Thesis Proposal:
Toward 3D reconstruction of static and dynamic objects

Enliang Zheng

Department of Computer Science
University of North Carolina at Chapel Hill

1 Introduction

Image-based 3D modeling is a process of scene reconstruction via a set of 2D images. Unlike standard image acquisition systems, which produce 2D images but lose important 3D information of the scene, image-based 3D modeling is an inverse process of recovering 3D information. With such extra 3D information in addition to color images, it arouses many applications including augmented/virtual reality [40], view synthesis [14], object recognition [9], and activity analysis [43]. Despite of many existing works, image-based modeling is still an active research topic in computer vision, as the task is inherently difficult and some of the issues still remain unaddressed.

The existing reconstruction algorithms aim at slightly different goals for the output. Based on the density of the reconstructed model, 3D reconstruction can be categorized into dense and sparse 3D reconstruction. While sparse 3D reconstruction recovers 3D information of a set of key points in the image, dense reconstruction aims at modeling the observed surface of the scene (represented as depthmaps or mesh models). Moreover, based on object motions, 3D reconstruction can also be classified as static object reconstruction and dynamic object reconstruction.

For my research, I take the geometrically registered images (either still images or images from videos) observing the same scene as input, and output 3D reconstructions, where the calibration parameters of images are estimated using existing structure from motion techniques [55]. The image typically captures a scene with both static objects (e.g. buildings and trees) and dynamic objects (e.g. people and cars). Given that both of them are important parts of the scene, my research starts from static object dense reconstruction and then extends to the more challenging area of dynamic object sparse reconstruction.

1.1 Static objects

Crowd-sourced images are widely available from Internet today, and static object reconstruction from these images has attracted much attention [20, 1, 23]. Though static object reconstruction has been studied extensively and improved upon over the years [27, 22, 8, 12, 47, 29], the 3D reconstruction using crowd-sourced images is still an on-going research [27, 23]. In this part of my work, I focus primarily on depthmap estimation using crowd-sourced images.

The depthmap represents the distance of the observed object surfaces relative to a camera. Key to the depthmap estimation is finding the correct 2D correspondences across images based on colors, because given 2D correspondences in different calibrated images, the task of reconstruction can be readily achieved through triangulation. However, finding the correct 2D correspondences is challenging due to a diversity of factors presented in crowd-sourced images, such as occlusions, calibration errors, illumination aberration, etc. In contrast, these challenges are not presented in

many previous works that takes as input the images captured under a controlled environment. To tackle this problem, it is commonly assumed there exist large subsets of images in the dataset that share similar capture properties with the reference image (the color image corresponding to the depthmap). Therefore, the task becomes one of finding a useful subset of images in the large dataset and using those images for reconstruction [27, 23].

The increasing availability of crowd-sourced datasets has explicitly brought efficiency and scalability to the forefront of application requirements. Computational complexity of depthmap estimation is typically high, since the process involves evaluations of a large number of depth hypotheses. This is especially true for large-baseline depthmap estimation, where the number of depth candidates may reach up to hundreds or thousands for each pixel. Much effort has been devoted to decrease the computational complexity by reducing the size of the depth hypothesis space [54, 8, 36, 44]. In particular, the PatchMatch depth hypothesis sampling scheme [8] has recently gained much interest due to its efficiency and effectiveness. This method initializes the depthmap with random values and then iteratively propagates good depth to the neighboring pixels, a process by which only a very small number of depth hypotheses are tested. I further explore the PatchMatch scheme in my research.

1.2 Dynamic objects

It is common that dynamic objects, such as people or cars, are present in crowd-sourced image collections. However, these dynamic objects are typically ignored in methods of static scene 3D reconstruction (e. g. [22, 27]), since those methods require concurrent capture of the object from multiple views to enable valid triangulation. In the problem of dynamic object reconstruction, no concurrent capture is assumed since otherwise it is the same to static object reconstruction. Therefore, dynamic object reconstruction is much more challenging.

Since the problem is difficult and fundamentally ambiguous, it is necessary to introduce some extra assumptions. Assumptions can be posed on scene geometry, object motion, capture pattern, etc. For instance, Bao et al. [4] estimate the depth of the dynamic object in a single image using the size of the detection box, with assumptions on the capture pattern and prior knowledge of the objects (i.e. the camera and the objects do not lie on the same plane, and the real object size is known). My research proposes to solve two different problems arising from different assumptions.

