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

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

\chapter{Introduction}

Imagery records how the world looks like by projecting the 3D scene into the image plane. However, the 3D information, which depicts the geometry of real objects, is lost during this capturing process. Conversely, the 3D information is key to many applications, such as augmented/virtual reality \cite{ventura2008depth}, robots and autonomous car navigation \cite{endres2012evaluation}, image-based rendering \cite{View\_interpolation1993} and image enhancement \cite{zhang2014personal}. Moreover, as additional information to RGB colors, 3D information is leveraged to improve performance of many computer vision tasks such as object classfication/recognition \cite{gupta2013perceptual} and human pose estimation \cite{CVPR\_kinect}. Therefore, there is a strong desire to recover reliable 3D information from 2D imagery.

3D information, when saved in computers, can be represented using 3D point cloud, 3D Polygon meshes or depthmaps. 3D point cloud is a set of data points in three-dimensional space representing the external surface of an object, and it can be classified as either dense or sparse based on its density. In contrast, 3D polygon mesh provides additional information of geometric topology among the 3D points. Comparing to these representations using values in a global coordinate system, a depthmap is a dense field of depth values indicating the distance of the observed surface relative to a camera. In practice, different representations are adopted based on the applications.

3D reconstruction from imagery, defined as a process that recovers 3D information from 2D image colors, is a traditional problem in 3D computer vision. Unlike the task of computer graphics that renders 2D imagery from 3D geometry, the inverse process of 3D reconstruction from imagery is more challenging since aiming at recovery of lost information inevitably introduces more ambiguities. Though 3D reconstruction has received heavy studies and improved on over the last few decades, it still remains a viable research area. This dissertation primarily focuses on the problems of dense static scene reconstruction and sparse dynamic object reconstruction from uncontrolled imagery.

\textbf{Dense static scene reconstruction}.

To attain the 3D information of a static scene, most of the existing works leverage 2D correspondences and available camera parameters for 3D triangulation. Though camera parameters can normally be estimated by structure from motion or offline calibration, obtaining 2D correspondences robustly from image colors still requires further exploration. The 2D correspondences are defined as pixels in different images that observe the same part of a 3D scene. Under the assumption of Lambertian surface, these 2D correspondences share the same or similar appearances/colors, and hence they have high color consistency.

For each point in one image, finding its correspondence in another image is attained by searching for candidate pixels with best color consistency along a line defined by 3D geometry (called epipolar line), and the positions of candidate pixels are determined by depth hypotheses generated in a valid range. Once the correspondence is found, the depth of the corresponding pixel is uniquely determined. However, estimating dense correspondences robustly is difficult since ambiguities arise in the cases of repetitive textures, homogeneous color regions, or occlusions along the epipolar line.

Recently, there is a growing interest in using the ever-increasing crowd sourced data (\ie~Internet collected photos) for reconstruction, since the large number of free data has intrigued many applications such as photo tour \cite{Snavely2} and image enhancement \cite{zhang2014personal}. With the non-controlled imagery as input, finding 2D correspondences based on colors is more challenging due to a diversity of factors such as heterogeneous resolution and scene illuminations, unstructured viewing geometry, scene content variability and image registration errors. To address these issues, it is normally assumed in the massive number of images, there are a subset of images sharing similar image characteristics, and then determining a suitable subset of images or pixels for correspondence search becomes essential \cite{Goesele07}.

Dense reconstruction typically has very high computational complexity, since in tradition the process involves exhaustive evaluations of a large number of depth hypotheses \cite{yang2003multi}. The increasing availability of crowd-sourced datasets has explicitly brought efficiency and scalability to the forefront of application requirements. Moreover, the high complexity of a method would impede its usage in less-powerful electronic devices such as smart phones. To this end, there is a compelling demand to develop efficient and scalable methods for dense reconstruction.

\textbf{Sparse dynamic object reconstruction}.

The problem of dynamic object reconstruction specifically aims at 3D reconstruction under the circumstance of non-concurrent image captures. To be more precise, the dynamic object is only observed by one image at each time instance. This poses an additional challenge compared to the problem of static scene reconstruction, since 3D triangulation becomes invalid and impossible with the unitary observation, even assuming 2D correspondences among non-concurrent images are correctly found. Probably due to the intrinsic difficulty, the state of the art for dynamic object reconstruction falls far behind that of static scene reconstruction.

The problem of dynamic object reconstruction is fundamentally under-constrained, and cannot be solved without further assumptions. To solve this problem, many existing works make various assumptions on scene geometry, object motion, capture pattern, \etc For instance, most non-rigid structure from motion (NRSFM) methods assume the 3D shapes of deforming objects lie in a low dimensional subspace, and hence any shape can be represented as a linear combination of $K$ shape bases \cite{Bregler\_CVPR2000,torresani2008nonrigid,dai2014simple}. Trajectory-based methods assume smooth motion of the dynamic objects across time \cite{Akhter\_NIPS08}. \citet{bao2011toward} assume non-coplanar layout of objects and cameras, plus a prior knowledge of the object size.

\citet{Hoiem\_CGRAPH2005} assume the scene only consists of the elements of sky, ground plane, and vertical buildings facades.

