can provide depth comparable to the quality of Kinect depth data.

1. Introduction

Using multi-modal cameras for surveillance systems has been popular since the year 2000 [1–4] since using cameras of different modalities, such as a pair of infrared and visible cameras, has advantages over using unimodal cameras in surveillance systems. These advantages include being able to work under low visibility or lighting conditions, better segregation of a target from the background, allowing a richer set of information like thermal signatures in the scene or the different reflectance properties of objects in different bands of the electromagnetic spectrum, etc. When considering to enhance the performance and usefulness of such multi-modal systems, the question of whether stereo-vision from multi-modal cameras can yield an accurate depth information or not has attracted well-deserved attention. One reason for this attention is that, for such systems, the distance of an intruder or the depth map of the scene under surveillance is very valuable.

A powerful method for computing depth from multiple cameras is stereo-vision. Stereo-vision [5,6] deals with computing depth by finding the corresponding pixels in different views. The correspondences, which are generally determined by comparing intensities of pixels, are used for computing the 3D positions using simple triangulation. It is one of the most studied problems of Computer Vision – for reviews, see [5,7–11]. Stereo-vision methods are mainly clustered around two main axes: *Sparse or feature-based* methods (*e.g.*, [8,12]) vs. *dense* methods (*e.g.*, [10,13]); and *local* methods (*e.g.*, [14,15]) vs. *global* methods (*e.g.*, [16,17]). The former grouping describes whether correspondences (and therefore the pixel disparities) are computed for all the pixels in the images (*i.e.*, the dense methods), or only for some reliable features (such as salient points, edges, corners and curves) extracted from the images. Regarding the latter grouping, local methods use only the local neighborhood and intensity information for finding stereo correspondences. Global methods, on the other hand, use global constraints to correct false correspondences that would be otherwise impossible to correct locally.

Although classical stereo-vision techniques have had tremendous success in terms of both accuracy and running time, they are not directly applicable in a multi-modal setting. The reason is that computing similarities between intensities of pixels or windows will not work using unimodal matching methods simply because the intensities of the corresponding pixels will be different. For example, an RGB-thermal image pair would have totally different intensities for corresponding pixels (see, *e.g.*, Fig. 1). This study aims to investigate how to compute reliable stereo correspondences for such an image pair and compute its depth information.

☆ This paper has been recommended for acceptance by Yehoshua Zeevi.

* Corresponding author.

E-mail addresses: mustafa.yaman@ceng.metu.edu.tr (M. Yaman), skalkan@ceng.metu.edu.tr (S. Kalkan).

http://dx.doi.org/10.1016/j.jvcir.2014.11.010
1047-3203/© 2014 Elsevier Inc. All rights reserved.

1.1. Related studies

Stereo-vision from multi-modal cameras was not studied much until the 2000s. The earliest of such studies, per the authors' knowledge, is from Egnal [15], who, influenced by Viola's studies of multi-modal registration [18], applied mutual information (MI) as the basic similarity measure for stereo correspondence. Egnal tested his method on images that were made multi-modal by red–blue filtering or altering the illumination of the different views. The results were promising and revealed the power of MI compared to standard correlation-based methods, especially on images with different spectral characteristics. However, using MI still not produced depth information of sufficient quality.

Fookes et al. extended the MI-based approach with adaptive windowing [19] and integrated prior probabilities using a 2D matching surface [20]. However, their methods were only tested on synthetically-altered unimodal images, which do not actually include different segmentation or the edge characteristics that genuine multi-modal images have. Nonetheless, Fookes's contributions are important for showing that stereo-vision using mutual information could be significantly enhanced when combined with other state-of-the-art stereo-vision techniques.

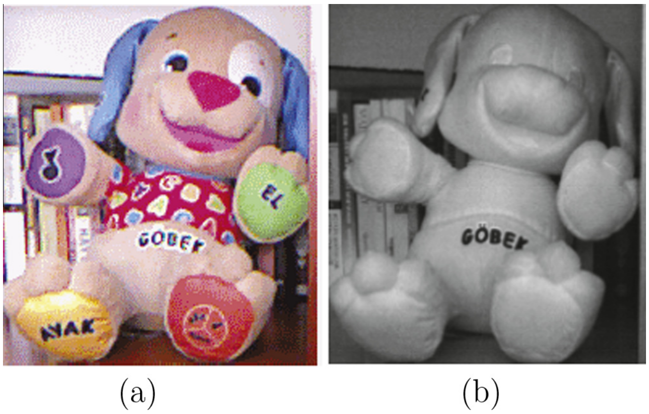

Fig. 1. An example illustrating the difficulty of finding correspondences in an IR-RGB image pair. (a) The RGB image. (b) The IR image.

