

論文の内容の要旨

論文題目 3D Reconstruction of Scattered Objects within RGB-D Images
(RGB-D 画像内に点在する物体の三次元再構成)

氏 名 サホロール ハムディ モ ハ メ ド サレフ
Sahloul, Hamdi Mohammed Saleh

Constructing 3D structures of real-world objects is a scientific problem for a wide variety of fields, including computer vision and robotics. In recent decades, hardware utilized in reconstruction, such as depth sensors and TOF (time of flight) cameras, have become widely affordable, enabling reconstruction to gain huge attention. Nowadays, 3D reconstruction is in active research to achieve goals ranging from aiding computer graphics, virtual and augmented reality to surveillance, robotic navigation, manipulation and objects recognition.

Various and numerous reconstruction applications in home and office environments as well as industrial environments have motivated many researchers to introduce different reconstruction methods. Briefly, there are three types of reconstruction methods: MVS (multi-view stereo), SfM (structure from motion), and dense VSLAM (visual SLAM (simultaneous localization and mapping)). 3D reconstruction from 2D images is an old problem, which have been tackled by various algorithms such as SfM and MVS. More importantly, after the development of RGB-D sensors in the last decade, VSLAM methods have emerged and made great momentum in the 3D reconstruction research society. Notably, KinectFusion [1] is one of the early VSLAM that have achieved real-time 3D construction using RGB-D sensor data. However, real-time 3D construction methods usually depends on marginal camera assumptions, and thus fail to reconstruct from RGB-D images of different view. For such

reasons, researchers are now revisiting MVS and SfM techniques and augmenting them with the new RGB-D images in order to achieve robustness and model completeness. MVS utilizes several extrinsically calibrated cameras system, while SfM depends on motion cues within its input 2D or 2.5D images by estimating their correspondences.

Importantly, the objective is to reconstruct accurate models covering all objects' surfaces and to segment them from their environments within reasonable time. An input of several RGB-D images is considered, where each RGB-D image captures a scene of various same- or different-type objects from numerous viewpoints. However, MVS and VSLAM are unable to cope with such input due to their camera motion constraints. Instead, if each input RGB-D image is segmented according to the captured objects and viewpoints, such that each segment is considered as an independent RGB-D image of its own, then a virtual camera motion between these segments can be assumed, and the problem can be reduced into a SfM problem. Although SfM does not have a camera motion limitation, it suffers from a time-accuracy trade off. More importantly, all reconstruction methods cannot segment an object model from its environment without texture, geometry, or motion assumptions. Yet, the discussions of both the time-accuracy trade off and the segmentation issue are deferred until more intuition is developed. Indeed, 3D reconstruction problem embarrasses a great deal of techniques, and for simplicity purposes, its study is divided into three categories based on their correspondence complexity: 1-1 (one-to-one), 1-M (one-to-many), and M-M (many-to-many). In 1-1 or model-view correspondence, one input, i.e. an RGB-D image, captures a single object from a viewpoint while another captures the same physical object yet from a different viewpoint, where the problem involves finding which parts of one viewpoint represent the same physical point in the other viewpoint, i.e. correct or inlier correspondences, in order to register the inputs together into a 3D model. On the other hand, if one input involves several different objects, then the correspondence problem is a 1-M or model-scene, which is more involved than the 1-1 relation, due to the additional need to localize the correct object. If several objects are observed in each input, then the correspondence problem is a M-M or scene-scene, in which there is a need to segment the scenes in order to reduce the problem to several 1-M ones. Importantly, although two inputs are employed in the above demonstration, the correspondence problem can involve arbitrarily larger number of inputs.

