Computer Vision and Image Understanding 117 (2013) 1736–1747

Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

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

A LSS-based registration of stereo thermal–visible videos of multiple people using belief propagation

Atousa Torabi *, Guillaume-Alexandre Bilodeau

LITIV, Department of Computer and Software Engineering, École Polytechnique de Montréal, P.O. Box 6079, Station Centre-ville, Montréal, Québec H3C 3A7, Canada

ARTICLE INFO

Article history:
Received 5 March 2012
Accepted 15 January 2013
Available online 12 July 2013

Keywords:
IR camera
Visible camera
Dense stereo matching
Belief propagation
Local self-similarity
Multimodal video registration

ABSTRACT

In this paper, we propose a novel stereo method for registering foreground objects in a pair of thermal and visible videos of close-range scenes. In our stereo matching, we use Local Self-Similarity (LSS) as similarity metric between thermal and visible images. In order to accurately assign disparities to depth discontinuities and occluded Region Of Interest (ROI), we have integrated color and motion cues as soft constraints in an energy minimization framework. The optimal disparity map is approximated for image ROIs using a Belief Propagation (BP) algorithm. We tested our registration method on several challenging close-range indoor video frames of multiple people at different depths, with different clothing, and different poses. We show that our global optimization algorithm significantly outperforms the existing state-of-the art method, especially for disparity assignment of occluded people at different depth in close-range surveillance scenes and for relatively large camera baseline.

1. Introduction

In the recent years, by reduction in the price of infrared sensors, there has been a growing interest in visual surveillance using thermal–visible imaging system for civilian applications. The advantages of jointly using a thermal camera with a visible camera have been discussed comprehensively in [1–4]. Combining visible and infrared information allows to better handling shadow, reflection, noise, misdetection, and missing information. The combined data enables better detection and tracking of people. Moreover, for human activity analysis, the joint use of thermal and visible data enables us to better detect and segment the regions related to the object that people may carry based on their temperature differences compared to the human body.

A fundamental issue associated to data fusion of close-range thermal–visible imaging is accurately registering corresponding information and features of images with dramatic visual differences. For a close-range scene, matching corresponding features in a pair of visible and thermal videos is much more difficult than for a long-range scene. People might be in very different sizes due to their distances to the camera, in different poses, and at different levels of occlusion. They might have colorful/textured clothes that are visible in color images, but not in thermal images. On the other hand, there might be some textures observable in thermal images caused by the amount of emitted energy from different parts of the human body that are not visible in a color image. Due to the high differences between thermal and visible image characteristics, finding correspondence for entire scene is very challenging. Instead registration is focused on the foreground ROIs.

The dense two-frame stereo correspondence is the only viable approach for registering possibly occluded objects at mutiple depths in the scene. Stereo matching is a well-studied subject for unimodal imaging system. An extensive taxonomy of two-frame stereo correspondence algorithms is described in [5]. However, this subject is new for multimodal visual surveillance applications. We summarize the problems associated to multimodal dense stereo as follows:

- *Dissimilar patterns.* This problem is specific to multimodal dense stereo. It is caused by the different types of image modalities. The corresponding regions in two images might be differently textured or one textured while the corresponding one is homogenous.
- *Depth discontinuities.* This difficulty is caused by segmentation results that contain two or more merged objects at different depths in the scene. In this case, correct disparities might be significantly different between neighboring pixels located on the depth boundaries.
- *Occlusions.* Some pixels in one view might be occluded in the other view. Therefore they should not be matched with pixels in the other view.

The main motivation of our proposed algorithm is the limitation of current approaches for registering occluded people ROIs. In this

* Corresponding author.
E-mail addresses: atousa.tora bi@polymtl.ca (A. Torabi), guillaume-alexandre.bilodeau@polymtl.ca (G.-A. Bilodeau).

1077-3142/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cviu.2013.01.016

paper we present a global optimization algorithm for partial image ROI registration. we formulate a multimodal stereo matching in a Markov Random Fields (MRFs) framework using color and motion information as smoothness assumptions in order to elegantly handle depth discontinuities, occlusions, and non-informative pixels caused by dissimilar patterns (corresponding pixels that do not contain similar visual information). Applying global optimization to multimodal stereo problem is challenging since most similarity measures, which are used for color images, are not viable for multimodal images. We integrate LSS as similarity measure in our global optimization algorithm.

The rest of the paper is organized as follows: The overview of the current multimodal registration approaches that gives insight about the limitations of exisiting methods is presented in Section 2. In Section 3, we describe the strengths of LSS as a viable image feature for matching thermal and visible images. In Section 4, the overview of our registration system is presented, and, in Section 5 each step of our algorithm is described in details. Our experiment is presented in Section 6 and demonstrate that our method is efficient for video surveillance applications and outperforms the current state-of-the-art method. Finally, in Section 7, we conclude this paper by describing the advantages and limitations of our algorithms.

