State of the Art and Challenges in Crowd Sourced Modeling

JAN-MICHAEL FRAHM, PIERRE FITE-GEORGEL, ENRIQUE DUNN, Chapel Hill

ABSTRACT

The recent ubiquitous availability of large crowd sourced photo collections (e.g. Facebook, Picassa, Flickr) has opened a wide field of new applications ranging from virtual tourism, cultural heritage preservation to intelligence and defense applications. The research in this active area was energized by the seminal PhotoTourism paper (Snavely et al., 2006) starting from registering 3000 photos on 24 PCs in two weeks. In only five years the community advanced to dense scene reconstruction from 3 million images on as single PC within 24 hrs. At the same time these crowd sourced photo collections are growing at exponential rates posing significant challenges for their long term use. Even the current state-of-the-art methods in scene reconstruction from photo collections lag behind today's demand to scale to the true size of larger city scale datasets like the 12 million images of New York. Even more challenging is the long term demand for world scale reconstructions required for virtual tourism, disaster response and large image based localization for navigation. In this paper we will discuss the goals that need to be reached in term of scalability, robustness against temporal and apperance variation, within the context of generating high quality dense 3D models from heterogenous input imagery. We will present the current state of the art to achieve the goal and some of the challenges that are left to conquer.

1. GOALS FOR CROWD SOURCED MODELLING

We identify three crucial performance goals to be achieved by automated crowd sourced modeling systems (CSMS) in order to fulfill current and future performance demands, while broadening the scope of their applicability:

- Scalable Performance: A CSMS will ideally scale with not more than linear complexity in the number of input images. Such performance would enable city scale reconstructions of 12 million images (number of images of New York on Flickr) or more, while still creating complete models.
- Robustness to Temporal and Appearance Variation: Large image collections capture complex variation due to physical modification of the scene or atmospheric perturbations that will modify the appearance of a scene being photographed. A reconstruction system needs to be able to cope and detect these variations to ease the scene interpretation.
- High quality dense 3D models from heterogeneous input: Images found in Internet photo-collection describe the scene very differently because the devices used and their positioning vary greatly. Dense reconstruction systems that use this type of images need to leverage this formidable variation to obtain the most accurate model possible.

In this paper we review the state of the art while keeping in mind these goals. We outline some of the approaches that have been presented to meet these challenges and describe some of the open problems that need to be solved in order to obtain a truly usable reconstruction system using crowd sourced modeling.

2. SCALABILITY

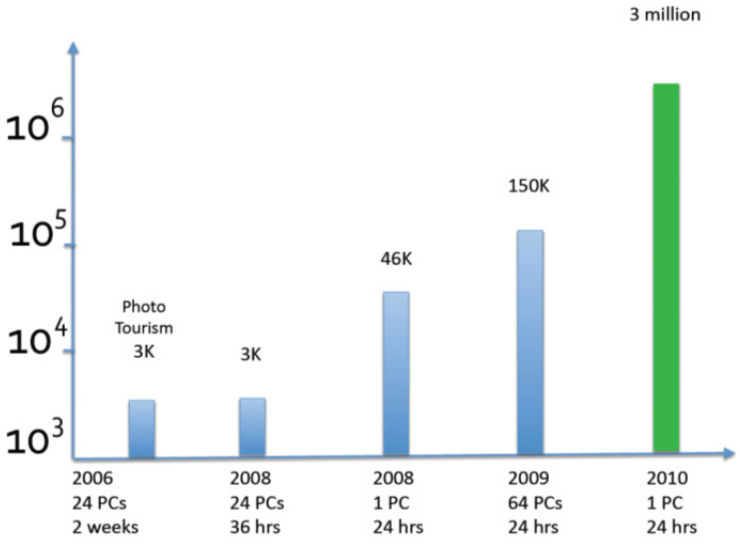

Figure 1: Performance review of current large-scale reconstruction algorithms. Benchmarking for the work of Frahm et al. (2010) is highlighted in green.