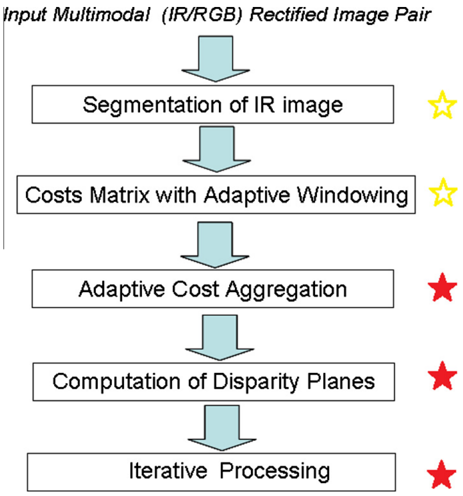
Input Multimodal (IR/RGB) Rectified Image Pair
Segmentation of IR image ☆
Costs Matrix with Adaptive Windowing ☆
Adaptive Cost Aggregation ★
Computation of Disparity Planes ★
Iterative Processing ★

Fig. 2. Overview of our method. The red (filled) stars are the extensions over the preliminary version of our work [28], and the yellow (empty) stars are the steps that are modified compared to our previous work. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Later, Krotosky and Trivedi [1–3] used mutual information for an infrared–visible camera pair in order to detect and track pedestrians. They applied mutual information for stereo correspondence within regions of interests (ROI) including human bodies, and proposed a disparity voting method for computing the final depth information of the corresponding regions as a significant restriction. Finally, this depth information was used to accurately register the multi-modal images for the ROIs.

In a very recent work on multi-modal stereo-vision, Campo et al. [21] proposed an MI-based method where the similarity measures were extended using the gradient information. They developed a multi-modal stereo rig (with thermal and visible cameras) and a database. The 3D depth results presented in their work were quite sparse for the scenes tested; however, their results are promising for showing that stereo-vision is possible from images with very distinct spectral characteristics.

Recently, a measure, called local self similarity (LSS), originally proposed for image template matching [22], has been applied as a thermal-visible stereo correspondence measure by Torabi and Bilodeau [23]. They implemented a ROI-based image matching system by tracking people in the scene according to their silhouettes, and compared it against MI-based similarity descriptors. In their first publication [24], they showed that the LSS measure outperforms MI and HoG (Histogram of Oriented Gradients). Later, they used the LSS measure in an energy minimization framework, enhancing the results when compared to their previous work [25]. In a recent study [26], with more data, they compared LSS and MI with (i) "traditional" descriptors such as SIFT, SURF, HOG, (ii) binary descriptors such as Census, Fast REtina Keypoint (FREAK) or Binary Robust Independent Elementary Feature (BRIEF) and (iii) direct comparisons of windows based on SSD, NCC. In their study, MI and LSS were shown to be the leading measures for ROI-based image matching of human silhouettes. MI outperformed LSS showing that it is still the best choice for multi-modal image windows matching; however, for smaller window sizes where the objects of interest were small or segmented into small fragments or there were many occlusions between objects, LSS performed better. On the other hand, LSS measure has not yet been tested for a dense disparity map estimation and still requires larger windows than is used in our study. Moreover, it is computationally more expensive, and performs poorly on uniform regions or small regions at salient points that are dissimilar to their neighboring regions [23]. Such regions constitute non-informative descriptors and for this reason, they are eliminated in the beginning of their method, which makes their method sparse, *i.e.*, not suitable for dense disparity map calculation.

1.2. The current study

In this article, we propose a new multi-modal stereo-vision method based on mutual information which can accurately generate *dense* disparity maps of images taken from cameras of different modalities. The method is compared to previous MI-based methods in the literature quantitatively and visually, and it is shown to outperform them. The contributions of the article are summarized as follows:

- Contribution of two datasets for evaluating multi-modal stereo-vision methods. One is based on cosine-transformed versions of the widely-used Middlebury Stereo Evaluation Dataset [27], and the other is collected from the RGB and IR cameras of a Kinect device.
- Adaptive computation of the window used in computing the cost matrix. The adaptively sized and shaped windows for matching the pixels are determined by the segments in the images, and in turn, these windows help generate a robust