AUTOMATIC VIDEO TO POINT CLOUD REGISTRATION IN A STRUCTURE-FROM-MOTION FRAMEWORK

Esteban Vidal , Nicola Piotto , Giovanni Cordara , Francisco Morán Burgos

ABSTRACT

In Structure-from-Motion (SfM) applications, the capability of integrating new visual information into existing 3D models is an important need. In particular, video streams could bring significant advantages, since they provide dense and redundant information, even if normally only relative to a limited portion of the scene. In this work we propose a fast technique to reliably integrate local but dense information from videos into existing global but sparse 3D models. We show how to extract from the video data local 3D information that can be easily processed allowing incremental growing, refinement, and update of the existing 3D models. The proposed technique has been tested against two state-of-the-art SfM algorithms, showing significant improvements in terms of computational time and final point cloud density.

Index Terms— 3D Reconstruction, SfM, Video Registration, Point Cloud Alignment

1. INTRODUCTION

Structure-from-Motion (SfM) algorithms applied to large unordered image collections have been proven to successfully recover the 3D model of a scene, as well as the camera locations [1] [2] [3] [4]. At the core of SfM frameworks, Bundle Adjustment (BA) [5] is usually adopted as the optimization step for the non-linear joint refinement of camera and point parameters. Unfortunately, BA can consume a significant amount of time as the number of images involved in the optimization grows. Although many strategies have been proposed to speed it up [6] [7] [8], time complexity is still a problem.

Another issue of most of the current SfM engines is related to the limited capability in integrating new information into pre-existing models. One pioneering work trying to solve this issue is reported in [1], a nd further improved in [9]. Here, the authors propose a way to estimate the new camera locations with respect to the existing point cloud, but without adding new 3D information to the model. Another interesting work is presented in [4], but it also does not provide 3D model refinement capability.

In light of these limitations, we propose a novel SfM framework, allowing a fast processing and integration of new visual information into existing 3D models. Moreover, instead of single images, we consider full rate video sequences, enabling 3D model refinement and update to be performed. In fact, due to their redundancy and high sampling rate, input from video frames can considerably increase the density of a portion of sparse point cloud. Our proposal focuses on the creation of local models out of the input video sequences. This stage can be carried out in parallel for multiple streams. Once the local models are computed, they are aligned to the base model through feature matching and robust pose estimation.

This paper presents our effort in improving the time complexity and extending the operational scenario of SfM systems by the following contributions:

- online video processing and local 3D information recovery;
- a novel approach to automatically recover the ratio between two point clouds of correspondences with unknown and arbitrary scale;
- effective refinement of the existing 3D model through integration of 3D information from different point clouds.

The effectiveness of the proposed framework has been validated on a real-world dataset, comprising both images and video streams. The obtained results show how our pipeline allows a fast 3D point cloud extension while providing a denser merged model.

2. RELATED WORK

Automatic registration of video streams into sparse point cloud aims at recovering the six-Degree-of-Freedom (6-DoF) pose of each video frame of a moving camera with respect to a given 3D model.

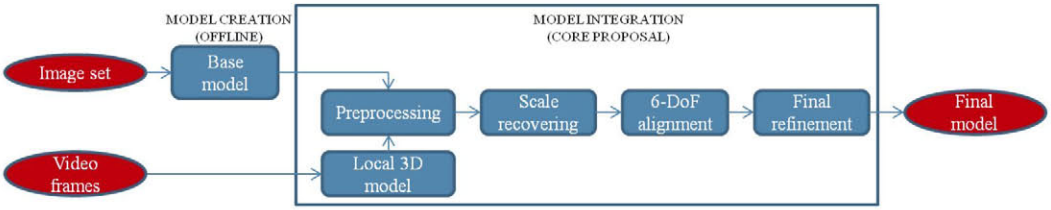

Fig. 1. Flowchart of the proposed solution.

In general, when 3D information is available, registration techniques based on Iterative Closest Point (ICP) [10] can be applied [11] [12]. One of the most representative works is reported in [11], where a technique for aligning video streams onto 3D sensor data is proposed. Motion stereo and SfM are used to create the 3D models from continuous video while a modification of the ICP algorithm is used to align the point cloud directly with the 3D sensor data. One of the advantages of ICP is that it does not require explicit feature extraction and matching. However, a good initial guess must be provided to ensure its convergence [13]. Although providing good results, ICP-based registration methods suffer from slow convergence rate and significant computational complexity: $O(N_m N_p)$ for point clouds of sizes M and P.

