Computer Vision and Image Understanding 000 (2016) 1–10

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Video registration in egocentric vision under day and night illumination changes

Stefano Alletto*, Giuseppe Serra, Rita Cucchiara

Dipartimento di Ingegneria "Enzo Ferrari", Università degli Studi di Modena e Reggio Emilia, Modena MO 41125, Italy

ARTICLE INFO

Article history:
Received 7 December 2015
Revised 25 July 2016
Accepted 19 September 2016
Available online xxx

Keywords:
Video registration
Egocentric vision
Visual matching

ABSTRACT

With the spread of wearable devices and head mounted cameras, a wide range of application requiring precise user localization is now possible. In this paper we propose to treat the problem of obtaining the user position with respect to a known environment as a video registration problem. Video registration, i.e. the task of aligning an input video sequence to a pre-built 3D model, relies on a matching process of local keypoints extracted on the query sequence to a 3D point cloud. The overall registration performance is strictly tied to the actual quality of this 2D-3D matching, and can degrade if environmental conditions such as steep changes in lighting like the ones between day and night occur. To effectively register an egocentric video sequence under these conditions, we propose to tackle the source of the problem: the matching process. To overcome the shortcomings of standard matching techniques, we introduce a novel embedding space that allows us to obtain robust matches by jointly taking into account local descriptors, their spatial arrangement and their temporal robustness. The proposal is evaluated using unconstrained egocentric video sequences both in terms of matching quality and resulting registration performance using different 3D models of historical landmarks. The results show that the proposed method can outperform state of the art registration algorithms, in particular when dealing with the challenges of night and day sequences.

1. Introduction

Egocentric vision, thanks to the widespread of cheap and powerful wearable cameras and devices, is increasing its spread among both researchers and consumers. Exploiting the unique first person perspective, many recent works have dealt with the study of self-gestures, social relationships or video summarization (Alletto et al., 2015; Betancourt et al., 2014; Lee and Grauman, 2015). While this new and unique perspective provides invaluable insights on the viewpoint of the user, challenging situations such as severe changes in the lighting of the environment or high motion blur occur and must be dealt with (Betancourt et al., 2015).

A relevant topic that has been recently studied but is yet to be brought to the egocentric field is video registration. That is, the task of precisely localizing an input sequence and, in the case of egocentric videos, the user, with regard to a pre-built 3D model (for example a building of historical interest). A precise estimation of the camera extrinsic parameters in a given timeframe, i.e. precise user localization, can be a significant starting point for several egocentric applications such as personalized tours in a city, assistive services or interactive environments.

The registration of images is a topic that has been widely studied in the past years (Li et al., 2012; Sattler et al., 2012; Schindler et al., 2007), on the other hand fewer works have dealt with the registration of video sequences and to the best of our knowledge the employment of egocentric videos has no precedent in literature. In fact, the unique perspective of first person camera views greatly differs from the ones employed in past works under several aspects. For example, fixed camera settings featuring cameras mounted on a van have been exploited, resulting in the acquisition of videos that display very constrained motion patterns and where the rigid setup provides accurate ground truth information about the extrinsic of the cameras used in the testing phase with regard to the ones used to build the Structure from Motion (SfM) model (Irschara et al., 2009; Kroeger and Van Gool, 2014). On the contrary, egocentric videos often display fast and unpredictable movements and can be acquired under very different conditions from the images or videos used to build the 3D model used in the registration.

Recent works (Sattler et al., 2012; Schindler et al., 2007) have established a standard pipeline for aligning images or video frames to a pre-build 3D model, which is based on two major steps: feature matching and camera localization. To address the

* Corresponding author.
E-mail addresses: stefano.alletto@unimore.it (S. Alletto), giuseppe.serra@unimore.it (G. Serra), rita.cucchiara@unimore.it (R. Cucchiara).

