Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Feature matching evaluation for multimodal correspondence

M. Gesto-Diaz [a,*], F. Tombari [b,c], D. Gonzalez-Aguilera [a], L. Lopez-Fernandez [a], P. Rodriguez-Gonzalvez [a]

[a] Cartographic and Land Engineering Department, University of Salamanca, Hornos Caleros 50, 05003 Avila, Spain
[b] DISI, University of Bologna, V.le del Risorgimento 2, Bologna, Italy
[c] CAMP, Technische Universität München (TUM), Boltzmannstr. 3, Garching b. München, Germany

ARTICLE INFO

Article history:
Received 13 August 2016
Received in revised form 4 May 2017
Accepted 8 May 2017

Keywords:
Features
Keypoints
Multimodal
Image matching
Detector
Registration
Descriptor

ABSTRACT

This paper proposes a study and evaluation of approaches aimed at image matching under different modalities, together with a survey of methodologies used for performance comparison in this specific context, and, finally, a novel algorithm for image matching. First, a new dataset is introduced to overcome the limitations of existing datasets, which includes modalities such as visible, thermal, intensity and depth images. This dataset is used to compare the state of the art of feature detectors and descriptors. Template matching techniques commonly used to carry out multimodal correspondence are also adapted and compared therein. In total, 28 different combinations of detectors and descriptors are evaluated. In addition, the detectors' repeatability and the assessment of matching results based on Receiving Operating Characteristic (ROC) curve associated to all tested detector-descriptor combinations are presented, highlighting the best performing pairs. Finally, a novel Adaptive Pairwise Matching (APM) algorithm created to improve the robustness of matching towards outliers is also proposed and tested within our evaluation framework.

1. Introduction

Determining similarity between visual data is necessary in many computer vision tasks (Viola and Jones, 2001; Belongie et al., 2002; Tissainayagam and Suter, 2005; Zitova and Flusser, 2003). Methods for performing these tasks are usually based on representing an image using some global or local image properties (*features*) and comparing them using a similarity measure. However, most of the existing methods are designed for matching images within the same modality or under similar imaging conditions. They often fail when are applied to data acquired from different sensor modalities or under different photometric conditions. In such cases the sought pattern may exhibit linear or non-linear variations in the tone mapping due to changes in illumination conditions, intrinsic camera parameters, viewing positions, different modalities, etc.

The majority of matching strategies of image pairs follow a methodology that has been well introduced in Zitova and Flusser (2003). This methodology encloses four steps: (i) feature detection, (ii) feature matching, (iii) transform model estimation and (iv) image resampling.

In this paper, combinations of different state-of-the-art detectors and descriptors are analysed to find which setup gives the best performance in matching pairs of images which exhibit strong tone mapping variations due to the aforementioned reasons. In addition, an adaptive pairwise matching (APM) approach is proposed, aimed at outlier rejection to refine the transformation estimation. To carry out this evaluation, a specific dataset is introduced, which includes relevant application-wise combinations of different modalities, such as depth data (acquired with a gaming sensor, Kinect II) paired with thermal images. In the case of Kinect II, a novel approach which registers directly depth to visible can avoid the registration errors (Gesto-Diaz et al., 2015) presented in the device between colour and depth and also allows to register the Kinect device with multiple devices with different modalities, such as thermal spectrum.

There have been different works on multimodal correspondence based on self-similarity. In Huang et al. (2011) different methods to build the self-similarity descriptor are compared and applied in multimodal images (visible against LiDAR and visible with different illumination conditions). In Bodensteiner et al. (2010) a comparison to match local patch regions using descriptors prone to multimodal image matching (MI and self-similarity) is

[*] Corresponding author.
E-mail addresses: mgesto@usal.es (M. Gesto-Diaz), federico.tombari@unibo.it (F. Tombari), daguilera@usal.es (D. Gonzalez-Aguilera), luisloez89@usal.es (L. Lopez-Fernandez), pablorgsf@usal.es (P. Rodriguez-Gonzalvez).

http://dx.doi.org/10.1016/j.isprsjprs.2017.05.007
0924-2716/© 2017 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

applied. The modalities in this case were visible against infrared or LiDAR. Heinrich et al. (2012) presented a new descriptor for matching images with different modalities based on the principles of self-similarity applied to medical imagery modalities. There are also other studies presenting new breakthroughs to tackle multimodal matching. For instance, in Kim et al. (2014) authors propose a new descriptor based on the frequency of self-similarity for matching near-infrared and visible images. In Senthilnath et al. (2013) a new feature matching descriptor, Discrete Particle Swarm Optimization (DPSO), is introduced and combined with one keypoint detector. In Tombari and Di Stefano (2014) proposed a new keypoint detector based on self-dissimilarity to find interest points, applying this methodology also on a multimodal dataset. Another contribution based on SIFT (Cheung and Hamarneh, 2007) introduces a new descriptor to match across medical images. Other studies (Torabi et al., 2011), perform a comparison of several descriptors used for multimodal matching (visible and thermal), but they were paired to only one specific feature detector. In Senthilnath and Prasad (2014) an interesting variation of the framework for matching multimodal images based on SIFT and a genetic algorithm for matching is presented.