When 3D information is not available, 2D-to-3D registration methods can be applied. Two main strategies have been proposed, namely *direct matching* [9] [14], and *image retrieval* [15]. In [9] a location recognition framework is presented where an image can be localized within a point cloud. Since every sample of the model is seen by several images, an average 2D descriptor is associated to each 3D point. When a query image is given, features are extracted and matched against the model average descriptors, providing a set of 2D-to-3D correspondences that are fed to a pose estimation algorithm. In [15] a dual process of [9] is proposed, a minimal set of virtual images that fully covers the point cloud is generated, and when a new query image is given, a vocabulary tree [16] retrieval approach is used to retrieve the most similar virtual images. Direct methods achieve a better localization performance than retrieval-based approaches as they are able to localize more query images [14]. However, this performance gain comes at the cost of memory consumption, since direct methods require to store the descriptors of the 3D points in memory.

When dealing with the registration of 3D point sets with large scale difference, the scale factor can be recovered independently [17], or obtained as part of a global optimization [18] [19] problem. In [17] the authors propose to characterize the scale of a given point cloud by a set of cumulative contribution rate curves obtained by performing principal component analysis on spin images. A variant of ICP is then used to register the curves of two point clouds and recover the scale. In [18] the authors propose to decompose the registration estimation in a sequence simpler steps: first, two rotation angles are determined by finding dominant surface normals, then the remaining parameters are found with RANSAC [20] followed by ICP and scale refinement. In [19] for automatic registration of images on approximate geometry is presented. Here, the point cloud is first roughly registered to the 3D object using a variant of the four points congruent sets algorithm. A global refinement algorithm based on mutual information is then applied to optimize the color projection of the aligned photos on the 3D object.

3. VIDEO-TO-POINT-CLOUD ALIGNMENT

An overview of the proposed framework is shown in Fig. 1. In the offline stage, the image set is fed into a SfM engine which provides the scene base model. Each sample of the point cloud is automatically augmented with an average 2D descriptor, in the same spirit of [9]. In the same way, the video frames are processed, and a number of local models are recovered from each sequence. The goal is thus to register and merge the base model with all the video local models. Relying on the average descriptors, standard feature matching techniques can be applied in order to find corresponding points. Given the correspondences, the scale factor between the point clouds is recovered. When the models are in the same scale, a RANSAC-based routine is run to find the rigid 6-DoF transformation minimizing the 3D distance between the points. The roto-translation matrix is then applied to the model points and cameras to align them. In a final step, duplicate points/cameras are pruned, and a BA is run to further refine the parameters.

3.1. Preprocessing

The preprocessing phase is introduced to find corresponding samples in the models to merge. To ease the point cloud registration, we follow the procedure proposed in [9]. In particular, we rely on the average descriptor attached to each model point to translate the problem of 3D-to-3D point matching to a more tractable 2D feature matching one, which can be solved with state-of-the art methods. Let $P_m = \{pm_i\}$ and $P_v = \{pv_j\}$ be respectively the base model cloud (of M points: $0 \leq i < M$) and the local model cloud extracted from a particular video sequence (of N points: $0 \leq j < N$). For each model point pm_i, $pv_j \in \mathbb{R}^3$ a 128-byte SIFT [21] average descriptor is

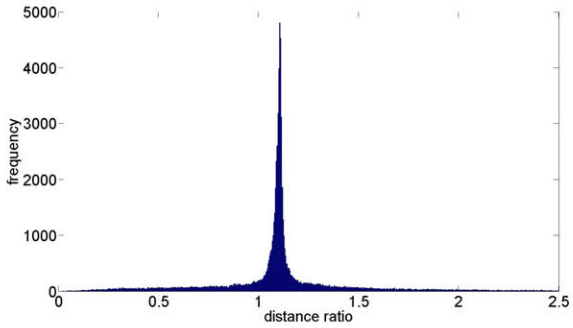

Fig. 2. Histogram of a vector of distance ratios $\vec{r}$.

