%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

\chapter{Related work}

3D reconstruction from 2D imagery has been studied intensively and extensively by many researchers in the computer vision community. In this section, I first review the works for camera parameter estimation, and then survey the works related to static and dynamic object reconstruction respectively.

\section{Camera Parameter Estimation}

Camera parameters are generally considered as an prerequisite for 3D reconstruction, since they provide the geometric relationship among multiple cameras. For instance, with this geometric information, the mapping from a 3D point to an image pixel can be uniquely determined.

Being seperated into two parts, the internal (intrinsic) camera parameters has focal length, principle point, skew parameter, and radial distortion that convert the space coordinates to image coordiantes, and the external (extrinsic) part describes a camera's rotation and translation relative to a global coordinate system \cite{Hartley2004}.

Given the importance of camera parameters in computer vision tasks such as 3D reconstruction, lots of works have been done in estimating camera parameters, a process also called camera calibration. Earlier works for camera calibration require a calibration object such as a planar checkerboard to be seen by the cameras \cite{conf/cvpr/SturmM99,zhang2000flexible,caltoolbox}, which impose a significant constraint in its real applications. Thanks to the recent development of techniques in structure from motion (SfM) \cite{Snavely2,snavely2008modeling,WuVSFM,wilson2013network,heinly2014\_duplicate\_structure,schoenberger2015paige,Heinly,heinly\_dissertation,zheng2015\_structureless\_resection}, camera calibration can be achieved by just leveraging 2D correspondences among multiple images.

Structure from motion is a pipeline that targets estimating the cameras parameters of the images observing a common static scene. A typical pipeline includes the main steps of feature extraction \cite{lowe2004\_sift,rublee2011\_orb,bay2008\_surf}, inlier correspondence searching \cite{raguram2013usac}, camera pose estimation \cite{nister2003\_five\_point,kneip2011novel,zheng2014general,zheng2015\_structureless\_resection} and bundle adjustment \cite{agarwal2010\_ba,wu2011\_multicore\_ba}. Recent works in structure from motion has exhibited its accuracy, efficiency and robustness to be applicable in most real scenarios \cite{Snavely2,WuVSFM}.

\section{Static Scene Reconstruction}

As a main research subject in 3D computer vision, there are tons of works addressing issues in static scene reconstruction. Early works mainly focus on depthmap estimation on binocular images \cite{boykov2001fast,Sun\_ECCV2002\_stereoBeliefProp,scharstein2002taxonomy,scharstein2007learning}.

In these works, the two images are rectified so that searching the correspondence of a pixel is executed on the same row in another image. In contrast, multiview depthmap estimation (MVDE) use multiple images to reduce the ambiguities in searching correspondences. Moreover, the redundant information among the estimated depthmaps can be leveraged to filter out outlier depth.

This section first discusses the most related works for multiview depthmap estimation and the associated issues such as robustness and efficiency, and then discusses briefly the methods generating a consistent point cloud or meshes.

\subsection{Multiview Depthmap Estimation}

Handling occlusion is important in depthmap estimation, and it firstly emerged in two view stereo \cite{Sun\_ECCV2002\_stereoBeliefProp, Sun\_CVPR2005\_stereo, Xiao:2008}. However, in these methods, the occluded pixel region is only marked with unknown depth due to the unavailable correspondence in another image.

In principle, the additional view redundancy available to MVDE can be leveraged to resolve occlusions. \citet{handle\_occlusion2001} explicitly address occlusion in multi-baseline stereo by only using the subset of the heuristically selected overlapping cameras with the minimum matching cost. The heuristic provides occlusion robustness as long as there is a sufficient number of unoccluded views (typically 50\%). \citet{MultiHypothesis\_ECCV2008} choose the best few depth hypotheses for each pixel, following with a MRF optimization to determine a spatially consistent depthmap. Their method chooses source images based on spatial proximity of cameras. \citet{Strecha\_BayesModelCVPR2004} handle occlusion in wide-baseline multi-view stereo by including visibility within a probabilistic model, where the depth smoothness is enforced on neighboring pixels according to the color gradient. The work by \citet{Strecha\_BayesModelCVPR2004} is further extended in \citet{CombinedDepthOutlier} where the depth and visibility are jointly modeled by hidden Markov random fields. In the work by \citet{CombinedDepthOutlier}, 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 contrast, our memory usage is linear with the number of images $K$.