The current state of the art algorithms address in parts the challenges of large scale reconstruction. There are two main classes of city scale reconstruction algorithms the first class uses bag of words methods to identify scene overlap, which is then used to bootstrap large-scale structure from motion registration to obtain the spatial camera registration and the sparse 3D point cloud of the scene, see (Snavely et al., 2006 and Agar-wal et al., 2009). These methods were shown to scale to the com-putation of a few hundred thousand images within 24 hrs of computation by using a cloud computer with 64 CPUs. Hence they currently cannot be considered scalable for typical city scale reconstructions, which consist of multiple million images. The second class of methods uses appearance based image grouping followed by epipolar geometry verification to identify overlapping images. Afterwards an iconic scene representation is used to bootstrap the efficient image registration through structure from motion followed by efficient dense reconstruction. These methods can scale to the processing of a few million images on a single PC within 24 hrs as shown in (Frahm et al., 2010). To achieve the scalability these methods compromise model completeness in parts. Both classes of methods fail to achieve location recall beyond the major tourist cites of a city.

Bag of words methods: These methods typically provide a higher degree of model completeness in the reconstruction than the appearance clustering approaches. In order to achieve higher completeness they compromise the computational complexity of the method, the seminal Photo-Tourism (Snavely et al., 2006) used exhaustive search to find image overlap. Agarwal et al. (2009) later improved the performance of the approach by using the intrinsic parallelism of the problem to deploy feature based recognition and query expansion. This improved the computational complexity of the overlap detection. Still this only enabled the processing of a few hundred images on a cloud computer (64 PCs). To maintain computational efficiency these methods typically compromise location recall and do not recover scenes captured by a small number of images. Given the latter and foremost the lack of scalability we opt not to develop an approach in this class but rather a novel approach that takes the appearance clustering methods one step further.

Appearance based clustering: The strong advantage of the appearance clustering methods is their scalability, which has been shown in (Frahm et al., 2010) to outperform the previous state of the art by at least three orders of magnitude. This is a result of the close to linear computational complexity of the method introduced. A lot of care was taken to limit the number of image comparison. Each image is only compared to a fixed number of candidates. In this way, the first step is to cluster the data in small manageable chunks. While the obtained spatial camera registration delivers the largest so far obtained models of the sites in the reconstructed city, the method compromises on model completeness not achieving as complete models as the bag of words methods due to their more

restrictive overlap search model. Further similar to the bag of words methods, these methods have a small location recall as they do not recover sites covered by a small number of images.

Incremental structure-from-motion is the traditional method to create large-scale 3D reconstruction from overlapping images. It is used by algorithms in both of the above classes of methods to obtain the spatial camera registration. Incremental structure-from-motion methods have two main drawbacks. First it is not scalable, as it requires repetitive error mitigation through increasingly larger bundle adjustment, which optimizes the 3D structure simultaneously with the pose of the cameras, see (Snavely et al., 2008) . Secondly, it suffers from drift as error is propagated in the reconstruction. This limits the coverage of reconstructed models (Crandall et al., 2011). In order to tackle the problem of efficiency, Ni et al. (2007), proposed to optimize independent sub-maps in parallel while optimizing the area of overlap in a global thread. This does not only split the problem in parallel units but also in smaller problems that can be optimized more efficiently. Snavely et al. (2008) propose to limit the number of cameras to optimize by using a skeletal set, which provides the same reconstruction accuracy. The skeletal set is determined by searching for the spanning tree of cameras that approximate the uncertainty obtained from the complete set. The major drawback for the computational complexity with the skeletal set is that it still requires a full pair-wise matching of the image connection graph. The approach of Frahm et al. (2010) deploys high-level appearance information to reduce the inherent redundancy in the data by obtaining an iconic scene representation of the photo collection, which represents the scene in a sparse manner, see (Li et al., 2010).
Both of the above classes of methods do not address the computational complexity of the error mitigation that is performed through bundle adjustment, which is cubic in the parameter space, nor its increasing numerical sensitivity for a growing number of cameras. In Steffen et al. (2010), the authors developed a relative bundle adjustment based on trifocal constraints, which overcomes both of these limitations. It has the advantages to on one hand limit the size of the state vector, as the 3D structure does not have to be parameterized. This effectively reduces the computational effort. On the other hand it addresses the numerical sensitivity in contrast to traditional approaches by enabling an equivalent not just an approximate differential formulation, which essentially prevents the accumulation of uncertainties that lead to an increasing condition number of the optimization problem. The novel bundle adjustment has shown to have constant uncertainty towards the rotation parameters and linearly increasing uncertainty for the translation due to the unavoidable scale error propagation and it is highly robust to noisy initialization. Recently Crandall et al. (2011) proposed to use a discrete continuous method to limit the number of bundle adjustments. It requires a complicated discrete optimization to initialize the continuous optimizer. Using this method limits the drift drastically and allows for large-scale reconstruction. Strecha et al. (2010) propose to limit drift by registering a set of smaller reconstruction to a street map of the environment. All of the above methods heavily deploy the EXIF information to estimate the focal length along with the other camera calibration parameters like principal point and aspect ratio, which limits the applicability of such methods in the case of crowd sourced imagery as targeted in this proposal. Gherardi et al. (2010) propose a hierarchical reconstruction that runs parallel reconstructions that are iteratively merged. It starts with projective reconstruction and upgrades to a Euclidean reconstruction when the reconstruction contains enough information.
By looking at the progress made we can see than scalable solution for matching have been proposed but completeness, high location recall and structure from motion for uncalibrated camera on large photo-collection are still open problems that should be tackled by researchers.