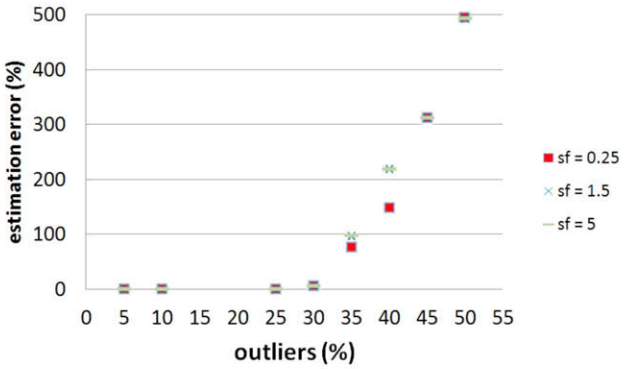

Fig. 3. Performance of the scale recovery strategy in terms of estimation error vs outlier percentage.

computed. Standard feature matching allows to find corresponding 3D points between P_m and P_v. These matching 3D points form two new clouds of correspondences $C_m = \{cm_k\}$ and $C_v = \{cv_k\}$ with $k = 0...K$.

3.2. Model integration: scale recovery

We recover the scale factor between the models adapting the distance ratio analysis of matching keypoints introduced in [22], and extended to the 3D case in [23]. In [22], in fact, in order to detect outliers after the keypoint matching Log Distance Ratio (LDR) of couple of matches is computed as $ldr(cm_i, cm_j, cv_i, cv_j) = \ln\left(\frac{||cm_i - cm_j||}{||cv_i - cv_j||}\right)$ with $cm_i \neq cm_j$ and $cv_i \neq cv_j$. We adapted the same concept to the scale factor recovery, dropping the ln operation, thus simply computing the distance ratio (DR). Assuming an error-free matching phase, the DR of a random pair of matches directly gives the scale between the point clouds. However, the presence of outlier matches forces a statistical analysis of the DR for all the potential pairs. To this aim, the DR is computed for each pair of matching points in C_m and C_v, producing a vector of DRs $\vec{r}$. An example of DR computed on real point cloud matching pairs is illustrated in Fig. 2. Finding the correct scale ratio between the clouds corresponds to locate the DR of the inlier matches. To this aim, the median DR value $M_{\vec{r}}$ is computed and the interval $inl_{\vec{r}} = [M_{\vec{r}} - MAD_{\vec{r}}, M_{\vec{r}} + MAD_{\vec{r}}]$ is considered, where $MAD_{\vec{r}} = median_i(|r_i - M_{\vec{r}}|)$ is the median absolute deviation. The assumption is that the DRs corresponding to the inlier matches are constrained in a small range, which resembles the real scale between the clouds. The final scale ratio is obtained by averaging the DRs in the $inl_{\vec{r}}$ range.

Fig. 3 shows the performance of the proposed scale recovery strategy: a reference cloud with 5000 points is considered, with varying scale factor (*sf*), and outlier matches percentage. As can be seen, the scale recovery is very accurate for reasonable outlier percentage below 25%.

3.3. Model integration: 6-DoF alignment and refinement

Once the scale ratio is recovered, a cloud P'_v scaled version of P_v is obtained in the same coordinate frame as P_m. The goal is now to estimate the 6-DoF roto-translation to align P'_v and P_m. Due to the presence of outlier matches, singular value decomposition is not an option to estimate that transformation; instead, a RANSAC-based routine has been chosen allowing a robust estimation. A new registered point cloud P''_v is obtained by applying the candidate transformation to P'_v. Finally, the base model P_m and the registered cloud can be merged into the final model F. Due to small alignment errors, some refinements are needed in the overlapping areas. In fact, while significant information is added to the final cloud F (i.e. all non-matching samples), duplicate points do appear. In order to avoid such duplicate points and improve matching robustness, matching points are relocated by averaging their local positions, descriptors, and color information.

In order to further consolidate the merged model, a final BA is applied to F, to refine the positions of cameras and 3D points, obtaining the final refined cloud F', which can be used as base model for future registrations.

4. RESULTS

In order to quantitatively assess the performance of our solution, we carried out a number of experiments comparing the time complexity and the final model density of our proposal with those of two state-of-the art SfM solutions. Due to their popularity, Bundler [1] and VisualSfM (VSfM) [4] have been selected as reference, however, the presented video integration method can be applied to any other SfM framework.

We collected a dataset of images and videos of the Rathaus building in Marienplatz, Munich. The set includes 492 pictures harvested from the web and 7 video streams acquired with a mobile phone (Huawei P6). The image set is used to create the base model, while the video streams serve as input for the model update.