\citet{Gallup08} 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 that work illustrates the benefits of fine grain data association strategies in multi-view stereo, it does not easily generalize to irregularly captured datasets.

Given the redundant information among multiple depthmaps, lightweight depthmap fusion removes outlier depth by leveraging the mutual depth consistency among multiple depthmaps. \citet{Shen\_TIP2013} computes the depthmap for each image using PatchMatch \mbox{stereo}, and enforces depth consistency over neighboring views. \citet{LeastCommitment\_3DIMPVT2012} follows a scheme similar to Camppbell \citet{MultiHypothesis\_ECCV2008} but select the final depth through a process enforcing mutual consistency across all depthmaps. These methods require the depthmaps of other views to be available, while in contrast we aim at developing a method that directly outputs an accurate depthmap.

\subsection{Robustness}

Robust stereo performance for crowd sourced data is an ongoing research effort.

Images downloaded from Internet (such as Flickr\footnote{https://www.flickr.com/} or Panoramio\footnote{http://www.panoramio.com/}) by searching key words are unstructured imagery with a large portion of unrelated images. To discern a suitable input datum for stereo, \citet{Frahm2010} use appearance clustering of a color augmented GIST descriptor \cite{oliva2001modeling} along with feature-based geometric verification. In contrast, the work by \citet{Heinly} discovers the image relationship in a streaming paradigm that register images to a vocabulary tree built online. However, even the unrelated images are purged, using the data for stereo is still challenging due to the heterogeneous capture characteristics.

To estimate the depthmap of an image, \citet{Frahm2010} select the most related images based on the number of sparse feature points shared in common. The depthmap is then estimated using the heuristic K-best planesweeping algorithm \cite{handle\_occlusion2001}. Due to the issues such as illumination difference and occlusion, their estimated depthmaps are of low quality.

\citet{InternetScaleMVS\_CVPR2010} use structure from motion (SFM) to purge redundant imagery but retain high resolution geometry. Their iterative clustering 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. \citet{Goesele07} address the viewpoint selection for crowd sourced imagery by building small-size image clusters using the cardinality of the set of common features among viewpoints and a parallax-based metric. This image-wide selection may not be robust to outlier camera pose estimates. After this, 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.

\subsection{Efficiency}

Efficiency is an important issue in depthmap estimation. Traditional methods on large baseline stereo generally involve exhaustive evaluations of a large number of depth hypotheses. The high complexity of computation is not only time-consuming \cite{yang2003multi,CombinedDepthOutlier,Gallup07,LeastCommitment\_3DIMPVT2012}, but also prohibitive in less powerful devices such as smart phones and tablets.

The recently proposed PatchMatch technique is an efficient sampling scheme. Though the scheme has no strict theory or proof of its working mechanism, it has been empirically shown that it always works very well in practice. The PatchMatch was firstly proposed to find approximate nearest neighbor matches between image patches in \citet{Barnes:2009:PAR}, and later \citet{patchMatchStereo1} introduce it to solve the two-view stereo problem. PatchMatch initializes each pixel with a random slanted plane at a random depth, and is followed by the propagations of different directions. The nearby and the current pixels' slanted planes are tested and the one with the best cost is kept. \citet{patchMatchStereo2} combine the PatchMatch sampling scheme and belief propagation to infer an MRF model that contains smoothness constraints.

By combining guided filter and PatchMatch, \citet{PMF\_Hongsheng} provide an efficient edge-aware filtering for correspondence field estimation, which can be applied in two-view stereo.

While the original PatchMatch stereo was a sequential method, \citet{patchMatchParallel} parallelize the algorithm by restricting the propagations to only horizontal and vertical directions. Our research further explore the potential of the PatchMatch in wide baseline stereo with large hypotheses space.

\subsection{Point Cloud and Mesh Generation}

By far, we only discuss the works focusing on depthmap estimation. Some other methods aim at generating a consistent 3D model (either point cloud or mesh) instead of depthmaps.

\citet{FURUKAWA\_PAMI2010} aim at reconstrucing a quasi-dense point cloud by densifying the sparse 3D points. They present an accurate Patch-based multiview stereo approach that starts from a sparse set of matched keypoints, which were repeatedly expanded until visibility constraints are invoked to filter out false matches. \citet{Zaharescu\_PAMI2011} propose a mesh evolution framework based on a new self-intersection removal algorithm.