Trajectory triangulation [38, 51] and nonrigid structure from motion (NRSFM) [11, 32, 2, 17, 26] reconstruct dynamic objects using an image sequence that captures a dynamic motion. While NRSFM only reconstructs the 3D shape of the object without absolute translation (i. e. the center of the shape), trajectory triangulation methods in [38, 51] are able to recover the 3D positions. Compared to just using a single image for reconstruction [4, 41, 19, 34, 39, 63], using the image sequence is able to leverage the extra information that the same dynamic object captured at two consecutive time instances is spatially close.

The existing works on trajectory triangulation [38, 51] require the information of image sequencing (i. e. the temporal order of the images) to work properly. Moreover, the reconstructability, which defines the reliability and accuracy of the triangulated trajectory on a particular dataset, depends on both the real object trajectory and the motion of the camera center [38, 51]. In practice, if the dynamic object is captured by a video, the sequencing information is readily available, but the reconstructability is poor given the small motion of the camera center. Conversely, if the dynamic object is captured by multiple cameras, either in the form of images or videos, the sequencing information is difficult to attain. Hence, image sequencing and reconstructability are competing factors in the natural scene capture, and cannot be achieved at the same time. My research assumes the sequencing is unavailable, and aims at estimating both the sequencing and trajectory jointly.

2 Expected contributions

I expect to make several contributions that advance the state of the art in static object reconstruction using crowd-sourced images and in dynamic object reconstruction. I have already published/submitted works that support this research ([62, 63]). The contributions include:

Pixel level view selection for depthmap estimation: Aiming at static object depthmap estimation using crowd-sourced images, I will propose a new probabilistic framework that jointly models pixel-level view selection and depthmap estimation given the local pairwise image photo-consistency. The corresponding graphical model is solved by EM-based view selection probability inference and PatchMatch-like depth sampling and propagation. The algorithm is able to efficiently generate accurate results on large crowd-sourced dataset.

Trajectory triangulation from unsynchronized videos: I will propose a new formulation targeting the sparse 3D reconstruction of dynamic objects observed by multiple unsynchronized video cameras with unknown temporal overlap. The main contribution is that the new formulation uses the compressive sensing (L_1 norm) to explore the sequencing information among the frames of unsynchronized videos.

Joint object class sequencing and trajectory triangulation: I will propose a new problem that reconstructs the motion path of a class of dynamic objects in a scene from an unordered set of images. I will tackle the problem through an optimization of a non-convex cost function over both the unknown 3D points and the unknown topology of the path. The non-convex optimization is solved by leveraging the Generalized Minimum Spanning Tree (GMST).

3 Related work and background

Many works related to static and dynamic object reconstruction have been proposed. The following sections outline several works in each of these research areas.

3.1 Static objects

Many research works have been devoted to robustly estimate depth. Depthmap estimation handling occlusion first emerged in two-view stereo [48, 49, 56]. Unlike two-view stereo, where computing the depth of the occluded pixel is infeasible due to lack of information, the partial occlusion presented in multi-view depth estimation (MVDE) can be solved by leveraging the additional view redundancy. Here, partial occlusion means the object is visible only in a subset of source images. Therefore, it is important to determine a subset of the source images for depth estimation of a particular pixel. Kang et al. [33] explicitly address occlusion in multi-baseline stereo by only using the subset of heuristically selected overlapping cameras with minimum matching cost. Campbell et al. [13] choose the best few depth hypotheses for each pixel, following with a Markov Random Field (MRF) optimization to determine a spatially consistent depthmap. Their method chooses source images based on spatial proximity of cameras. The heuristic provides occlusion robustness as long as there is a sufficient number of unoccluded views (typically 50%). Strecha et al. [45] handle occlusion in wide-baseline multi-view depth estimation by including visibility within a probabilistic model, where the depth smoothness is enforced on neighboring pixels according to the color gradient. The work of Strecha et al. [45] is further extended in [46], where the depth and visibility are jointly modeled by hidden MRFs. In [46], the memory used for visibility configuration of each pixel is 2^K,

which grows exponentially with respect to the number of input images K. Hence, the approach is limited to very few images (three images in their evaluation).