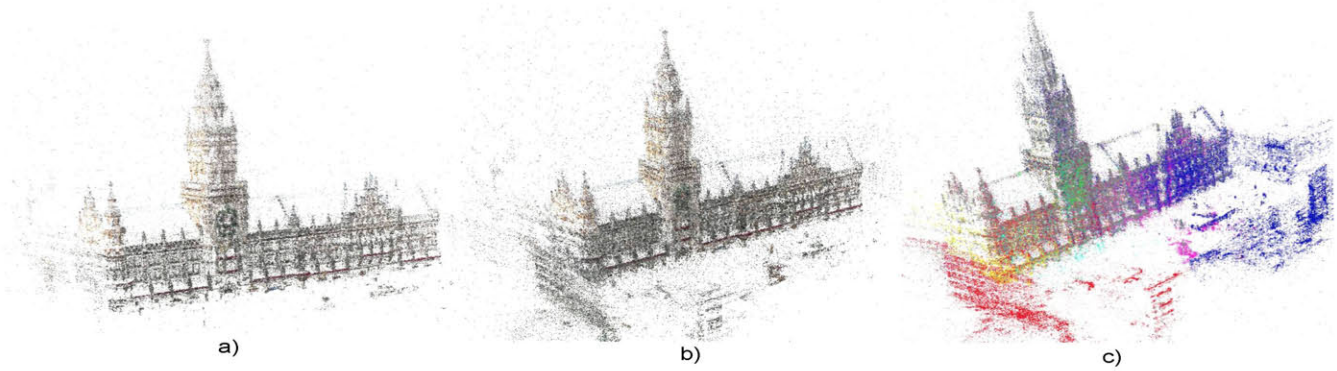

Fig. 4. a) Base point cloud model, and b) Bundler's final model vs. c) ours, in which colors map different videos.

	Bundler				VSfM			
Dataset (# images)	Method	Total points	Time (s)	Time gain	Method	Total points	Time (s)	Time gain
(A) Base model (492) + *facadeL* (66)	Ours	248032	6935 (130)	30.08%	Ours	116099	2186 (18)	22.01%
	Bundler	180646	9918		VSfM	110111	2803	
(B) A (558) + *facadeR* (64)	Ours	263115	7392 (126)	31.84%	Ours	128881	2354 (13)	10.90%
	Bundler	253718	10845		VSfM	121294	2642	
(C) B (622) + *glock* (73)	Ours	280334	7917 (127)	23.99%	Ours	139836	2632 (8)	34.07%
	Bundler	271005	10416		VSfM	131497	3992	
(D) C (695) + *tower* (49)	Ours	292886	8204 (127)	22.77%	Ours	148772	2700 (8)	32.74%
	Bundler	274119	10623		VSfM	141412	4014	
(E) D (744) + *clock* (56)	Ours	298925	8563 (134)	30.08%	Ours	153263	3016 (24)	20.23%
	Bundler	252887	12246		VSfM	143430	3781	
(F) E (800) + *dragon* (131)	Ours	327910	9622 (212)	38.89%	Ours	165276	3541 (11)	29.34%
	Bundler	305114	15745		VSfM	156636	5011	
(G) F (931) + *marien* (319)	Ours	372209	12285 (267)	65.94%	Ours	191871	5369 (23)	24.62%
	Bundler	336951	36068		VSfM	179443	7123	

Table 1. Computational efficiency of our technique vs. both Bundler and VSfM. All the experiments have been run on a machine equipped with 2 x 2.7GHz Xeon CPUs, 8 cores/CPU and 128GB of RAM.

The comparison against the selected baselines [1] [4] is carried out considering the base model as the reference to be incrementally updated with the local model information. Since neither Bundler nor VSfM can add information to the model, they run the entire SfM pipeline with an increasingly enlarged set of input images. The simulations show how our proposal allows an efficient processing of the local model and point cloud registration, significantly reducing the computation time while preserving the model accuracy and increasing point density.

