Available online at www.sciencedirect.com
ScienceDirect
Computer Vision and Image Understanding 106 (2007) 270–287
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

Mutual information based registration of multimodal stereo videos for person tracking ☆

Stephen J. Krotosky *, Mohan M. Trivedi

Computer Vision and Robotics Research Laboratory, University of California, San Diego, 9500 Gilman Dr. 0434, La Jolla, CA 92093-0434, USA

Received 15 September 2006; accepted 23 October 2006
Available online 20 December 2006
Communicated by James Davis and Riad Hammoud

Abstract

Research presented in this paper deals with the systematic examination, development, and evaluation of a novel multimodal registration approach that can perform accurately and robustly for relatively close range surveillance applications. An analysis of multimodal image registration gives insight into the limitations of assumptions made in current approaches and motivates the methodology of the developed algorithm. Using calibrated stereo imagery, we employ maximization of mutual information in sliding correspondence windows that inform a disparity voting algorithm to demonstrate successful registration of objects in color and thermal imagery. Extensive evaluation of scenes with multiple objects at different depths and levels of occlusion shows high rates of successful registration. Ground truth experiments demonstrate the utility of the disparity voting techniques for multimodal registration by yielding qualitative and quantitative results that outperform approaches that do not consider occlusions. A basic framework for multimodal stereo tracking is investigated and promising experimental studies show the viability of using registration disparity estimates as a tracking feature.

Keywords: Thermal infrared sensing; Multisensor fusion; Person tracking; Visual surveillance; Situational awareness

1. Introduction

A fundamental issue associated with multisensory vision is that of accurately registering corresponding information and features from the different sensory systems. This issue is exasperated when the sensors are capturing signals derived from totally different physical phenomena, such as color (reflected energy) and thermal signature (emitted energy). Multimodal imagery applications for human analysis span a variety of application domains, including medical [1], in-vehicle safety systems [2] and long-range surveillance [3]. The combination of both types of imagery yields information about the scene that is rich in color, depth, motion and thermal detail. Once registered, such information can then be used to successfully detect, track and analyze movement and activity patterns of persons and objects in the scene.

At the heart of any registration approach is the selection of the most relevant similarity metric, which can accurately match the disparate physical properties manifested in images recorded by multimodal cameras. Mutual Information (MI) provides an attractive metric for situations where there are complex mappings of the pixel intensities of corresponding objects in each modality, due to the disparate physical mechanisms that give rise to the multimodal imagery [4]. Egnal has shown that mutual information is a viable similarity metric for multimodal stereo registration when the mutual information window sizes are large enough to sufficiently populate the joint probability histogram of the mutual information computation [5]. Further investigations into the properties and applicability of mutual information for windowed correspondence measure has been done by Thevenaz and Unser [6]. Challenges lie in obtaining these appropriately sized window regions for

☆ This research is sponsored by the Technical Support Working Group (TSWG) for Combating Terrorism, DHS and the U.C. Discovery Grant.
* Corresponding author.
E-mail addresses: krotosky@ucsd.edu (S.J. Krotosky), mtrivedi@ucsd.edu (M.M. Trivedi).

1077-3142/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2006.10.008

computing mutual information in scenes with multiple people and occlusions, where a balanced tradeoff between larger windows for matching evidence and smaller windows for registration detail is needed.

This paper presents the following contributions: we first give a detailed analysis of current methods to multimodal registration with a comparative analysis that motives our approach. We then present our approach for mutual information based multimodal registration. A disparity voting technique that uses the accumulation of disparity values from sliding correspondence windows gives reliable and robust registration and an analysis of several thousand frames demonstrates its success for complex scenes with high levels of occlusion and numbers of objects occupying the imaged space. An accuracy evaluation to ground truth measurements is presented and a comparative study using practical segmentation methods illustrates how the occlusion handling of the disparity voting algorithm is an improvement over previous approaches. Additionally, a basic framework for person tracking in multimodal video is presented and a promising experimental study is given to illustrate the use of the disparities generated from multimodal registration as a feature for tracking. We then discuss current algorithmic issues and potential resolutions for future research.

2. Multimodal registration approaches: comparative analysis of algorithms

In a multimodal, multicamera setup, because each camera can be at a different position in the world and have different intrinsic parameters, objects in the scene can not be assumed to be located at the same position in each image. Due to these camera effects, corresponding objects in each image may have different sizes, shapes, positions, and intensities. In order to combine the information in each image, it is required that the corresponding objects in the scene be aligned, or registered. Sensory measurements can then be fused or features combined in a variety of ways that can fuel algorithms that take advantage of the information provided from multiple and differing image sources [7]. Experiments in our previous work [2] have offered analysis and insight into the commonalities and uniqueness of the multimodal imagery. Multimodal image registration approaches vary based on factors such as camera placement, scene complexity and the desired range and density of registered objects in the scene. In order to better understand the algorithmic details of the various multimodal registration techniques, it is important to outline the underlying geometric framework for registration. Much of the multiple view geometry properties derived in this paper are adapted from Hartley and Zisserman [8].

Given a two camera setup with camera center locations C and C', a 3D point in space can be defined relative to each of the camera coordinate systems as $P = (X, Y, Z)^{\mathrm{T}}$ and $P' = (X', Y', Z')^{\mathrm{T}}$, respectively. The coordinate system transformation between P and P' is:

$$P' = RP + T \tag{1}$$

where R is the matrix that defines the rotation between the two camera centers and T is the translation vector that represents the distance between them. Additionally, the projection matrices for each camera are defined as K and K', where the projected points on the image plane are the homogeneous coordinates $p = (x, y, 1)$ and $p' = (x', y', 1)$.

Let π be a plane in the scene parameterized with N, the surface normal of the plane and d_π is the distance from the camera center C. Then a point lies on that plane if $N^{\mathrm{T}}P = d_\pi$. The homography induced by π is $P' = H_P P$ where:

$$H_P = R - T\frac{N^{\mathrm{T}}}{d_\pi} \tag{2}$$

Applying the projection matrices K and K', we have $p' = Hp$, where $H = K' H_P K^{-1}$ giving

$$H = K'\left(R - T\frac{N^{\mathrm{T}}}{d_\pi}\right)K^{-1} \tag{3}$$

This homographic transformation describes the transformation of points only when the points lie on the plane π (e.g. $N^{\mathrm{T}}P = d_\pi$). When a point does not lie on this plane, then an additional parallax component needs to be added to transformation equation to accommodate the projective depth of other points in the scene relative to the plane π. It has been shown in [8] that the transformation that includes the additional parallax term is:

$$p' = Hp + \delta e' \tag{4}$$

where e' is the epipole in C' and δ is the parallax relative to the plane π. The epipole is the intersecting point between the image plane and the line containing the optical centers of C and C'. The equation in (4) has effectively decomposed the point correlation equation into a term for the induced planar homography (Hp) and the parallax associated with points that do not satisfy the planar homography assumption ($\delta e'$). It is within this framework that we will describe the registration techniques used for multimodal imagery. Fig. 1 illustrates the main approaches to multimodal image registration that will be analyzed. Additionally, Table 1 provides a summary of references utilizing these approaches and indicates the assumptions, methods and limitations in each.

2.1. Infinite homographic registration

In Conaire et al. [9] and Davis and Sharma [3], it is assumed that the thermal infrared and color cameras are nearly colocated and the imaged scene is far from the camera, so that the deviation of pedestrians from the ground plane is negligible compared to the distance between the ground and the cameras. Under these assumptions, an infinite planar homography can be applied to the scene and all objects will be aligned in each image.