On the other hand, several works have presented registration alternatives to feature-based matching, most of them employed in medical imaging. The idea of these approaches is to use some kind of similarity among images, one of these approaches is Mutual information (MI) developed by Viola and Wells (1997). When MI became popular some novel approaches were inspired, such as the approach developed by Wachowiak et al. (2004) where a method to register images based on the normalization of the MI was presented. The approach presented by Hel-Or et al. (2014) is a method for pattern recognition in images with different modalities, with an inspiration on MI but with a different approach, called Multi Tone Mapping (MTM).

To the best of our knowledge, there is no work in literature which takes into consideration a range of modalities (visible, thermal, LiDAR intensity and depth images).

The detectors used in this work are the following ones: the well established SIFT (Scale Invariant Feature Transform) (Lowe, 1999) and SURF (Speeded-Up Robust Features) (Bay et al., 2006), and more recent approaches such as ORB (Oriented FAST (Features from Accelerated Segment Test) (Rosten and Drummond, 2006) and Rotated BRIEF (Binary Robust Independent Elementary Features) (Calonder et al., 2010; Rublee et al., 2011) MSD (Maximal Self-Dissimilarities) (Tombari and Di Stefano, 2014). These detectors are used in combination with several descriptors. Each descriptor originally proposed together with the introduced detectors is used. Except for the case of MSD that was proposed without any specific descriptor. Some descriptors also are included in this work that have already been used for multimodal correspondence in previous works, e.g., LSS (Local Self Similarity) (Shechtman and Irani, 2007) and HOG (Histogram Oriented Gradients) (Dalal and Triggs, 2005). Furthermore, MI (Mutual information) (Viola and Wells, 1997) and MTM (Multi-Tone Mapping) (Hel-Or et al., 2014) are two template matching solutions extensively used for multimodal correspondence. To include them into our evaluation framework these two popular techniques are adapted to work like a descriptor. The 4 detectors (MSD, ORB, SIFT and SURF) combined with the 7 descriptors (HOG, LSS, MI, MTM, ORB, SIFT and SURF) provide 28 possible combinations for the comparison.

Finally, a novel Adaptive Pairwise Matching (APM) for pairwise image matching is proposed. This method automatically selects the best correspondences to determine the transformation estimation between a pair of images, including an outlier removal method, that can be RANSAC (Random Sample Consensus) (Fischler and Bolles, 1981) or LMedS (Least Median of Squares) (Zhang et al., 1995).

Importantly, public datasets for quantitative evaluation of algorithms for multimodal correspondence are quite limited. For this reason, we start by adopting the methodology and structure of the dataset presented in Mikolajczyk and Schmid (2005) (hereinafter referred to as Oxford dataset), which represents a reference benchmark for pairwise image matching, but only contains pairs of images acquired with an optical camera. Then, we add several image pairs acquired under different modalities, hence obtaining a bigger dataset, which we plan to publicly release upon publication of this work.[1] For what concerns the performance evaluation, we present comparative results in terms of keypoint repeatability for the evaluated detectors, as well as ROC curves for the evaluated descriptors, and the number of images registered using the proposal outlier rejection method compared to RANSAC and LMedS.

This paper has been structured as follows: Section 2 presents the materials used to create the image dataset for multimodal matching and the methods used to compare the detectors, descriptors and the novel matching algorithm. The experimental results are reported and discussed in Section 3, while final remarks and conclusions are drawn in Section 4.

2. Materials and methods

2.1. Multimodal dataset

Fig. 1 shows a set of 10 image pairs that we have added to those originally included in the Oxford dataset to perform our evaluation. This accounts for a total of 15 images pairs.

The modalities taken into account in this dataset are: thermal, visible, LiDAR intensity and depth images from a gaming sensor, Kinect II. The first four image pairs are visible with LiDAR. The fifth image pair is depth image with thermal. Finally, five more images pairs combining visible with LiDAR intensity images were included.

These image pairs are from two different sources: one is a synthetic, created virtually from a 3D real object with different acquired modalities (the first four image pairs). For these cases of synthetic images, it can be easily controlled rotation, scale and shear to evaluate the behaviour of the detectors and descriptors in the images affected for scale, rotation and shear. Please note, that in this case, the ground truth is perfectly defined by the transformation applied. In the remaining cases, the images have been extracted from real cases, where a method to obtain the ground truth with the transformation between each pair was used. In these cases, the sensors and positions used for acquiring the images are not the same, such as thermal and visible images; therefore, the intrinsic and extrinsic parameters are different. In addition to this, the tone mapping changes in a non-linear way due to changes in illumination conditions and its different modality. To quantitatively compare different solutions on this dataset, the ground truth represented by the transformation between each image pair needs to be obtained. The transformation used as ground truth is the fundamental matrix (Luong and Faugeras, 1996)

$$xFx^{0^T} = 0 \tag{1}$$

where x and x^0 are vectors of matching points presented in both images expressed in homogeneous coordinates.

This matrix (Eq. (1)) provides the transformation for a set of matching points between a pair of images. The methodology for the estimation of the fundamental matrix is the same as in Mikolajczyk and Schmid (2005). The fundamental matrix between the reference image and the other image in a particular dataset is

[1] Available for reviewing at http://tidop.usal.es/dataset/datasetmultimatching.7z.