The Advantages of a Joint Direct and Indirect VSLAM in AR

Georges Younes
University of Waterloo
Waterloo, Canada
gyounes@uwaterloo.ca

Daniel Asmar
American University of Beirut
Beirut, Lebanon
da20@aub.edu.lb

John Zelek
University of Waterloo
Waterloo, Canada
jzelek@uwaterloo.ca

Abstract

At its core, Augmented Reality (AR) localizes a user's pose in a scene and accordingly augments the user's view with virtual objects.
While many methods for user tracking and scene understanding were proposed, Visual Simultaneous Localization And Mapping (VSLAM) prevailed as the back-end for a realistic and unrestricted AR experience.
In this paper, we extend our previous work on joint Direct and Indirect VSLAM to perform AR. The advantages of having both representations concurrently for AR are thoroughly discussed, resulting in an improved pose tracking accuracy and robustness, a more detailed scene representation while gaining the ability to handle lighting variation through auto-exposure in real-time. We also argue that our VLSAM system can be run with only a single camera, thus simplifying the hardware necessary.

1. Introduction

Augmented Reality (AR) is a process that injects virtual components onto a live view of the real world in real-time. The main objective is to provide the user with an enriched and informative experience of their surroundings. Due to its wide range of applicability in multidisciplinary domains such as tourism, education, gaming, military, and advertisement, etc., AR emerged over the past years from research labs to become a multi-billion dollar industry.
However, despite its wide adoptability, many software and hardware related challenges still limit the user experience and remain subject to active research. In particular, to achieve a plausible augmentation, AR applications must accurately and actively track a user point of view and their surrounding environment to render the virtual parts from the correct perspective.
To that end, a multitude of methods were developed over decades, with early solutions suggesting specialized infrastructure anchors in the scene that track distinct light-or-infrared emitting beacons placed on users. Other methods were also developed using Inertial Measurement Units (IMU) and magnetic sensors to track the user's pose. However, computer vision methods evolved to replace the aforementioned methods as cheaper, more accurate and infrastructure-less alternatives for pose tracking. The interested reader is referred to [2] for a historical overview and detailed survey on AR.
Early vision based solutions to pose tracking, referred to as Marker-Based, involved adding known markers in the scene, such that when observed, provide a spatial datum from which the observer point of view is determined and then used as a reference to insert the virtual components. However, aside from the inconvenience of physically modifying the scene, Marker-Based methods require the marker to be always visible and fail when it goes out of view. Marker-less approaches address these limitations by extracting artificial markers (referred to as features) from the camera feed and uses them to recover both the camera pose and the scene structure concurrently; no markers are physically placed in the scene and the map of features can be expanded on the go. The core technology behind this ability is known as Visual Simultaneous Localization and Mapping (VSLAM) where accuracy and robustness are key in achieving a realistic AR experience. While VSLAM can be performed using various vision based sensors (*e.g.* RGB-D sensors, stereo cameras, etc.), single cameras are ubiquitous, cheap, have a low weight, and are passive sensors *i.e.* they do not project rays into the scene, making them power efficient and usable both indoors and outdoors. For these reasons, our proposed system employs a monocular camera; nevertheless, the ideas put forward generalize to both RGB-D and stereo cameras.
In this work, we propose an extension to our previously proposed hybrid monocular VSLAM system [15] dubbed UFVO (Unified Formulation for Visual Odometry) to perform AR. In particular, we leverage the presence of various types of features in UFVO to achieve high accuracy, increased robustness to texture-deprived environments, motion blur, and illumination variations, while having control over the density of the map.

2. Background

Over the past two decades, the research community has put forward an ample amount of VSLAM solutions that can be categorized as either Direct, Indirect or Hybrid, with each extracting different types of features and exhibiting different but often complementary properties.

2.1. Direct vs. Indirect

Both Direct and Indirect methods extract features from images and associate them with descriptors. Direct methods sample pixels with a relatively large intensity gradient and associate them with a patch of pixels surrounding their sampled location. In contrast, Indirect methods sample corners and associate them with higher dimensional descriptors. The different types of extracted features require different objective functions to minimize, with Direct methods resorting to photometric residuals (pixel intensity difference) and Indirect methods resorting to geometric residuals (pixel distance between matched features). The photometric objective function assumes constant brightness from one picture to another and thus is brittle under illumination changes and can only handle small inter-frame motion. In contrast, the geometric objective function assumes that a feature can be unambiguously matched across images and as such is brittle in corner-deprived environments caused by texture deprived regions or motion blur. Indirect methods also suffer from reduced accuracy as their resolution is limited to that of the pixel location from which a corner was extracted, whereas Direct methods can integrate over the image domain to achieve subpixel accuracy. DSO [3] is currently considered state of the art in Direct methods whereas ORB SLAM 2 [9] is currently considered state of the art in Indirect methods.