In addition to pixel-level view selection for handling occlusion [45, 46, 27], the benefits of fine-grain data association have also been demonstrated in the work of Gallup et al. [24]. They present a variable-baseline and variable-resolution framework for MVDE, exploring the attainment of pixel-specific data associations for capture from approximately linear camera paths. While this work illustrates the benefits of fine-grain data association strategies in multi-view stereo, it does not easily generalize to irregularly captured datasets.

To improve depthmap accuracy, some methods rely on mutual depth consistency across multiple depthmaps. Shen [42] computes the depthmap for each image using PatchMatch stereo and removes outlier depths by enforcing depth consistency over neighboring views. Hu and Mordohai [30] follow a scheme similar to Campbell et al. [13], but select the final depth through a process enforcing mutual consistency across all depthmaps. Both of these works require the depthmaps of other views to be available, which may be a constraint for real applications.

Other stereo methods aim at generating a consistent 3D model instead of depthmaps. Furukawa et al. [22] present an accurate patch-based MVS approach that starts from a sparse set of matched keypoints that are repeatedly expanded until visibility constraints are invoked to filter out false matches. Zaharescu et al. [58] propose a mesh evolution framework based on a new self-intersection removal algorithm. Jancosek et al. [31] propose a method that additionally reconstructs surfaces that do not have direct support from the input 3D points by exploiting visibility in 3D meshes.

Robust stereo performance for the more challenging crowd-sourced data is an ongoing research effort. Frahm et al. [20] discern a suitable input datum by using appearance clustering using a color-augmented GIST descriptor along with feature-based geometric verification. Furukawa et al. [23] use structure from motion (SFM) to purge redundant imagery but retain high resolution geometry. Their iterative clustering approach merges sparse 3D points and cameras based on visibility analysis. Although intra-cluster image partitioning is not performed, the cluster size is limited in an effort to maintain computational efficiency. Goesele et al. [27] address the problem of viewpoint selection for crowd-sourced imagery by building small-sized image clusters using the cardinality of the set of common features among viewpoints and a parallax-based metric. In their method, images are resized to the lowest common resolution in the cluster. Pixel depth is then computed using four images selected from the cluster based on local color consistency.

Because the depth hypothesis space is huge in wide-baseline stereo, and because evaluating all these hypotheses is computationally expensive, much effort has gone into reducing the hypothesis space without degrading the quality of the resulting depthmaps. Hierarchical stereo (HS) [36, 44] is a multi-resolution approach for deterministic search space reduction. Wang et al. [54] apply Sequential Probability Ratio Test (SPRT) mode to reduce the depth (or disparity in two-view stereo) hypothesis space.

Another popular method to reduce the depth hypothesis space is the recently proposed PatchMatch depth sampling scheme [8]. It is first introduced to solve the two view stereo problem in [8]. PatchMatch initializes each pixel with a random slanted plane at random depth and then propagates well-fitting pixel depths. The nearby and current pixels' slanted planes are tested, and the one with the best cost is kept. Besse et al. [7] combine the PatchMatch sampling scheme and belief propagation to infer an MRF model that has smoothness constraints. By combining guided filter and PatchMatch, Lu et al. [35] provide an efficient edge-aware filtering for correspondence field estimation, which can be applied in two-view stereo. While PatchMatch stereo was originally a sequential method, Bailer et al. [3] parallelize the algorithm by restricting the propagations to only horizontal and vertical directions. I propose to further explore the potential of PatchMatch in wide-baseline stereo with a large hypotheses space.

3.2 Dynamic objects

The problem of 3D reconstruction from a single image is inherently ambiguous and difficult without further assumptions. Some works predict the depth of a monocular image by training a model that leverages monocular cues such as texture gradient, light and shading, etc. [41, 19, 34]. Saxena et al. [41] compute depthmaps from a single still image by using a hierarchical multi-scale Markov Random Field (MRF) that incorporates several hand-designed features. Eigen et al. [19] and Liu et al. [34] use a deep neural network as the model so that the heuristic feature design step is avoided. All these methods generate reasonable depthmaps, yet their results are generally far away from ground truth. In man-made scenes with mainly orthogonal facades (called a Manhattan world [16]), 3D reconstruction from a single image can be simplified to finding 3D lines and planes within the scene. Delage et al. [18] use a MRF model to identify the different planes and edges in the scene, as well as their orientations. Then, an iterative optimization algorithm is applied to infer the planes' positions. Ramalingam et al. [39] reconstruct the 3D lines in a Manhattan scene from an image using linear programming that identifies a sufficient minimal set of least-violated line connectivity constraints.

