2016 IEEE International Symposium on Mixed and Augmented Reality Adjunct Proceedings

Improving Localization under Varying Illumination

Alexander Plopski*
Nara Institute of Science and Technology

Tomohiro Mashita
Osaka University

Akira Kudo
Osaka University

Tobias Höllerer
University of California, Santa Barbara

Kiyoshi Kiyokawa
Osaka University

Haruo Takemura
Osaka University

ABSTRACT

Localizing the user from a feature database of a scene is a basic and necessary step for presentation of localized augmented reality (AR) content. Commonly such database depicts a single appearance of the scene, due to time and effort required to prepare it. To account for appearance changes under different lighting we propose to generate the feature database from a simulated appearance of the scene model under a number of different lighting conditions. We also propose to extend the feature descriptors used in the localization with a parametric representation of their changes under varying lighting conditions. We compare our method with a standard representation and matching based on L_2-norm in a simulation and real world experiments. Our results show that our approach achieves a higher localization rate with fewer feature points and a lower process cost.

1 INTRODUCTION

Successful presentation of geo-allocated augmented reality (AR) content, e.g., in form of labeling or guidance systems, is dependent on successful user-localization. To present consistent AR this requires to not only determine the general position of the user, but an accurate 6DOF pose. Commonly, this is achieved through feature databases that represent the recorded environment. The appearance of these features however strongly depends on the environment illumination. The result is a decreased localization success rate if the illumination deviates from the time the database was recorded at. Addressing this through multiple reconstructions of the scene under different conditions leads to extensive databases, which cannot be stored easily on a mobile device that is used to view the content and may require an extensive time-span to record. Furthermore, to capture representations of possible changes in the environment would require an equally long time-span.

We address these issues by introducing a dual approach. First, we propose to use the continuously improving rendering capabilities of off-the-shelf rendering engines to generate virtual representations of the scene under varying lighting conditions. This method is viable for a large variety of scenes as the number of illumination sources is limited and can be obtained either manually, from construction plans, in case of buildings, or from the time of the day and the geographic coordinates. Second, we propose a new descriptor that captures appearance variations of feature points and can be constructed either from multiple recording, or as is done in this paper a simulation of various conditions of the scene.

2 RELATED WORK

The main contribution of our paper is an approach for generation of feature databases for localization and a feature matching approach that extends feature descriptors to account for feature variability.

*e-mail:plopski@is.naist.jp

Figure 1: Examples of images from 4 datasets with overlayed edges of the estimated pose.

2.1 Feature Matching

Over the past years a variety of descriptors have been developed to provide an efficient way to represent detected features.

Matching of SIFT [6] descriptors of detected corners based on L_2-norm has proven robust against orientation, scale and partially illumination changes. These descriptors have also found application in a variety of localization [2, 7] solutions. With the rise of mobile computing, modified descriptors further improved the localization results by incorporating the orientation of the phone [4] or the spatial size of the features [5].

Our work is in the spirit of the above work in that an extension of the commonly used features is applied to further improve the robustness of the matching. However, we differ from previous work in that the extension is based on the variance of the feature's appearance instead of additional sensor information.

2.2 Database Acquisition

To evaluate localization methods researchers have proposed and developed various methods to generate ground-truth information as well as acquire a representative feature database.

The developed methods use reconstruction of the target scene with Structure-from-Motion [7] or from large image databases [2]. Kurz et al. [5] use a laser scanner to recover a dense point-cloud representation of the environment and generate synthetic views by projecting the model into the camera.

To reduce the number of stored features and improve the representability, Irschara et. al [2] generate additional virtual views of the scene and keep the smallest subset that covers the targeted viewing area.

Kurz et al. [3] select a representative subset by sorting the features based on the number of successful detection They recover a representative feature subset according to the method of [3].

Our method resembles [2] and [5] in that a simulation and a dense 3D model is used to generate the feature database. Contrary to their assumption of a static appearance, we model the scene under varying illumination conditions.

978-1-5090-3740-7/16 $31.00 © 2016 IEEE
DOI 10.1109/ISMAR-Adjunct.2016.22

3 Mahalanobis Distance Based Matching

The varying appearance of a feature point under different lighting conditions can be seen as a cluster of feature vectors with a certain degree of variance of the descriptor parameters. In that sense, feature matching resolves the task of classifying a descriptor with a best-fit cluster. Hereby, we propose to use the Mahalanobis distance to account for covariance within each cluster. Furthermore, the stochastic representation of the feature cluster allows to approximate the feature descriptor for non-observed appearances. Let a feature cluster P be composed of m feature vectors $\boldsymbol{x}_i$, $i = 1 \ldots m$, that describe the appearance of the feature under varying conditions, e.g., camera pose and scene illumination. The mean of the cluster $\boldsymbol{\mu}_P$ and it's covariance matrix $\boldsymbol{\Sigma}_P$ are defined as

$$\boldsymbol{\mu}_P = \frac{1}{m}\sum_{k=1}^{m} \boldsymbol{x}_k, \tag{1}$$

$$\boldsymbol{\Sigma}_P = \frac{1}{m}\sum_{k=1}^{m} (\boldsymbol{x_k} - \boldsymbol{\mu_P})(\boldsymbol{x_k} - \boldsymbol{\mu_P})^{\mathsf{T}}. \tag{2}$$

The distance of a feature vector $\boldsymbol{x}$ to P is defined as

$$dist^{mah}(\boldsymbol{x}, P) = \sqrt{\frac{1}{m}(\boldsymbol{x} - \boldsymbol{\mu}_P)^T \boldsymbol{\Sigma}_P^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_P)}. \tag{3}$$

In some cases feature vectors in a cluster may display no width in some directions. To improve robustness against such cases, and to reduce the complexity, we apply Principal Component Analysis (PCA) to each cluster. The result is a compact vector whose elements have a strong descriptive power.