When developing methods for dynamic object reconstruction, in addition to making a valid assumption, having less but general assumptions is vital to enable the methods to work more universally and robustly in real scenarios.

For the problem of dynamic object reconstruction from a monocular image sequence, it is typically assumed smooth object motion. This is a reasonable assumption given the typical objects motions. To reconstruct the dynamic objects under this assumption, it is important that sequencing information is available. The sequencing information is defined as the temporal order of images being taken. It tells that a moving 3D point is spatially close, if the temporal stamp of two capturing images are temporally close. Alternatively, we can also define the sequencing information in terms of spatial proximity. \emph{In effect, it is this spatial proximity that is leveraged by most existing methods for dynamic object reconstruction.} However, this sequencing information is typically unavailable among crowd sourced images or videos. While most of existing works assume such information being available \cite{Park\_ECCV2010,Park\_ICCV2011,Valmadre\_ECCV2012,Valmadre\_CVPR2012}, my research in dynamic object reconstruction skip this assumption by focusing on joint estimation of image sequencing and 3D points.

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

\section{Thesis Statement}

Robust and efficient depthmap estimation for Internet collected photos can be attained through a framework of joint view selection and depthmap estimation, and accurate dynamic object reconstruction from unstructured images or unsynchronized videos is achieved by joint estimation of image sequencing and 3D points.

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

\section{Outline of Contributions}

This dissertation contributes significantly to advance the state-of-the-art techniques for the problems of static scene reconstruction and dynamic object reconstruction, and it builds on our published works \cite{zheng2014patchmatch,zheng2014joint,zhengiccv\_2015}.

\begin{description}

\item[PatchMatch Based Joint View Selection and Depth Estimation:]

Chapter \ref{ch:patchmatch} focuses on the problem of depthmap estimation using the Internet collected photos. The non-controlled input imagery presents the practical challenges such as heterogeneous scene illuminations and unstructured viewing geometry.

Therefore, it is vital to determine a subset of images or pixels in the dataset for robust depth estimation. Moreover, the ever-increasing crowd-sourced datasets have explicitly brought

efficiency and scalability to the forefront of application requirements.

To solve this problem, we propose a probabilistic framework for joint view selection and depth estimation at the pixel level. Our new method attains more complete depthmaps comparing to the state-of-the-art method for Internet collected photos \cite{Goesele07}.

To increase the efficiency and scalability, our framework seamlessly incorporate the PatchMatch scheme \cite{patchMatchStereo1} to reduce the size of depth hypotheses set. Also, the memory requirement of our framework scales linearly with respect to the number of source images, as opposed to exponentially \cite{CombinedDepthOutlier}.

Moreover, the method is designed to process each row or column of the reference image independently for easy parallelization and GPU implementation.

\item[Joint Object Class Sequencing and Trajectory Triangulation:]

Chapter \ref{ch:jost} targets the problem of reconstructing the 3D positions of dynamic objects from a set of unstructured images. Each dynamic object is observed by the images once so that the traditional 3D triangulation for static scenes is impossible. To tackle the fundamentally under-constrained problem, we assume all the objects of the same class (\eg~pedestrians or cars) move in a common path in 3D space. Then our method estimates the 3D positions of the dynamic objects by triangulating the trajectory formed by all the objects moving in the common path.

To the best of our knowledge, no current methods have solved this challenging problem.

Our method uses the object detection outputs as a general feature for each dynamic object, as opposed to typical images features such as points or edges. In solving the problem, recovering the sequencing information, which is defined as the topology of the trajectory in this specific problem, is vital for trajectory triangulation. We propose to jointly estimate the sequencing information and the 3D points, which is posed as minimizing a nonconvex function. To minimize the function, we propose a novel discrete-continuous optimization approach based on the generalized minimum spanning tree (GMST).

\item[Dynamic Object Reconstruction from Unsynchronized Videos:]

Chapter \ref{ch:video\_l1} aims at the similar problem of dynamic object reconstruction, but using unsynchronized video streams as input. To handle this unconstrained problem, we observe that any shape at one time instance is a linear combination of other shapes (self-expression), under the assumption of smooth object motion. The problem is then solved by learning a self-expressive dictionary, which is defined as a collection of temporally varied structures.

The main contribution of this paper is solving the new problem of dynamic object reconstruction without temporal order information across video streams (also called sequencing information). This is contradictory to the existing works that strictly rely on available sequencing information \cite{Park\_ECCV2010,Valmadre\_CVPR2012}.

Moreover, to the best of our knowledge, we are the first to use self-expression prior for the dynamic object reconstruction. This prior has the potential to be used in the traditional non-rigid structure from motion (NRSFM) problem, where most existing methods use the assumption that any shape is a linear combination of $K$ fixed shape bases \cite{dai2014simple,Bregler\_CVPR2000}. In learning the dictionary, we propose an efficient solver based on the alternating direction method of multipliers (ADMM) \cite{boyd2011distributed}.

\end{description}

Each of these contributions addresses the issue of 3D reconstruction from 2D imagery. The next three chapters describe each method in detail, and Chapter \ref{ch:discussion} conclude the dissertation with potential extensions to our works and possible future research directions.