In Fig. 4, the base model, and the models obtained by Bundler and by our approach are rendered side by side for subjective comparison. Table 1 reports relevant details: the size of the image database, the total number of points of the final model, the computation time required by the different approaches and the time gain using our method. The time in brackets is the one required for the alignment of the two point clouds, assuming they are pre-calculated. As can be seen, our proposal significantly reduces the time, although we have added the time required to build the base and local models to the one spent for the point cloud registration. However, in a more realistic scenario, where the base model can be assumed to be available, and a video sequence is provided for the update, the effective computation load would be only function of the video length, and the registration time. The time gain in this more realistic scenario is 4-60X faster than the selected baselines (75-98% gain), depending on the length of the video stream and the number of correspondences.

5. CONCLUSIONS

We have proposed a method to integrate information from video sequences into reference 3D point clouds. First, a global, sparse, base model is built thanks to a SfM engine. Then, the videos are used to generate several local, denser 3D models. The scale between local and base model is recovered through statistical analysis of the distance ratio of matching samples, and then the local model is registered to the base one exploiting the local information of the images. The testing phase has focused on comparing the time complexity of the proposed approach against two state-of-the-art SfM methods [1] [4], showing the effectiveness of our solution.

6. REFERENCES

[1] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3D," *ACM transactions on graphics (TOG)*, vol. 25, no. 3, pp. 835–846, 2006.

[2] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski, "Building rome in a day," *Communications of the ACM*, vol. 54, no. 10, pp. 105–112, 2011.

[3] M. Farenzena, A. Fusiello, and R. Gherardi, "Structure-and-motion pipeline on a hierarchical cluster tree," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1489–1496.

[4] C. Wu, "Towards linear-time incremental structure from motion," in *3DTV-Conference, 2013 International Conference on*. IEEE, 2013, pp. 127–134.

[5] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment - a modern synthesis," in *Vision algorithms: theory and practice*, pp. 298–372. Springer, 2000.

[6] S. Agarwal, N. Snavely, S. M. Seitz, and R. Szeliski, "Bundle adjustment in the large," in *Computer Vision–ECCV 2010*, pp. 29–42. Springer, 2010.

[7] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz, "Multicore bundle adjustment," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3057–3064.

[8] D. J. Crandall, A. Owens, N. Snavely, and D. P. Huttenlocher, "Sfm with mrfs: Discrete-continuous optimization for large-scale structure from motion," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 12, pp. 2841–2853, 2013.

[9] Y. Li, N. Snavely, and D. P. Huttenlocher, "Location recognition using prioritized feature matching," in *Computer Vision–ECCV 2010*, pp. 791–804. Springer, 2010.

[10] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Robotics-DL tentative*. International Society for Optics and Photonics, 1992, pp. 586–606.

[11] W. Zhao, D. Nister, and S. Hsu, "Alignment of continuous video onto 3D point clouds," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1305–1318, 2005.

[12] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz, "Aligning point cloud views using persistent feature histograms," in *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*. IEEE, 2008, pp. 3384–3391.

[13] R. B. Rusu, Z. C. Marton, N. Blodow, and M. Beetz, "Persistent point feature histograms for 3D point clouds," in *Proc 10th Int Conf Intel Autonomous Syst (IAS-10), Baden-Baden, Germany*, 2008, pp. 119–128.

[14] T. Sattler, B. Leibe, and L. Kobbelt, "Fast image-based localization using direct 2D-to-3D matching," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 667–674.

[15] A. Irschara, C. Zach, J-M. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2599–2606.

[16] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 2, pp. 2161–2168.

[17] B. Lin, T. Tamaki, B. Raytchev, K. Kaneda, and K. Ichii, "Scale ratio icp for 3D point clouds with different scales," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, 2013, pp. 2217–2221.

[18] D. Novak and K. Schindler, "Approximate registration of point clouds with large scale differences," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 1, no. 2, pp. 211–216, 2013.

[19] M. Corsini, M. Dellepiane, F. Ganovelli, R. Gherardi, A. Fusiello, and R. Scopigno, "Fully automatic registration of image sets on approximate geometry," *International journal of computer vision*, vol. 102, no. 1-3, pp. 91–111, 2013.

[20] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[22] S. S. Tsai, D. Chen, G. Takacs, V. Chandrasekhar, R. Vedantham, R. Grzeszczuk, and B. Girod, "Fast geometric re-ranking for image-based retrieval," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 1029–1032.

[23] N. Piotto and G. Cordara, "Statistical modelling for enhanced outlier detection," in *Image Processing (ICIP), 2014 21th IEEE International Conference on*. IEEE, 2014.