A typical way for 3D mesh generation is to fuse the depthmaps into a consistent model by leveraging the redundant information across the depthmaps. \citet{gallup20103d,gallup2010heightmap} develop heightmap-based fusion methods that work well for planar object surfaces such as building facades. \citet{zach2008fast} tackles the surface reconstruction task in a variational formulation.

Given that all these methods are volumetric-based and hence memory inefficient, \citet{zheng2012efficient} instead propose to compress the volume of interest by Haar wavelet, and hence much less memory is required. \citet{JAN} propose a method that reconstructs surfaces that do not have direct support in the input 3D points by exploiting visibility in 3D meshes. Their method has been shown to work pretty robustly on the textureless regions.

\section{Dynamic Object Reconstruction}

The following sections outline the related works of trajectory triangulation, image sequencing, articulated object reconstruction, non-rigid structure from motion (NRSFM), and single-view reconstruction.

\subsection{Trajectory Triangulation}

\citet{avidan2000trajectory} first coined the task of trajectory triangulation, which is defined as reconstruction of a moving point from monocular images. That is, each dynamic point is observed only by one camera at a time. Their method assumes the dynamic point moves along a simple parametric trajectory, such as a straight line or a conic section. Apparently, this is a rather strict constraint to impede its application in real scenarios.

In contrast, some other methods \cite{Park\_ECCV2010,Valmadre\_CVPR2012,ZhuCL\_CVPR11,park20153d} focus on a more general model by only assuming a smooth motion of dynamic objects.

\citet{Park\_ECCV2010} represent the trajectory with a linear combination of low-order discrete cosine transform (DCT) bases, and the trajectory is triangulated by estimating the coefficients of the linear combination. There are two fundamental limitations of their method as observed by \citet{Valmadre\_CVPR2012}. First, there is no automated scheme to determine the optimal number ($K$) of DCT bases. Second, the correlation between the object trajectory and the camera motion inherently limits the reconstruction accuracy. To overcome the first limitation, \citet{park20153d} select $K$ by checking the consistency of the reconstructed trajectory in an N-cross validation scheme.

Alternatively, \citet{Valmadre\_CVPR2012} propose a new method without using DCT bases. They estimate the trajectory by minimizing the trajectory's response to a bank of high-pass filters. To overcome the second limitation, \citet{ZhuCL\_CVPR11} propose to incorporate the 3D structures of a number of key frames to enhance the reconstructability. However, obtaining those key-frame 3D structures requires manual interaction. All the methods \cite{Park\_ECCV2010,Valmadre\_CVPR2012,ZhuCL\_CVPR11} require the sequencing information of the images, but in natural capture setups, the availability of sequencing information and high reconstructability typically cannot be fulfilled simultaneously \cite{ZhuCL\_CVPR11,park20153d}.

\subsection{Sequencing and Synchronization}

Sequencing information is important in trajectory triangulation. Recently,

\citet{Basha\_ECCV2012, Basha\_ICCV2013} target the problem of determining the temporal order of a collection of photos without recovering the 3D structure of the dynamic scene. The method by \citet{Basha\_ECCV2012} relies on two images taken from roughly the same location to eliminate the uncertainty in the sequencing. \citet{Basha\_ICCV2013} later introduce a solution that leverages the known temporal order of the images within each camera. Both of these methods assume that dynamic objects move closely to a straight line within a short time period, but in practice, points can deviate considerably from the linear motion model, especially when the temporal discrepancy between images is large.

Video synchronization has attracted much attention in the computer vision community \cite{Tuytelaars\_CVPR,shrestha2010synchronization,rao2003view}. Those methods have various constraints such as camera motion, availability of sound, and number of videos.

\subsection{Articulated Object Reconstruction}

Trajectory triangulation suffers from the reconstructability problem of inaccurate reconstruction if the camera motion is relatively smaller than the object motion \cite{park20153d}.

In the case of 3D reconstruction of articulated objects, we can enforce an additional constraint that the distances between joint points (according to the topology) are fixed, which helps to reduce ambiguities in reconstruction.

