Image and Vision Computing 28 (2010) 1203–1219

Contents lists available at ScienceDirect

Image and Vision Computing

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

A multi-resolution area-based technique for automatic multi-modal image registration

Peter Bunting [a,*], Frédéric Labrosse [b], Richard Lucas [a]

[a] Institute of Geography and Earth Sciences, Aberystwyth University, Aberystwyth SY23 3DB, UK
[b] Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, UK

ARTICLE INFO

Article history:
Received 25 February 2008
Received in revised form 26 May 2009
Accepted 15 December 2009

Keywords:
Image registration
Remote sensing
Multi-modal
Correlation coefficient
Network

ABSTRACT

To allow remotely sensed datasets to be used for data fusion, either to gain additional insight into the scene or for change detection, reliable spatial referencing is required. With modern remote sensing systems, reliable registration can be gained by applying an orbital model for spaceborne data or through the use of global positioning (GPS) and inertial navigation (INS) systems in the case of airborne data. Whilst, individually, these datasets appear well registered when compared to a second dataset from another source (e.g., optical to LiDAR or optical to radar) the resulting images may still be several pixels out of alignment. Manual registration techniques are often slow and labour intensive and although an improvement in registration is gained, there can still be some misalignment of the datasets. This paper outlines an approach for automatic image-to-image registration where a topologically regular grid of tie points was imposed within the overlapping region of the images. To ensure topological consistency, tie points were stored within a network structure inspired from Kohonen's self-organising networks [24]. The network was used to constrain the motion of the tie points in a manner similar to Kohonen's original method. Using multiple resolutions, through an image pyramid, the network structure was formed at each resolution level where connections between the resolution levels allowed tie point movements to be propagated within and to all levels. Experiments were carried out using a range of manually registered multi-modal remotely sensed datasets where known linear and non-linear transformations were introduced against which our algorithm's performance was tested. For single modality tests with no introduced transformation a mean error of 0.011 pixels was identified increasing to 3.46 pixels using multi-modal image data. Following the introduction of a series of translations a mean error of 4.98 pixels was achieve across all image pairs while a mean error of 7.12 pixels was identified for a series of non-linear transformations. Experiments using optical reflectance and height data were also conducted to compare the manually and automatically produced results where it was found the automatic results out performed the manual results. Some limitations of the network data structure were identified when dealing with very large errors but overall the algorithm produced results similar to, and in some cases an improvement over, that of a manual operator. We have also positively compared our method to methods from two other software packages: ITK and ITT ENVI.

1. Introduction

In many situations, the quality and quantity of information that can be extracted from the fusion of data coming from multiple sources can be significantly increased when compared to the information that any of the datasets would individually provide. Remote sensing applications requiring data fusion include feature classification [32], change detection [12] and product integration. In all these cases, fusion requires accurate registration of the data [12].

To register data, manual methods have been used for some time within the remote sensing community. These typically involve the identification of common tie points between the pair of images to be registered. These might include road or river junctions or the edges of forest plantations, but they mark the same location in both images, thereby tying these positions together [5]. Manual registration is generally a slow and repetitive task, particularly as the operator can only identify tie points where features are distinct in both images. As such, tie point locations are often biased to regions of the image that are texturally diverse (e.g., urban areas, forest blocks). However, tie points should ideally be uniformly distributed across the image and as densely as possible, which is not always possible (e.g., where gradual changes in pixel value

* Corresponding author. Tel.: +44 1970621861.
E-mail addresses: pete.bunting@aber.ac.uk (P. Bunting), ffl@aber.ac.uk (F. Labrosse), rml@aber.ac.uk (R. Lucas).

0262-8856/$ - see front matter © 2010 Published by Elsevier B.V.
doi:10.1016/j.imavis.2009.12.005

occur or no features exist). As a result, the accuracy of georegistration cannot be guaranteed for all images or even all parts of one image. Because of these difficulties of the manual registration, an automated approach is therefore likely to be of benefit to the remote sensing community.

Classical automatic approaches mimic the manual process by identifying and aligning matching features, such as SIFT [29] or corners, within the two images considered. In many cases, the density of such features (corresponding to, e.g., roads and buildings) within the images acquired over forested areas will be low (see Section 3). This implies that methods only matching features are likely to perform poorly for such datasets as they will only sparsely specify the registration between images. Moreover, because the datasets correspond to images of different modalities taken from different view points at different points in time, it is likely that precisely matching features will not exist between the various images.