As opposed to depth recovery from a single image, non-rigid structure from motion (NSFM) aims to recover a deforming object's structure as well as camera motion given corresponding 2D points in a sequence of images. Tomasi and Kanade [50] propose to achieve rigid structure from motion through matrix factorization under an affine camera assumption. An important extension of the well-known Tomasi-Kanade factorization is the work by Bregler et al. [11], which tackles the NSFM problem by assuming the shape of the object can be represented by a linear combination of low-order shape bases. Due to the fact that the shape bases are not unique, Xiao et al. [32] prove that using only rotation constraints results in ambiguous and invalid solutions. To solve this shape ambiguity, most existing works rely on different prior knowledge specific to the problem at hand. Recently, Dai et al. [17] solved the problem by introducing a prior-free method. It should be noted that all these methods require a certain number of points to be available in each frame in order to well capture the underlying shape of the deforming object.

As a dual method to the shape basis assumption, Akhter et al. [2] proposed that the smooth trajectory of each point across time can be restricted to a low-dimensional subspace and represented by a linear combination of Discrete Cosine Transform (DCT) bases. However, this method fails completely if image frames are randomly shuffled, in contrast to the method proposed by Dai et al. [17], which requires no temporal information. Moreover, the above NSFM formulations assume affine cameras with orthographic projection. Projective NSFM, in contrast, is much more difficult, as the unknown depths of the points present one more set of unknown parameters. This problem has not yet been completely solved [28, 53].

Another set of problems closely related to NSFM focuses on 3D reconstructions of dynamic objects given camera registrations. These camera registrations are typically computed from structure from motion (e.g. VisualSFM [55]) that leverages the static background scene. Park et al. [38] reconstruct the 3D trajectory from a sequence of images by approximating the trajectory with a number of low-order DCT bases (similar to Akhter et al. [2]). Their method recovers accurate 3D trajectory, but it has two major flaws, as noted in [51]: (1) The predefined number of bases for each image sequence is difficult to set, and (2) the accuracy of 3D trajectory reconstruction is fundamentally limited by the correlation between the trajectory of 3D points and the motions of camera centers. This high correlation of object and camera motion commonly occurs in real captures and thus degrades the reconstruction results. Valmadre et al. [51] recover the trajectory by minimizing the response of the trajectory to a set of high pass filters. Their method, in contrast to Park et al. [38], requires no basis size but still suffers under the correlation between object and

camera motion.

The works by Park et al. [37] and Valmadre et al. [51] are further extended to estimate the 3D motion of an articulated object, where a child 3D point remains a fixed distance with respect to its parent point. (For instance, the distance between 3D points on the left foot and the left knee is fixed.) Based on [38], Park et al. [37] reconstruct 3D articulated motion with the constraint that a trajectory remains at a fixed distance with respect to a parent trajectory. This improves the reconstructibility over their earlier approach [38], but involves solving an NP-hard quadratic programming problem. Valmadre et al. [52] develop a dynamic programming approach which scales linearly in the number of frames to overcome solving the quadratic programming problem. All of these methods require a given temporal order of the captured frames, which may not be available in some situations. Taking one step forward, my research focuses on the dynamic object reconstruction without temporal information.

Recovering temporal order of the images, also called image sequencing, is a nontrivial task. Basha et al. [5, 6] has recently proposed two methods that determine the temporal order of photos taken by a set of cameras. In [5], they compute the partial orderings for a subset of the images by analyzing the dynamic features present in the subset. The method inherently relies on two images taken from the same static camera to eliminate the uncertainty in the sequencing. Their later work [6] proposes to enforce the constraint of a known order for the images taken by each camera. While these methods provide useful insight to the difficult problems, they still suffer from various limitations.

4 Research Plan

The following sections provide an overview and details for the proposed research to be completed for the thesis.