2. Related works

In the thermal–visible video surveillance research context, the majority of the image registration approaches are related to global image registration that globally transform a reference image on the second image. Krotosky and Trivedi give a comparative survey of multimodal registration approaches [6]. Global transformation approaches, either extract low-level image features such as edge features [7], or temporal–spatial features such as object trajectories [8,9] to estimate a transformation matrix that transforms one image on another with the assumption that all the objects in the scene approximately lie in one depth plane. A few works in literature cover a video registration method appropriate for close-range people monitoring. These methods have been categorized as partial image ROI registration [6].

In previous partial image registration approaches excluding ours [10,11,4], MI is the only similarity measure used in local dense correspondence algorithm for human monitoring applications [6,12,13]. The accuracy of MI as a similarity metric is directly affected by the MI window sizes. For unsupervised human monitoring applications, obtaining appropriate MI window sizes for the registration of multimodal pairs of images containing multiple people with various sizes, poses, distances to cameras, and different levels of occlusion is quite challenging. In the video surveillance context, Chen et al. proposed a MI-based registration method for pairs of thermal and visible images that matches windows on foreground regions in the two images with the assumption that each window contains one single depth plane [12]. In their method, the problem of depth discontinuity inside an ROI was not addressed. Later, Krotosky and Trivedi proposed a MI-based disparity voting (DV) matching approach [6]. Their method, for each ROI column, computes the number of votes related to each disparity and assigns a disparity with maximum votes. Their method theoretically considers depth discontinuities that may occur between neighboring columns, but it ignores vertical depth discontinuity where the pixels on a column belong to multiple depths. For example, two people with different heights, where the shorter person is in front of the taller one. To the best of our knowledge, in our context of visual surveillance, all the existing methods for multimodal stereo matching are local correspondence approach.

Recent global stereo algorithms have achieved impressive results by modeling disparity image as Markov Random Field (MRF) and determining disparities simultaneously by applying energy minimization method such as belief propagation [14–16], and graph cuts (GC) [17,18]. Tappen and Freeman have shown that GC and BP produce comparable results using identical MRF parameters [19]. Sun et al. proposed a probabilistic framework to integrate into BP model, additional information (e.g., segmentation) as soft constraints [14]. Moreover, they have shown that the powerful message passing technique of BP deals elegantly with textureless regions and depth discontinuity problems. Later, Felzenszwalb and Huttenlocher proposed an efficient BP algorithm that dramatically reduced the computational time [15]. Their method is interesting for time sensitive applications like video surveillance. More recently, different extension of this efficient BP was proposed in several works [20,21].

In our previous work, we have shown that Local Self-Similarity (LSS), as a similarity measure, is viable for thermal–visible image matching and outperforms various local image descriptors and similarity measures including MI, especially for matching corresponding regions that are differently textured (high differences) in thermal and visible images [11]. Also we presented an extensive study of MI and LSS as similarity measure for human ROI registration in [4]. In [10,4], we proposed a LSS-based local stereo correspondence using disparity voting approach for close-range multimodal video surveillance applications. In this work, we adopt LSS as similarity measure in an energy minimization stereo model using the efficient BP model [15].

3. MI and LSS for multimodal image registration

Mutual information (MI) is the classic dense similarity measure for multimodal stereo registration. The MI between two image windows L and R is defined as

$$\mathrm{MI}(L,R)=\sum_{l}\sum_{r}P(l,r)log\frac{P(l,r)}{P(l)P(r)}, \tag{1}$$

where $P(l,r)$, is the joint probability mass function and $P(l)$ and $P(r)$ are the marginal probability mass functions. $P(l,r)$ is a two-dimensional histogram $g(l,r)$ normalized by the total sum of the histogram. $g(l,r)$ is computed as for each point, the quantized intensity levels l and r from the left and right matching windows (L and R) increment $g(l,r)$ by one. The marginal probabilities $P(l)$ and $P(r)$ are obtained by summing $P(l,r)$ over the grayscale or thermal intensities.

Local Self-Similarity (LSS) is a descriptor that capture locally internal geometric layout of self-similarities (i.e., edges) within an image region (i.e., human body ROI) while accounting for small local affine deformation. Initially, this descriptor has been proposed by Sechtman and Irani [22]. LSS describes statistical co-occurrence of small image patch (e.g. 5 × 5 pixels) in a larger surrounding image region (e.g. 40 × 40 pixels). First, a correlation surface is computed by a sum of the square differences (SSD) between a small patch centered at pixel p and all possible patches in a larger surrounding image region. SSD is normalized by the maximum value of the small image patch intensity variance and noise (a constant that corresponds to acceptable photometric variations in color or illumination). It is defined as

$$S_p(x,y)=exp\left(-\frac{SSD_p(x,y)}{max(var_{noise},var_{patch})}\right). \tag{2}$$

Then, the correlation surface is transformed into a log-polar representation partitioned into e.g. 80 bins (20 angles and 4 radial intervals). The LSS descriptor is defined by selecting the maximal value of each bin that results in a descriptor with 80 entries. A