2.2. Hybrid Methods

Hybrid methods were introduced to leverage the advantages of both Direct and Indirect methods. However, most proposed systems fall short from diminishing all of their shortcomings. For example [8] run both DSO and ORB SLAM 2 sequentially with the camera pose and local map estimated from DSO and a global map built from ORB SLAM 2. However such implementation is subpar in both efficiency and accuracy. On the other hand [6] extracts corners with ORB descriptors ([12]) to perform loop closure but does not use them for pose estimation. Finally, [14] estimate the camera pose using DSO and triggers an Indirect based failure recovery if the Direct tracking fails, however it may not always detect tracking failure.

3. Proposed approach

3.1. Unified Formulation for Visual Odometry (UFVO)

UFVO address all limitations of previous hybrid methods by homogeneously distributing both types of features (Direct and Indirect) across an image, and by concurrently associating two types of descriptors with Indirect features (as shown in figure 1).

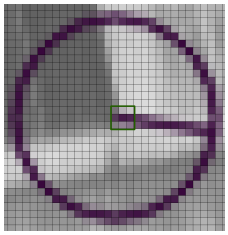

Figure 1. Patch of pixels descriptor(encapsulated by the green square) and ORB descriptor with orientation (encapsulated by the mauve circle) associated with the same Indirect corner.

3.1.1 Pose Tracking

Both Direct and Indirect features are used to compute the camera pose in a joint optimization using:

$$\underset{T}{argmin}\ e_{pho}(T, X_d, X_i) + \lambda e_{geo}(T, X_i), \qquad (1)$$

where T $\in Se(3)$ is the camera pose relating the current frame to the last added keyframe, X_d and X_i are the set of Direct and Indirect map points respectively, and λ is the relative weight between the photometric and geometric residuals. λ is controlled by a logistic utility function defined in [15], which starts by assigning a large weight λ to the geometric residuals and subsequently reduces it as the optimization progress such that the Direct residuals eventually dominate. This allows UFVO to handle large inter-frame motions while benefiting from the sub-pixel accuracy of Direct methods. Furthermore, the logistic utility function keeps track of the number of inlier geometric residuals within the optimization and accordingly reduces λ in feature deprived environments, allowing UFVO to handle textureless regions and motion blur using the photometric residuals only. The entire pose estimation process takes 14ms per frame on a core i7-8700 Intel CPU. In contrast, DSO requires 6 ms and ORB SLAM 2 requires 23 ms.
The ability of UFVO to handle large inter-frame motions is tested in [14] where the feature assisted Direct pose estimation was able to handle motions as large as that of state of the art in Indirect methods. On the other hand, UFVO's accuracy is tested in [15] where it outperformed state of the art in Direct, Indirect and other Hybrid systems, over an exhaustive set of sequences covering challenging indoor and outdoor environments.

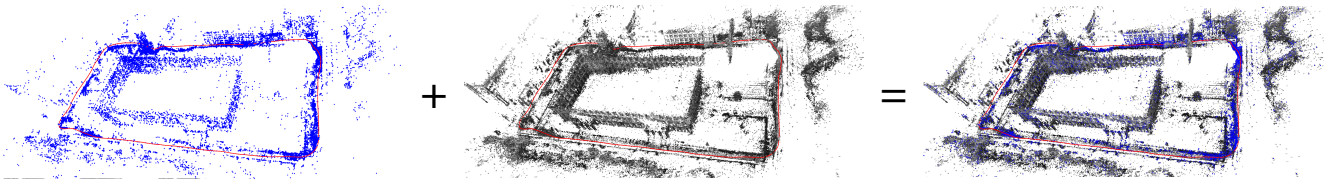

Figure 2. The Indirect features are shown in the left image (in blue), the Direct features are shown in the middle image (in black) and the Joint map representation from UFVO is shown in the right image.

3.1.2 Mapping

Unlike typical Indirect VSLAM in the literature such as [9] and [7], etc., UFVO adopts an inverse depth parametrization for both Direct and Indirect features and uses the photometric descriptors to triangulate them in a filter formulation. Aside from its ability to triangulate both representations at virtually no extra computational cost than that of DSO, the filter based triangulation avoids the numerical instability typically associated with n-view triangulation methods when the views are separated by small baselines. The end result is a single map with both representations as shown in figure 2.

3.2. Global Map and Density Control

A faithful representation of the environment allows for realistic interactions between the real world and the augmented parts by integrating physics such as collision and occlusion detection. Accordingly, the reconstructed map using VSLAM is usually processed to generate a dense scene representation by either meshing the VSLAM map into a continuous surface [10] and/or voxelizing it into a 3D grid [11], in which the augmented interactions take place. The accuracy and density of the VSLAM map are then essential for exact and detailed scene representation.