3. ROBUSTNESS TO TEMPORAL AND APPEARANCE VARIATION

Figure 2: Example of appearance variation.

The role of illumination is dominant in the image generation process. Atmospheric perturbation, time of the day or artificial light will modify the appearance of the picture even if it was taken from the same location with the same camera. Such variations are always present in crowd source photograph because they are captured across time. These photo-collections offer a great source of information to model the captured scene. Unfortunately most algorithm used for dense reconstruction are not invariant to such variation. Goesele et al. (2007) propose two layers selection process. First each candidate images selects neighbors based on features matching, view-angles, and scale changes, and then at pixel level they propose a iterative disparity estimation using region growing starting from triangulated points obtained from structure from motion as they are expected to be more reliable. This method copes well the variation of camera locations presents in reconstruction of photo-collections unfortunately it does not offer a solution for image that have large illumination differences. In order to tackle this challenge we propose to select the images for multi-view stereo based on a clustering using color and gist features. Images presents in each obtained cluster are usually captured under similar condition. This allows us to create dense model from large collection of images capture in an uncontrolled environment. When such a dense model is available, Haber et al. proposed to estimate the reflectance and the incident illumination per-image thus allowing correct texturing of the estimated 3D model. In order to cope with erroneous model resulting from uncertain depth map, Goesele et al. propose to reduce visual artifact by using non-photorealistic rendering for background information. The background pixel color is estimated using random sampling color points along viewing rays. This offers a pleasing ghosting effect that smoothes out artifacts.

Color is not the only variation that is present in photo-collections; structural evaluation is also captured by images acquired across time. This is especially the case for 3D reconstruction from cities obtained using images spanning several decades, as new buildings are erected and older ones are destroyed. Schindler et al. propose to infer image acquisition time based on the presence or absence of certain landmark. They do not only estimate a time frame but evolution of a city skyline by estimating temporal visibility probability of each landmark. The goal of handling appearance variation has not yet been reached. Even if we can estimate dense model from varying illumination, every modeling systems only estimate a single texture, thus only representing an average view of the scene where there should be multiple for example switching between day and night.

4. HIGH QUALITY DENSE 3D MODELS FROM HETEROGENOUS INPUT

Dense geometry estimation from controlled binocular and multiview stereo sequences is an active and well studied problem. However, multiview stereo (MVS) for CSMS poses the challenge that the assumptions of uniform viewpoint distribution, homogenous sampling scale and scene illumination are not fulfilled across the entire datum. In an effort to mitigate these issues (while maintaining high quality dense models), state of the art approaches have addressed the aspects of intelligent datum segregation while developing their systems around more robust 3D geometric representations.