Based on the previous work by \citet{Park\_ECCV2010}, the authors further reconstruct 3D articulated motion with the constraint that a trajectory remains at a fixed distance with respect to its parent trajectory \cite{Park\_ICCV2011}. Their work shows the improvement of the reconstructibility over their earlier approach \cite{Park\_ECCV2010}.

However, the formulation involves solving an NP-hard quadratic programming problem, which is intractable in the case of a large number of input images.

To conquer the limitation, \citet{Valmadre\_ECCV2012} develop a dynamic programming approach that is guaranteed to solve the problem in a timely manner.

As opposed to articulated object reconstruction, our research focuses on reconstructing more general dynamic objects.

\subsection{Non-rigid SfM}

One class of related works solve the non-rigid structure from motion (NRSFM) problem, which targets simultaneous recovery of camera motion and 3D structure using an image sequence. These methods typically start from a set of 2D correspondences across frames. As an important extension of the well-known Tomasi-Kanade factorization \cite{tomasi1992shape}, \citet{Bregler\_CVPR2000} tackle the NRSFM problem through matrix factorization, with the assumption that deforming non-rigid objects can be represented by a linear combination of low-order shape bases. It is later shown by \citet{Xiao\_ECCV2004} that utilizing only orthogonality constraints on the camera rotation is not enough, and a basis prior is required to uniquely determine the shape bases. However, \citet{akhter2009defense} discover that in spite of the inherent ambiguity in the shape bases, the 3D shape itself can be uniquely recovered without ambiguity. Not until very recently, \cite{dai2014simple} propose a new prior-free method that estimates the shape matrix without explicitly recovering the shape bases, which is achieved by minimizing the rank (nuclear norm) of the shape matrix.

As a dual method to the above shape-based methods, \citet{Akhter\_NIPS08} propose the first trajectory-based NRSFM approach, which leverages DCT bases to approximately represent point trajectories. While shape-based approaches typically do not require sequencing information, trajectory-based approaches completely fail if image frames are randomly shuffled \cite{dai2014simple}.

At first glance, it seems that the NRSFM problem targets a more complete problem than the trajectory triangulation problem since the former additionally assumes unknown camera poses.

However, these approaches assume orthographic or weak perspective camera models, and it has been shown empirically that the extension of these methods to the projective camera model is not straightforward \cite{Park\_ECCV2010}. There are works for projective non-rigid shape and motion recovery based on tensor estimation \cite{hartley2008perspective,vidal2006nonrigid}, but this challenging problem is still under on-going research. Moreover, the NRSFM methods only recover the shape of the object without absolute translation due to the inherent ambiguity arising from the unknown shape translation and the unknown camera translation.

\subsection{Single Image Reconstruction}

While trajectory triangulation and NRSFM methods estimate 3D points from an image sequence, some other works target the problem of 3D reconstruction from a single image. Since there is only one shot of the object, the object motion, either static or dynamic, becomes irreverent for the reconstruction.

Some works focus on 3D reconstruction of Manhattan world \cite{Coughlan\_ICCV99}, which is defined as

man-made scenes with mainly orthogonal facades. In this scenario, 3D reconstruction from a single image can be simplified to finding the 3D lines and planes within the scene. The work by \citet{Delage\_ISRR2005} uses 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. \citet{Ram\_ICCV2013} reconstruct the 3D lines in a Manhattan scene from a single image using linear programming that identifies a sufficient minimal set of least-violated line connectivity constraints.

There are other approaches mainly relying on supervised learning. \citet{Hoiem\_CGRAPH2005} label the image regions as ground, vertical, and sky with a pre-trained classifier, and then ``cut and fold" the image into a pop-up model like children's pop-up books. The method is limited to the application of outdoor scenes containing simple ground and vertical structures. \citet{Saxena\_IJCV2008} propose a method for computing a depthmap from a single still image by using a hierarchical multi-scale Markov Random Field (MRF) that incorporates several features. The features are manually designed and the parameters of the MRF model are trained using ground truth depths. Instead of manually choosing features, \citet{eigen2014depth} recently propose to estimate the depthmap of a single image by employing two deep network stacks: one that makes a coarse global prediction based on the entire image, and another that refines this prediction locally. Due to its wide applications, single depthmap estimation using learning is currently still an active research topic.