4.1 Pixel level view selection for depthmap estimation

Multi-view depthmap estimation methods strive to determine a view-dependent depthfield by leveraging the local photoconsistency of a set of overlapping images observing a common scene. Achieving highly accurate depthmaps is inherently difficult even for well-controlled environments where factors such as viewing geometry, image-set color constancy, and optical distortions are rigorously measured and/or corrected. In contrast, practical challenges for robust depthmap estimation from non-controlled input imagery (e.g. crowd-sourced images) include heterogeneous resolution and scene illuminations, unstructured viewing geometry, scene content variability, and image registration outliers. Moreover, the increasing availability of crowd-sourced datasets has explicitly brought efficiency and scalability to the forefront of application requirements, while implicitly increasing the importance of data association management when processing such large-scale datasets.

Assuming there is a subset of source images suitable for depthmap estimation of a given reference image, we will propose to jointly determine the subset and estimate the depthmap on the pixel level. If the depth is known, we can easily choose the subset of source images based on color consistency. Conversely, the depth estimation becomes easy if the subset of source images is available. However, both pieces of information are not available in reality. To accomplish this chicken-and-egg problem, I will propose a probabilistic framework for depthmap estimation that jointly models pixel-level view selection and depthmap estimation given pairwise image photoconsistency. The corresponding graphical model will be solved by EM-based view selection probability inference and PatchMatch-like depth sampling and propagation [8]. That is, we estimate the image selection while the depth is fixed, and vice versa. The insight leveraged by the method is that at correct depth hypotheses,

the photo-consistency measure of the neighboring pixels with respect to a source image is spatially smooth.

The algorithm is easily parallelizable and will be implemented in GPU. I will evaluate the accuracy of the algorithm on the standard benchmark datasets [47]. The qualitative results on large crowd-sourced datasets will be shown and compared to the method by Goesele et al. [27]. The speed of the algorithm will be compared to the planesweeping algorithm [33] that requires searching in a large depth hypothesis space.

A working prototype has developped and this research was published in [62].

4.2 Trajectory triangulation from unsynchronized videos

We target the problem of sparse reconstruction of dynamic objects captured by a variety of unsynchronized video cameras. This type of setup is common in real life. For instance, several people might use their own video cameras to capture an event, such as people dancing. Solving this problem is nontrivial because the cameras are not synchronized, meaning the temporal order of the frames is only known within each video sequence independently. Hence, for a successful reconstruction, it is required to recover the temporal alignment across video sequences, accounting for the potentially different and unknown frame rates of the cameras.

We will use the observation that since the dynamic objects move smoothly, the 3D shape of the object at time t can be approximated by a linear combination of the shape at time $t-1$ and time $t+1$; the shapes at other time instances are less related. Despite the unknown image sequencing information, we can find the linear coefficients of the most related shapes at time $t-1$ and time $t+1$ via compressive sensing (L_1 norm). The problem will be modeled as a biconvex function, which is optimized in an alternating manner. That is, we optimize over 3D shape while fixing sequencing information, and vice versa. As the general solver is typically slow for large problems, and to make the algorithm more scalable, we will design a special solver based on alternating direction method of multipliers (ADMM) optimization [10].

The algorithm will be tested on both synthetic and real datasets. I will generate synthetic videos by projecting the real motion capture dataset into the virtual images. Since the ground truth is known, the results will be quantitatively evaluated. For the real datasets, we will use both publicly available and self-captured data for qualitative evaluations.

This work was submitted to ICCV 2015 [64], but still requires more experiments.

4.3 Joint object class sequencing and trajectory triangulation

I will reconstruct the motion path of a class of dynamic objects through a scene from an unordered set of images. Once the motion path is reconstructed, the 3D positions of the dynamic objects is reconstructed. As this problem is ambiguous without introducing any prior knowledge, we assume objects of the same class move in a common path in the real scene. For instance, pedestrians walk on the sidewalk, and vehicles drive in road lanes. I will leverage standard object detection techniques to identify object instances within a set of registered images. Each of these object detections defines a single 2D point with a corresponding viewing ray. The set of viewing rays attained from the aggregation of all detections belonging to a common object class will then be used to estimate a motion path denoted as the object class trajectory.