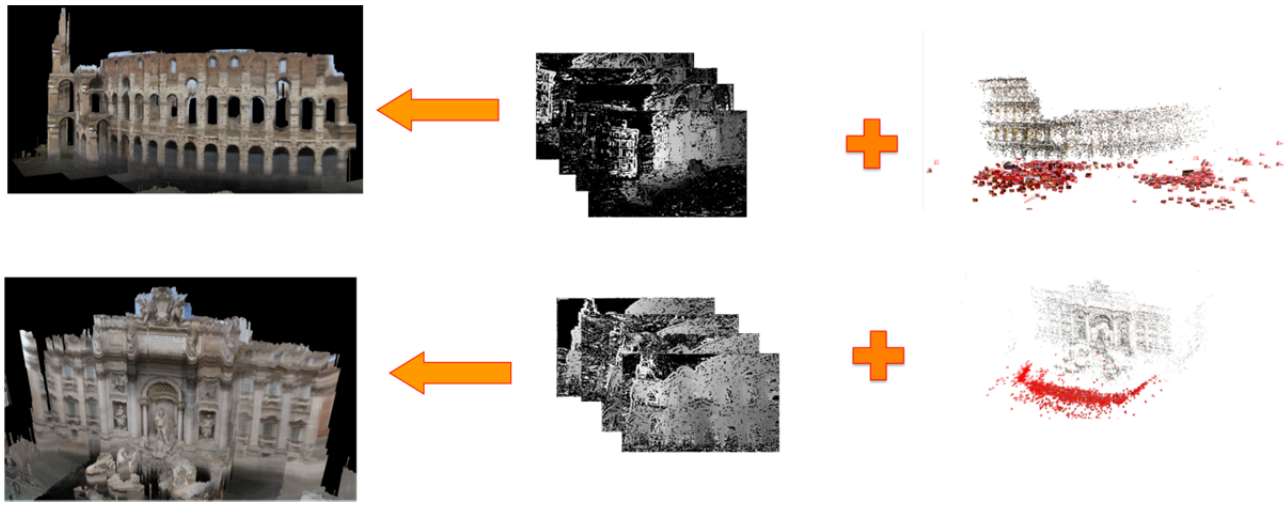

Figure 3: Dense geometry generation. On the right: SfM estimates and the result of image based clustering. Middle: Individual depthmaps obtained for different images in the cluster. At left: Textured 3D model obtained through heightmap based data fusion.

View Selection: Global image datum variability (in terms of effective scale, coverage and illumination) can be mitigated by performing partial model reconstruction on image clusters. These clusters aim at grouping images observing the same scene content while sharing common scale and global image appearance. Although such segregation may be directly obtained from preceding SfM modules, it is important to note that favorable viewpoints for sparse structure estimation may not provide an optimum datum for dense reconstruction. Examples of general viewpoint selection schemes for dense reconstruction include the works of Hornung et al. (2008) as well as Dunn and Frahm (2009). Addressing the problem of view selection in the context of crowd source imagery, Goesele et al. (2007) utilized the cardinality of the set of common features among viewpoints along with a parallax metric to build image clusters. Images are then rescaled to the lowest resolution view in the cluster and decisions on image inclusion into the MVS were performed on the pixel level basis according to photo-consistency thresholding. The work in Kanatani (2010) used SfM estimates to discard redundant images while retaining high resolution geometry. In this instance clustering was performed iteratively by merging sparse 3D points and viewpoints based on visibility relations. While no intra-cluster image segregation was made, the cluster size was bounded as a function of available computational resources. In Frahm et al. (2010) it was demonstrated how direct appearance based clustering utilizing color augmented GIST descriptors in conjunction with geometric verification provided a suitable datum for MVS. In Gallup et al. (2008) the concept of variable-baseline and variable resolution was introduced for stereo. This offers an elegant mechanism for heterogeneous image capture integration in MVS systems and may be readily be modified to operate within an image cluster oriented framework instead of the image sequence addressed in the original work.

Robust 3D Representations: Variations in ambiance illumination variations and/or partial scene occlusions, are common hindrances to robust dense geometry reconstruction that may not be completely resolved by selective MVS processing. Accordingly, recent work has studied dense structure representations capable of retaining fine structure detail while imposing implicit or explicit regularization on the obtained model surface. All of the following works implemented some form of input data clustering, but differ in the mechanisms used for surface generation. The work of Goesele et al. (2007) implemented clustering followed by region growing mechanisms for dense geometry estimation. In this work, surface normals were controlled explicitly within the photoconsistency function optimization. Recently, Furukawa et al. (2010) was able to implement region growing in city-scale input data. The work of Ackermann et al. (2010) uses MVS techniques to create a partial reconstruction of the scene, which serves as scene intrinsic reference geometry. The authors then transfer surface normals from reconstructed to unreconstructed regions based on robust photometric matching. The work presented in Frahm et al. (2010) addresses surface regularization through the use of a multiple layer heightmap based depthmap fusion module. In this way, water tight models of arbitrary geometry can be efficiently and robustly computed. Moreover, this type of regularization is especially well suited to urban scenes where vertical facades are abundant.