http://dx.doi.org/10.1016/j.cviu.2016.09.010
1077-3142/

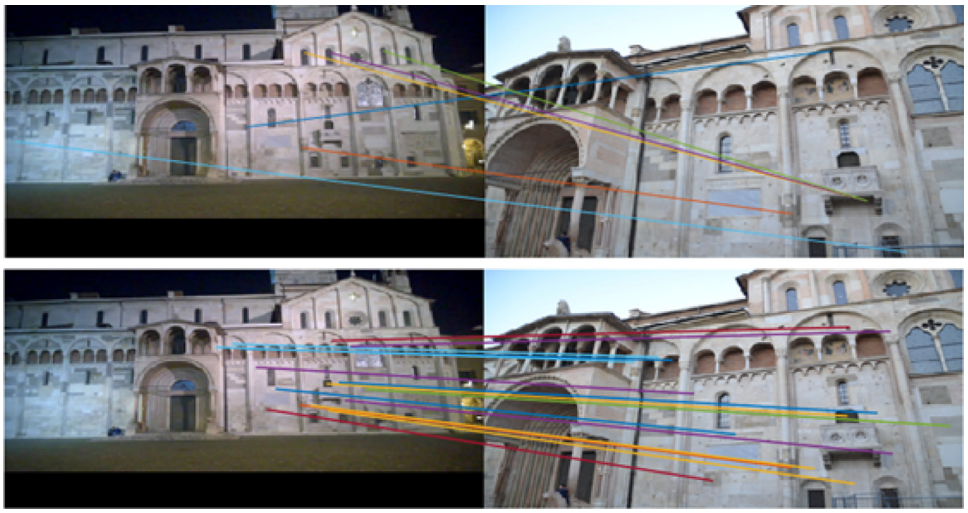

Fig. 1. Samples of the matching results. a video frame acquired in a night sequence (left) compared to a model image (right). Top: SIFT standard matching technique; bottom: the proposed approach.

first stage of the pipeline, a widely adopted approach is to extract SIFT feature keypoints and descriptors from a query image and then robustly match them against the descriptors composing the 3D model (Sattler et al., 2012). These correspondences form the 2D-3D matches that will be used to estimate the camera location in terms of rotation and translation matrices, using a Perspective-n-Point (PnP) algorithm often enclosed in a RANSAC loop (Schindler et al., 2007). The extrinsic parameters estimation, while using algorithms such as RANSAC in order to gain robustness against outliers, strongly depends on the quality of the initial feature matching between the query and the model. In fact, the resulting registration performance decreases if the number and quality of the correspondences found is not sufficient. A major challenge in the video registration from an egocentric perspective derives from the fact that first person videos can span multiples time of the day, and can result in being acquired during the night. Substantial experiments show how matching images acquired during the night against a 3D model built from a collection of images collected in normal, daytime lighting conditions, results in very poor matches, both in terms of quality and number of outliers.

To address the issues deriving from a poor match between query and reference images, recent methods focus on the improvement of the registration results using synthetic views or complex a-posteriori optimizations techniques (Irschara et al., 2009; Kroeger and Van Gool, 2014). Here, on the contrary, we propose to address the problem at its source and intervene on the matching procedure itself. In particular, we design a novel matching technique that aims at improving the number of scored matches while jointly decreasing the number of outlier. To do so, we propose a novel embedding space that maps local descriptors, its spatial arrangement and temporal robustness of employed keypoints in order to produce a descriptor robust to steep changes in lighting conditions. Our experiments show that this matching technique results in an increase in scored matches in both night and day sequences and in a subsequent improve in registration, without the need of a-posteriori optimization. Fig. 1 displays an example of the results obtained by standard SIFT matching on a night-day matching, and compares it with results achieved from our method.

The main contributions of this paper are the proposition of a novel embedding space that takes into account in its design the challenges posed by steep changes in illumination. This embedding space combines local feature descriptors with a representation of their surroundings based on the covariance of densely sampled features. This formulation is further extended to include temporal coherence by tracking local keypoints over a short time to assess their robustness and over-time stability. Finally, we experimentally show that our video registration proposal can cope with the challenges of night sequences with only a small loss in performance and display improved results when compared to current video registration state of the art methods.

2. Related work

Several approaches deal with the task of image registration treating it as an image retrieval problem, matching the query image against a database of images with annotated localization, i.e. their rotation and translation matrices aligning them to the desired 3D model (Bourmaud and Giremus, 2015; Li et al., 2012; Sattler et al., 2012; Schindler et al., 2007; Torii et al., 2015b; Zamir and Shah, 2014). These approaches tend to be slower due to the high number of comparisons required and can produce a localization that is only accurate at the scale of the single images in the database, but can benefit from established image retrieval methods. Schindler et al. (2007) deal with the task of city scale image localization using a bag-of-words representation of street view images. Similarly, Hays and Efros (2008) compute coarse geo-location information of a query image by matching it to a set of Flickr geo-tagged images. While these approaches can achieve significant performance in scenarios of large-scale localization such as city-scale, their localization is precise at most as the geo-location of the used images and GPS position is often not accurate enough when localizing a camera with regard to a model of a single building.

Aiming at the improvement of localization performance, the use of the 3D structure of the surrounding environment has been recently employed (Sattler et al., 2011). In fact, thanks to the recent advancements in Structure from Motion techniques, 3D models can be obtained by a small set of images and can be build on a city-scale with even with consumer computers (Wu, 2011). This results in a shift in paradigm where the descriptors computed on the query image are matched directly to the descriptors of the 3D point-cloud instead of having the intermediate step of matching with the images used to build said point-cloud. To most widely employed descriptors used are the local-invariant SIFT descriptors (Lowe, 2004), which are robust to scale variations and to moderate changes in viewpoint. Despite this progresses, the approaches that rely on the matching of interest points succeed only under moderate changes in visual appearance. Image registration in sequences where severe changes in lighting occur due to the

Please cite this article as: S. Alletto et al., Video registration in egocentric vision under day and night illumination changes, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.09.010