As is shown in Park et al. [38] and Valmadre et al. [51], temporal sequencing information should be known for valid trajectory triangulation. In object class trajectory, the sequencing will be defined as the topology of the objects along the 3D path. Therefore, I propose to jointly estimate the sequencing and triangulate the object class trajectory.

The problem will be solved by formulating a non-convex function over unknown sequencing and 3D dynamic object positions. The optimization will be achieved in two steps: In the first step, a discrete and approximate version of the original problem is constructed and solved by finding a Generalized Minimum Spanning Tree (GMST) on a multipartite graph. In the second step, the solution from the first step is further refined through a continuous convex optimization over the 3D points. The algorithm will be tested on both synthetic and real datasets.

A working prototype has developped and this research was published in [63].

5 Remaining Tasks

The following is a list of remaining tasks to be completed before graduation. The absolute requirements are the completion of a final project in COMP 790-134, Machine Learning with Discriminative Methods, taught by Professor Alex Berg. The project is to design a noisy depthmap filter using a deep neural network.

- Finish trajectory triangulation from unsynchronized videos, and write a journal paper for it
  - Add the mechanism to handle the case of occasional missing 2D measures.
  - Include one more experiment using a real dataset.
  - Explore a faster optimization solver based on the method by Chen et al. [15].
- Open-source the code for the pixel-level view selection and depthmap estimation [62]
  - The global view selection in [27] should be included in the code to handle the case when the number of source images is large and there is not enough GPU memory for the whole algorithm.
  - The raw depthmap of the code contains noise, which should be removed. The final project in COMP 790-134 can be incorporated into the code.
  - Incorporate the code into the structure from motion software of Johannes Schönberger for easier code usage and depth visualization.
- Journal version of the pixel-level view selection and depthmap estimation
  - Replace the slower normalized cross correlation (NCC) with the faster census transform [57] for color consistency measure. To achieve this, the likelihood function should be changed accordingly.
  - Include a filter to remove the noise in the depthmap.

6 Proposed Oral Examination Topics

The following are sample topics and related citations to be discussed in my oral examination:

- Static object stereo
  - Goesele et al. [27], Bleyer et al. [8], Furukawa and Ponce [22], Campbell et al. [12], Kang et al. [33], Strecha et al. [46], Shen [42], Jancosek and Pajdla [31], Gallup et al. [24], Frahm et al. [20], Furukawa et al. [23], Gallup et al. [25]
- Trajectory triangulation and image sequencing
  - Park et al. [38], Park and Sheikh [37], Valmadre and Lucey [51], Valmadre et al. [52], Basha et al. [5], Basha et al. [6]

- Nonrigid structure from motion
  - Bregler et al. [11], Jing et al. [32], Akhter et al. [2], Dai et al. [17], Garg et al. [26]

7 Published Works

The following is a list of the works that I authored or contributed to during my graduate research:

2011

[60] E. Zheng, R. Raguram, P. Fite-Georgel, and J-M Frahm. Efficient generation of multi-perspective panoramas. In *3DIMPVT*, 2011

2012

[61] E. Zheng, E. Dunn, R. Raguram, and J-M Frahm. Efficient and scalable depthmap fusion. In *BMVC*, 2012

2013

[21] J-M Frahm, J. Heinly, E. Zheng, E. Dunn, P. Fite-Georgel, and M. Pollefeys. Geo-registered 3d models from crowdsourced image collections. *Geo-spatial Information Science*, 2013

2014

[62] E. Zheng, E. Dunn, V. Jojic, and J-M Frahm. Patchmatch based joint view selection and depthmap estimation. In *CVPR*, 2014

[63] E. Zheng, K. Wang, E. Dunn, and J-M Frahm. Joint object class sequencing and trajectory triangulation (jost). In *ECCV*. 2014

2015

[64] E. Zheng, D. Ji, E. Dunn, and J-M Frahm. Sparse dynamic 3d reconstruction from unsynchronized videos. In *ICCV*, 2015, submitted

[65] E. Zheng, K. Wang, E. Dunn, and J-M Frahm. Minimal solvers for 3d geometry from satellite imagery. In *ICCV*, 2015, submitted

[65] E. Zheng and C. Wu. Structure from motion using structure-less resection. In *ICCV*, 2015, submitted

Bibliography

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. Seitz, and R. Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011.