For each cluster our database thus contains $\mathbf{P}$, the position of the point described by it, $\boldsymbol{\mu}_P$, $\boldsymbol{\Sigma}_P$ and a projection matrix that maps the feature descriptor onto the principle component space.

4 Experiment

In our experiments we use SIFT feature descriptors to compute. Our environment was the Vienna concert hall model and the ground plane from the City of Sights dataset [1]. The model was printed out on heavy paper and reinforced with card board. Using such a model has the benefit that an accurate 3D model and texture of the target is already known and can be used with the simulation to construct the database. We have recorded 5 sequences of the model at different days and times of the day (Fig. 1) with an iPhone5S with the video mode set to 720p and 3 images/frame. For each recording we aligned the model with its position in the simulation with a compass.

4.1 Database from Simulation

To acquire the ground truth for each frame and verify our idea for database acquisition, we simulated the appearance of the model at the time of the recording with Unity[1], where the width of the model was assumed to be 40 m, and extracted a dataset of 5000 features, as described in [3]. The camera was then localized with 10000 iterations of RANSAC and the correctness was verified for every frame. Hereby, we displayed the contours of the estimated model pose in the image, as shown in Fig. 1, and manually verified the quality of the alignment. This suggests that our approach of creating the dataset for localization from a simulation can be used in an actual scenario.

4.2 Feature Matching

We conducted a comparison with L_2-norm matching. To acquire comparable databases, we synthesized the appearance of the model with different sun positions and illumination colors, and record its appearance for 50 pre-defined camera poses. A feature point $\mathbf{P}$ is represented by its cluster in our database, and potentially multiple SIFT descriptors for L_2 matching, where the descriptors were selected according to [3]. For this comparison we used test dataset taken under 5 different lighting conditions in the real environment and the number of principal components is defined as 16. We assume that a camera localization is correct, if it is within 0.5 m from the ground truth pose. Figure 2 shows the result with varying number of reference points. The localization rate with the L_2-norm increases with the number of reference points. As shown in Fig. 2, it outperforms our method for more than 200 reference points. However, the localization of our method remains relatively constant independent of the number of used reference points. Additionally, our method performs faster than L_2 norm.

[1] www.unity3d.com

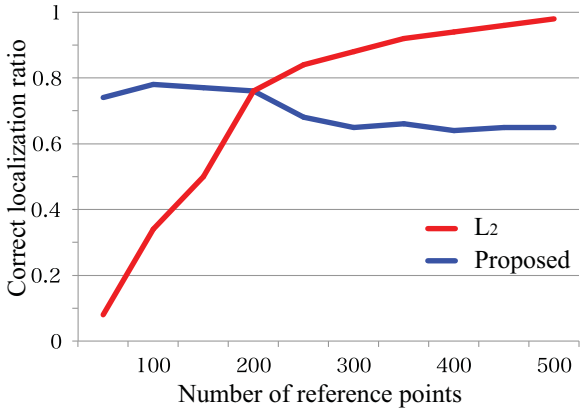

Figure 2: Localization results for the proposed method and L_2-norm matching over all images.

5 Conclusion

We have presented a novel method for localization that accounts for the impact varying lighting impacts the scene appearance. We use a simulation to predict the appearance of the scene under different illumination conditions and improve the robustness of the feature matching by accounting for the variance of feature descriptors. Our method performs faster and achieves a higher localization ratio than localization based on L_2-norm. In the future we want to investigate if better lighting simulation and subdivision of the database, e.g., by time or weather, can improve the results.

Acknowledgements

This work was partly supported by JSPS KAKENHI Grant Number JP16H02858 and JP16K16100.

References

[1] L. Gruber et al. The city of sights: Design, construction, and measurement of an augmented reality stage set. In *Proceedings of ISMAR*, pages 157–163, 2010.

[2] A. Irschara et al. From structure-from-motion point clouds to fast location recognition. In *Proceedings of CVPR*, pages 2599–2606, 2009.

[3] D. Kurz et al. Representative feature descriptor sets for robust handheld camera localization. In *Proceedings of ISMAR*, pages 65–70, 2012.

[4] D. Kurz et al. An outdoor ground truth evaluation dataset for sensor-aided visual handheld camera localization. In *Proceedings of ISMAR*, pages 263–264, 2013.

[5] D. Kurz et al. Absolute spatial context-aware visual feature descriptors for outdoor handheld camera localization. In *Proceedings VISAPP*, pages 36–42, 2014.

[6] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60(2):91–110, 2004.

[7] J. Ventura and T. Höllerer. Wide-area scene mapping for mobile visual tracking. In *Proceedings of ISMAR*, pages 3–12, 2012.