Instead, the approach adopted here used pixel-based metrics that compare windows from the entire images allowing tie points (corresponding to the centre of the windows) to be aligned on a topologically regular grid. This avoids the explicit extraction and matching of features from the images which is usually expensive and requires prior assumptions to be made regarding the content of the images. Such matching of windows works even if the images present very poor contrast or slow changing values, places where extracting features is unreliable, if not impossible. However, this does not mean that any such window will be suitable, only that their spatial density will be higher than that of features that can be extracted from such images (see Section 4.3). Moreover, given an appropriate window size, the matching captures enough information for it to perform adequately even under changing viewing conditions.

2. Background

Much work has already been carried out to achieve the goal of fully automated image registration, particularly within the fields of medical imaging and robotic vision, although the field of remote sensing is now starting to see real progress [44]. Two main methods have been adopted in image registration: (1) area matching techniques and (2) feature extraction and matching techniques. Area matching techniques [1,35,41,39,19] use an image similarity measure (e.g., correlation coefficient) to match windows of data from the two images, identifying tie points where the two windows match. Such techniques require no prior knowledge of the scene and operate even when features are not clearly defined (e.g., forests and deserts). Feature extraction and matching techniques [40,11] initially require the extraction of features and a separate process to then match and align those features. Such feature matching techniques often make assumptions on the type of features available for extraction (e.g., building corners or road junctions) or require specialist feature extraction techniques to be deployed for different data types (e.g., optical and radar). Moreover, the various stages of the process (particularly the matching of features) are usually computationally expensive. For these reasons an area-based technique was selected for this work and the remaining review will concentrate on such techniques. A fuller review of the field may be found in [44,2].

Area-based methods implicitly perform the feature extraction and matching steps associated with the alternative feature matching techniques by matching windows of data from the two images using a pixel-based similarity metric. Divergence in these techniques has occurred through the use of a number of similarity metrics and search strategies, multiple scales and resolutions to further improve accuracy, and complex data structures to help guide the search and relate neighbouring tie points.

Similarity measures can be divided into distinct groups including (1) image pixel intensity measures (e.g., correlation coefficient, Euclidean distance in image space; [25]), (2) transform domain (e.g., wavelets or Fourier transformation; [10,27]) and (3) joint histogram probability measures (e.g., mutual information [41]). Attempts have been made to merge some of these groups of metrics. For example, Woods et al. [42,43] proposed similarity measures combining image pixel values and joint histogram probabilities. Although the method showed promise against measures that only considered image pixel values on multi-modal data [19], they failed to show advantages over measures only considering joint histograms. Many studies relating image pixel intensity values focus on the correlation coefficient [34,38,6,14], although when using images with the same intensity range, Euclidean and Manhattan distances can also be deployed [1,25]. The correlation coefficient method has proved to be very effective at matching single modality data [6] although, as shown in the review of Inglada and Giros [19], is often less successful when applied to multi-modal datasets.

Methods in the frequency domain tend to be robust to image noise, which could equate to changes within the scene (e.g., fallen tree or building knocked down) and differences in illumination [2]. These often occur, particularly where temporal baselines between the image acquisitions are longer. Also, and because of the speed of the Fast Fourier Transformation (FFT), these methods have the potential to be faster than correlation based methods, particularly if hardware implementations of the FFT are used. However, similar to correlation, these methods failed to match multi-modal imagery where pixel values and patterns can change significantly between modalities.

A new class of similarity metrics based around the joint histogram of the two images [39] was therefore identified because of the difficulty in registering multi-modal datasets. Such joint histogram based measures make use of the probability of each pixel intensity pair and, although being slower to compute, offer good matching performance over multi-modal data [19]. A commonly used measure derived from the joint histogram is Mutual Information (MI; [41,39]) and has been seen by some as the leading method for multi-modal registration [44].

All measures of image similarity require the identification of a global optimum of the similarity between windows from the two images to identify the position of best image-to-image correspondence of the windows. A number of methods for identifying the global optimum have been proposed in the literature, including restricted exhaustive searches, hill climbing, simulated annealing [17,18,15], genetic algorithms [6], hierarchical grid [7] and other specialist search methods attempting to create a function representing the surface which can be used to guide the search direction [39,8,23]. Exhaustive searches are very time consuming although guarantee to identify the global optimum. To reduce the computation time, a restriction of the search domain can be applied if the global solution is known to exist within a short distance of the starting point. Hill climbing strategies on the other hand are fast as only those positions leading to an optimum are calculated, but the result often identifies a local rather than global optimum because such methods are of the "greedy" type. To solve this problem, a number of searches can be executed from a number of different random starting positions. The best result is then taken as the global optimum.

A number of studies have extended the problem to achieve subpixel accuracy of registration [16,36,8,27,25,20]. There are two common ways for this to be achieved; the first is through interpolation of the image data [33,9] while the second interpolates the similarity measure function [25]. Image interpolation has a number of drawbacks. Firstly, interpolating the image data increases the size of the search space. Secondly and more importantly, the