[2] I. Akhter, Y. A. Sheikh, S. Khan, and T. Kanade. Nonrigid structure from motion in trajectory space. In *NIPS*, 2008.

[3] C. Bailer, M. Finckh, and H. P. A. Lensch. Scale robust multi view stereo. In *ECCV*, 2012.

[4] S. Bao, M. Sun, and S. Savarese. Toward coherent object detection and scene layout understanding. *Image and Vision Computing*, 2011.

[5] T. Basha, Y. Moses, and S. Avidan. Photo sequencing. In *ECCV*, 2012.

[6] T. Basha, Y. Moses, and S. Avidan. Space-time tradeoffs in photo sequencing. In *ICCV*, 2013.

[7] F. Besse, C. Rother, and J. Kautz. Pmbp: Patchmatch belief propagation for correspondence field estimation. In *BMVC*, 2012.

[8] M. Bleyer, C. Rhemann, and C. Rother. Patchmatch stereo-stereo matching with slanted support windows. In *BMVC*, 2011.

[9] L. Bo, K. Lai, X. Ren, and D. Fox. Object recognition with hierarchical kernel descriptors. In *CVPR*, 2011.

[10] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 2011.

[11] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *CVPR*, 2000.

[12] N. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *ECCV*. 2008.

[13] N. D. F. Campbell, G. Vogiatzis, C. H. Esteban, and R. Cipolla. Using multiple hypotheses to improve depthmaps for multi-view stereo. In *ECCV*, 2008.

[14] S. Chen and L. Williams. View interpolation for image synthesis. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, 1993.

[15] Y. Chen, J. Mairal, and Z. Harchaoui. Fast and Robust Archetypal Analysis for Representation Learning. In *CVPR*, 2014.

[16] J. M. Coughlan and A. L. Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *ICCV*, 1999.

[17] Y. Dai, H. Li, and M. He. a simple prior-free method for non-rigid structure-from-motion factorization. *CVPR*, 2012.

[18] E. Delage, H. Lee, and A. Y. Ng. Automatic single-image 3d reconstructions of indoor manhattan world scenes. In *International Symposium on Robotics Research*, 2005.

[19] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014.

[20] J-M Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y. Jen, E. Dunn, B. Clipp, S. Lazebnik, et al. Building rome on a cloudless day. In *ECCV 2010*. 2010.

[21] J-M Frahm, J. Heinly, E. Zheng, E. Dunn, P. Fite-Georgel, and M. Pollefeys. Geo-registered 3d models from crowdsourced image collections. *Geo-spatial Information Science*, 2013.

[22] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *PAMI*, 2010.

[23] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Towards Internet-scale multi-view stereo. In *CVPR*, 2010.

[24] D. Gallup, J.-M. Frahm, and P. Mordohaiand M. Pollefeys. Variable baseline/resolution stereo. In *CVPR*, 2008.

[25] D. Gallup, M. Pollefeys, and J. Frahm. 3d reconstruction using an n-layer heightmap. In *Pattern Recognition*. 2010.

[26] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *CVPR*, 2013.

[27] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz. Multi-view stereo for community photo collections. In *ICCV*, 2007.

[28] R. Hartley and R. Vidal. Perspective nonrigid shape and motion recovery. In *ECCV*. 2008.

[29] Heiko Hirschmuller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *CVPR*, 2007.

[30] X. Hu and P. Mordohai. Least commitment, viewpoint-based, multi-view stereo. In *3DIMPVT*, 2012.

[31] M. Jancosek and T. Pajdla. Robust, accurate and weakly supported-surfaces preserving multi-view reconstruction. In *CVPR*, 2011.

[32] X. Jing, C. Jinxiang, and K. Takeo. A closed-form solution to non-rigid shape and motion recovery. In *ECCV*, 2004.

[33] S. Kang, R. Szeliski, and J. Chai. Handling occlusions in dense multi-view stereo. In *CVPR*, 2001.

[34] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. *arXiv preprint arXiv:1411.6387*, 2014.

[35] J. Lu, H. Yang, D. Min, and M. N. Do. Patch match filter: Efficient edge-aware filtering meets randomized search for fast correspondence field estimation. In *CVPR*, 2013.