5. OUTLOOK AND CLOSING REMARKS

The application 3D reconstruction techniques for large unstructured image datasets is a very active research area that is being fueled by the increasing of availability of crowd sourced data. Current research efforts entail making careful design decisions in order to address the data association problem inherent in these vast photo collections. While the scale of the problem is formidable, a closer inspection also reveals a significant qualitative difference in the input datum. These differences in data scope and scale require also different approaches to integrating 3D reconstruction systems. In this communication we have focused on scalability, robustness and model quality as driving goals and challenges in this relatively new research area. However, suitable benchmarks for all of these goals need to be adopted within the research community in order to perform quantifiable comparisons.

6. REFERENCES

J. Ackermann, M. Ritz, A. Stork, and M. Goesele. Removing the Example from Photometric Stereo by Example. In: Proceedings of the ECCV 2010 Workshop on Reconstruction and Modeling of Large-Scale 3D Virtual Environments.

S. Agarwal, N. Snavely, I. Simon, S. Seitz, and R. Szeliski. Building rome in a day. IEEE International Conference on Computer Vision (ICCV), pages 1-8, Jul 2009.

M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. European Conference on Computer Vision (ECCV), IV: 778-792, Jul 2010.

D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. CVPR, 2011.

E. Dunn, and J.-M. Frahm. Next best view planning for active model improvement. In: Proceedings of British Machine Vision Conference, 2009.

P. Fite-Georgel, T. Johnson, and J.-M. Frahm. City-scale reality modeling from community photo collection. ISMAR Workshop on Augmented Reality Super Models, pages 1-4, Oct 2010.

J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building rome on a cloudless day. European Conference on Computer Vision (ECCV), pages 1-14, Jun 2010.

Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Towards Internet-scale Multi-view Stereo. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 2010.

D. Gallup, J.-M. Frahm, P. Mordohai, and M. Pollefeys. "Variable Baseline/Resolution Stereo", IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2008.

R. Gherardi, M. Farenzena, and A. Fusiello. Improving the efficiency of hierarchical structureand-motion. Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 1594-1600, 2010.

M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz (2007): Multi-View Stereo for Community Photo Collections. Proceedings of ICCV.

M. Goesele, J. Ackermann, S. Fuhrmann, C. Haubold, R. Klowsky, D. Steedly, and R. Szeliski (2010): Ambient Point Clouds for View Interpolation. Proceedings of ACM SIGGRAPH.

T. Haber, C. Fuchs, P. Bekaert, H.-P. Seidel, M. Goesele, and H. P. A. Lensch. Relighting Objects from Image Collections. Proc. CVPR, June 2009.

J. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1-8, 2008.

A. Hornung, B. Zeng, and L. Kobbelt. Image selection for improved multiview stereo. In: Proccedings of CVPR, pages 1-8, 2008

A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From structurefrom-motion point clouds to fast location recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1-8, Mar 2009.

X. Li, C. Wu, C. Zach, S. Lazebnik, and J.-M. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. European Conference on Computer Vision (ECCV), pages 1-14, Jul 2008.

Y. Li, N. Snavely, and D. P. Huttenlocher. Location recognition using prioritized feature matching. European Conference on Computer Vision (ECCV), pages 1-14, Sep 2010.

K. Ni, D. Steedly, and F. Dellaert. Out-of-core bundle adjustment for large-scale 3d reconstruction. IEEE International Conference on Computer Vision (ICCV), 2007.

R. Raguram and J.-M. Frahm. RECON: Scale-Adaptive Robust Estimation via Residual Consensus. To appear ICCV, 2011.

A. Saxena, S. H. Chung, and A. Ng. Learning depth from single monocular images. Advances in Neural Information Processing Systems, 18: 1161, 2006.

A. Saxena, M. Sun, and A.Y. Ng. Make3d: Learning 3d scene structure from a single still image. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 824-840, 2008.

G. Schindler and F. Dellaert (2010): Probabilistic Temporal Inference on Reconstructed 3D Scenes. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).

G. Schindler, F. Dellaert, and S. B. Kang (2007): Inferring Temporal Order of Images From 3D Structure. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).

F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". Computer Vision – ECCV 2002, pages 414-431, 2002.

N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), pages 835-846, 2006.

N. Snavely, S. Seitz, and R. Szeliski. Skeletal graphs for efficient structure from motion. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1-11, May 2008.

R. Steffen, J.-M. Frahm, and W. Forstner. Relative bundle adjustment based on trifocal constraints. ECCV workshop on Reconstruction and Modeling of Large-Scale 3D Virtual Environments, 2010.

C. Strecha, T. Pylvanainen, and P. Fua. Dynamic and scalable large scale image reconstruction. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1-8, Mar 2010.