Notwithstanding, each of these correspondence problems has its own challenges. Firstly, the success in solving 1-1 correspondence problems depends heavily on the input viewpoints, for which if two inputs have been taken from very different angles, i.e. relatively-large viewpoint difference, the common surfaces captured within these inputs drastically diminishes, rendering their matching unattainable. Even in the moderately overlapping-surface cases, algorithms desperately struggle to detect the same physical point in such inputs due to perspective changes. To explain this limitation, there is a need to introduce some algorithmic aspects first. A correspondence between two inputs is usually estimated by detecting some interesting or outstanding points, i.e. features, from both inputs

and matching them using their intrinsic descriptions, e.g. color gradient or neighborhood curvature. However, feature detection and description algorithms lack enough repeatability and distinctiveness under large viewpoint difference due to the inconsistencies observed around the same physical point within its observed very-different perspectives. For this, in the first contribution, a wrapper is proposed around existing algorithms, i.e. wrapper, for improving the invariance of feature-based algorithms against viewpoint difference. The idea is to rotate and translate surfaces to an invariant perspective, thus improving the feature detection and description. Another important benefit from improving an algorithm's maximal viewpoint difference, i.e. the viewpoint invariance, is to decrease the number of required frames for covering the whole object, thus the consumed time on capture and reconstruction, thus addresses the previously deferred time-accuracy trade off. Secondly, as for the 1-M correspondences, even with highly repeatable and distinctive algorithms for feature detection and description, incorrect correspondences, i.e. outliers, are mostly contaminating the correspondences set due to noise, similar surfaces, occlusions ...etc. Filtering such contaminated, i.e. putative, correspondences by rejecting the outliers is a very distinct and well-known problem. Accordingly, in the second contribution, a rapid voting-based pattern recognition scheme is proposed to rank the single-structure putative correspondences according to their likelihood being inliers, thus enabling application-specific outlier rejection, which contributes in addressing the time-accuracy trade off as well. Thirdly, the segmentation problem embedded with the M-M correspondence problem gets more intractable due to the scarce information on the spatial shapes of each object, where the aim is to actually solve such problem for 3D reconstruction to begin with. The multi-structure hypotheses modeling approach in the third contribution extends that of the previous contribution, to enable modeling the motion of several objects within their environment, even under high outliers rate.

As for results, firstly, the detectors and descriptors wrapper improved their viewpoint invariance to different levels, depending on the scene geometric curvatures, while unwrapped existing, i.e. 2D, feature algorithms fail to accommodate average viewpoint difference beyond 33.3° . Objects with distinct surface discontinuities were on average matchable up to 52.8° , and the overall average for all evaluated datasets was 45.4° . Similarly, out of 140 combinations involving 20 feature algorithms and various objects with distinct surface discontinuities, only a single 2D feature algorithm exceeded the goal of 60° viewpoint difference in just two combinations, as compared with 19 different feature algorithms succeeding in 73 combinations when wrapped in the proposed wrapper. Furthermore, the proposed approach operates robustly in the presence of input depth noise, even that of low-cost commodity depth sensors, and well beyond. Secondly, the proposed correspondences voting scheme scored $97.0\% \pm 12.9\%$ on the PR (precision-recall)-AUC (area under curve) metric measuring the correctly scored correspondences in average of all the experiments, while the two state-of-the-art schemes scored $74.2\% \pm 22.2\%$ and $78.3\% \pm 26.4\%$ respectively. Furthermore, the proposed scheme did not require more than $41.5\% \pm 12.5\%$ of the time consumed by the fastest state-of-the-art scheme.

Likewise, the proposed voting scheme also demonstrated high robustness against occlusions and scarce inliers. Thirdly, the multi-structure hypothesis generation method outperformed famous multi-structure hypothesis generation methods in hypotheses precision, points recall, and speed aspects. Quantitatively, with 750 generated hypotheses and inliers rate as low as 3.5% on a total of 50 experiments, the proposed method scored $69.6\% \pm 11.5\%$ precision and $69.1\% \pm 5.4\%$ recall, while the remaining methods scored no higher than 12% on both metrics. Furthermore, the proposed method did not require more than 6.45ms per each hypotheses, which is about 1.16% of the time required by a sophisticated method, and about 268% of the time required by random hypothesis generation, the most simple method. Importantly, the balance between precision and recall, as well as the execution time is totally controllable in the proposed method, as it performs in a progressive manner, with the first generated hypotheses having high precision probability. By limiting the generated hypotheses to 200 in the above mentioned experiments, the execution time reduces to one fourth, while the precision and recall of proposed method becomes $93.9\% \pm 4.2\%$ and $32.4\% \pm 8.0\%$, respectively, in compare to $6.6\% \pm 2.1\%$ and $3.8\% \pm 1.0\%$ for the best compared method.