[36] G. Van Meerbergen, M. Vergauwen, M. Pollefeys, and L. Van Gool. A hierarchical symmetric stereo algorithm using dynamic programming. *IJCV*, 2002.

[37] H. Park and Y. Sheikh. 3d reconstruction of a smooth articulated trajectory from a monocular image sequence. In *ICCV*, 2011.

[38] H. Park, T. Shiratori, I. Matthews, and Y. Sheikh. 3d reconstruction of a moving point from a series of 2d projections. In *ECCV*. 2010.

[39] S. Ramalingam and M. Brand. Lifting 3d manhattan lines from a single image. In *ICCV*, 2013.

[40] B. Reitinger, C. Zach, and D. Schmalstieg. Augmented reality scouting for interactive 3d reconstruction. In *Virtual Reality*, 2007.

[41] A. Saxena, S. H. Chung, and A. Y. Ng. 3d depth reconstruction from a single still image. *IJCV*, 2008.

[42] S. Shen. Accurate multiple view 3d reconstruction using patch-based stereo for large-scale scenes. In *TIP*, 2013.

[43] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 2013.

[44] M. Sizintsev. Hierarchical stereo with thin structures and transparency. In *Computer and Robot Vision*. IEEE, 2008.

[45] C. Strecha, R. Fransens, and L. V. Gool. Wide-baseline stereo from multiple views: a probabilistic account. In *CVPR*, 2004.

[46] C. Strecha, R. Fransens, and L. V. Gool. Combined depth and outlier estimation in multi-view stereo. In *CVPR*, 2006.

[47] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *CVPR*, 2008.

[48] J. Sun, H.-Y. Shum, and N.-N. Zheng. Stereo matching using belief propagation. In *ECCV*, 2002.

[49] J. Sun, Y. Li, S. B. Kang, and H.-Y. Shum. Symmetric stereo matching for occlusion handling. In *CVPR*, 2005.

[50] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 1992.

[51] J. Valmadre and S. Lucey. General trajectory prior for non-rigid reconstruction. In *CVPR*, 2012.

[52] J. Valmadre, Y. Zhu, S. Sridharan, and S. Lucey. Efficient articulated trajectory reconstruction using dynamic programming and filters. In *ECCV*, 2012.

[53] R. Vidal and D. Abretske. Nonrigid shape and motion from multiple perspective views. In *Computer Vision–ECCV 2006*. 2006.

[54] Y. Wang, K. Wang, E. Dunn, and J. Frahm. Stereo under sequential optimal sampling: A statistical analysis framework for search space reduction. In *CVPR*, 2014.

[55] C. Wu. Visualsfm: A visual structure from motion system. In *http://homes.cs.washington.edu/~ccwu/vsfm/*, 2011.

[56] J. Xiao, J. Chen, D.-Y. Yeung, and L. Quan. Learning two-view stereo matching. In *ECCV*, 2008.

[57] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *ECCV*. 1994.

[58] A. Zaharescu, E. Boyer, and R. P. Horaud. Topologyadaptive mesh deformation for surface evolution, morphing, and multi-view reconstruction. In *PAMI*, 2011.

[59] E. Zheng and C. Wu. Structure from motion using structure-less resection. In *ICCV*, 2015, submitted.

[60] E. Zheng, R. Raguram, P. Fite-Georgel, and J-M Frahm. Efficient generation of multi-perspective panoramas. In *3DIMPVT*, 2011.

[61] E. Zheng, E. Dunn, R. Raguram, and J-M Frahm. Efficient and scalable depthmap fusion. In *BMVC*, 2012.

[62] E. Zheng, E. Dunn, V. Jojic, and J-M Frahm. Patchmatch based joint view selection and depthmap estimation. In *CVPR*, 2014.

[63] E. Zheng, K. Wang, E. Dunn, and J-M Frahm. Joint object class sequencing and trajectory triangulation (jost). In *ECCV*. 2014.

[64] E. Zheng, D. Ji, E. Dunn, and J-M Frahm. Sparse dynamic 3d reconstruction from unsynchronized videos. In *ICCV*, 2015, submitted.

[65] E. Zheng, K. Wang, E. Dunn, and J-M Frahm. Minimal solvers for 3d geometry from satellite imagery. In *ICCV*, 2015, submitted.