On one hand, Indirect methods are limited to sparse map representations since they rely on the saliency of extracted features to establish matches from one image to another, which decreases as the extracted features get closer to each other. On the other hand, Direct methods do not require this restriction and can sample features from any pixel with an intensity gradient in an image. Therefore Direct methods density can be varied from sparse to semi-dense.

On another note, keeping track of a global Direct map is memory inefficient and computationally exhaustive; typical Direct methods are either limited to odometry [3] which keeps track of a small subset of the map only, or rely on Indirect feature encoding to query the global map [4].

The presence of both Direct and Indirect features in UFVO provides a semi-dense scene representation (using the Direct map points), essential for an accurate dense reconstruction and subsequently resulting in realistic physics interactions. Whereas the user pose is estimated using the Indirect formulation globally and a joint Direct and Indirect features locally. The global map here allows for pose tracking in AR applications where a scene was mapped beforehand similar to the work presented in [13].

3.3. Robustness to Illumination Variation

During regular operation, a camera is allowed to adjust its exposure to accommodate varying lighting conditions. However, for the sake of brightness constancy, auto exposure is typically turned off in an AR session so that the camera does not lose track of the map. In an effort to increase the robustness of Direct VSLAM to varying lighting conditions, [5] proposed an offline photometric calibration process to estimate the camera response function and vignetting map. If a photometric calibration for a camera is then available along with the exposure times per frame, they can be used to map the image intensities to the scene irradiance, which is independent of the camera response, thereby exploiting auto-exposure to maintain the brightness constancy assumption. However, obtaining a photometric calibration and exposure times is a daunting task, nevertheless [1] showed that if features can be reliably matched between photometrically distorted images in a video sequence, they can be used to recover both the photometric calibration and the exposure time per frame in real-time.

The required feature matches are inherently available in UFVO through its Indirect features that use ORB descriptors to establish correspondences across frames. Therefore, our proposed joint formulation can be exploited to perform real-time photometric calibration and allow for auto-exposure in an AR session using the already available feature matches and the optimization described in [1].

4. Conclusion

A joint Direct and Indirect formulation for VSLAM offer many advantages for AR; from improved tracking accuracy and robustness to better detailed and more precise scene representation, the end result is an overall improved and more realistic AR experience.

While this paper has analyzed the ways in which such improvements can be achieved, we are currently developing our joint formulation to capitalize on said advantages and achieve a realistic AR experience in challenging outdoors environments.

References

[1] P. Bergmann, R. Wang, and D. Cremers. Online photometric calibration of auto exposure video for realtime visual odometry and slam. *IEEE Robotics and Automation Letters (RA-L)*, 3:627–634, April 2018.

[2] Mark Billinghurst, Adrian Clark, and Gun Lee. A survey of augmented reality. *Found. Trends Hum.-Comput. Interact.*, 8(2-3):73–272, Mar. 2015.

[3] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):611–625, March 2018.

[4] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 834–849, Cham, 2014. Springer International Publishing.

[5] J. Engel, V. Usenko, and D. Cremers. A photometrically calibrated benchmark for monocular visual odometry. In *arXiv:1607.02555*, July 2016.

[6] Xiang Gao, Ying Wang, Nikolaus Demmel, and Daniel Cremers. Ldso: Direct sparse odometry with loop closure. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2198–2204, 2018.

[7] Georg Klein and David Murray. Parallel tracking and mapping for small AR workspaces. In *Proc. Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07)*, Nara, Japan, November 2007.

[8] Seong Hun Lee and Javier Civera. Loosely-coupled semi-direct monocular SLAM. *CoRR*, abs/1807.10073, 2018.

[9] Montiel J. M. M. Mur-Artal, Raúl and Juan D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[10] R. A. Newcombe and A. J. Davison. Live dense reconstruction with a single moving camera. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1498–1505, June 2010.

[11] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, Oct 2011.

[12] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571, Nov 2011.

[13] Georges Younes, Daniel Asmar, Imad Elhajj, and Howayda Al-Harithy. Pose tracking for augmented reality applications in outdoor archaeological sites. *Journal of Electronic Imaging*, 26(1):1 – 12 – 12, 2016.

[14] Georges Younes, Daniel Asmar, and John Zelek. Fdmo: Feature assisted direct monocular odometry. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, pages 737–747. INSTICC, SciTePress, 2019.

[15] Georges Younes, Daniel C. Asmar, and John S. Zelek. A unified formulation for visual odometry. *CoRR*, abs/1903.04253, 2019.