Geometry-aware methods for efficient and accurate 3D reconstruction

Thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy in Computer Science and Engineering

by

Rajvi Shah
201399663
rajvi.shah@research.iiit.ac.in

International Institute of Information Technology
(Deemed to be University)
Hyderabad - 500 032, INDIA
December 2020

To

Friendships

International Institute of Information Technology

Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Geometry-aware methods for efficient and accurate 3D reconstruction" by Rajvi Shah, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Adviser: Prof. P J Narayanan

Acknowledgments

I sincerely thank my advisor Prof. P J Narayanan for his technical guidance and for providing me invaluable opportunities for teaching and mentorship, attending workshops and conferences, and research collaborations. I am thankful to him for believing in me and for extending care and unwavering support through difficult times. I am thankful to the Google India PhD Fellowship program for providing me generous financial support and Vivek Kwatra for mentoring me during my internship at Google. I also thank Chris Sweeney for his technical help during my PhD. I am thankful to my co-authors Aditya (Deshpande), Vanshika, Aditya (Singh), Ishit, Saumya, Siddhartha, Saurabh, Parikshit, for their technical collaboration and for burning the midnight oil with me during paper submission deadlines.

During the course of my PhD, I have come to know many incredible people who have contributed greatly to my professional and personal growth. In addition to my collaborators, I also thank Srinath, Nitish, Mihir, Udit, Vidyadhar, Kaustav, Pramod, Lipika, Parijaat, Tejaswinee, and other lab-mates who I met during the early years of my PhD, for their technical help and for creating a vibrant and joyful place of work. I would like to thank my fellow PhD students in CVIT: Anand, Nataraj, Yashaswi, Praveen, Aditya (Arun), Suranjana, Sourabh D, Vijay, Arunava, Aniket, Pritish, for technical and academic help, for organizing and participating in study groups, summer schools, and tech talk series, and overall for fostering a great environment for learning and collaboration; Parikshit, Revanth, for passing on passion for Graphics and Computer Games, for many brain-storming sessions, and for starting fun lab rituals. I cherish our many board-game sessions, car rides, impromptu chess challenges, and lunch outings.

I am utmost thankful to Saurabh (Saini), my philosopher friend, for being (literally and figuratively) by my side all these years, but most importantly for philosophizing, even to my protesting ears. I thank Aniket for teaching me to to be strong and authentic and to look out for the vulnerable. I am thankful to Pritish for lending me his unwavering optimism for life and for supporting me through adversarial times. Tejaswinee, my musical friend, I am thankful to her, for being my remote source of love, support and my voice of reason, for inspiring me to persevere and find joy.

I thank my parents and sisters for having faith in me and for their silent support, Dinesh and his family for their love and support in early years of my PhD, Shilpi and Anu for being my local sisterhood. I thank my therapist Ms. Sandhya Luther for teaching me coping skills and for helping me remain functional through some difficult times.

I thank CVIT and IIIT admin staff for helping me with paperwork and numerous logistical requests. I also thank the lab and hostel housekeeping staff, mess workers, and canteen staff for their tireless work that eased my everyday life for all these years. I am also greatly thankful to all the CVIT and IIIT faculty members and larger IIIT community for creating an atmosphere of scientific and academic fervor and for providing a safe space for expression and exploration of progressive thought and ideals.

For all the people I have mentioned here, there are many more people who I may have missed, who I am thankful to for directly or indirectly helping and supporting me through my PhD journey and in my time at IIIT.

Abstract

Advancements in 3D sensing and reconstruction has made a huge leap for modeling large-scale environments from monocular images using structure from motion (SfM) and simultaneous localization and mapping (SLAM) algorithms. SfM and SLAM based 3D reconstruction has applications for digital archival and modeling of real-world objects and environments, visual localization for geo-tagging and information retrieval, and mapping and navigation for robotic and autonomous driving applications.

In this thesis, we address problems in the area of large-scale structure from motion (SfM) for 3D reconstruction and localization. We introduce new methods for improving efficiency and accuracy of state-of-the-art pipeline for structure from motion. Large-scale SfM pipeline deals with large unorganized collections of images pertaining to a particular geographical site. These image collections are formed by either retrieving relevant images using textual queries from the Internet, or can be captured for the specific purpose of 3D modeling, mapping, and navigation. Internet image collections tend to be more noisy and present more challenges for reconstruction as compared to datasets captured with specific intention to reconstruct. In this thesis, we propose methods that help with organizing these large, unstructured, and noisy images into a structure that is useful for SfM methods, a match-graph (or a view-graph). We first propose a geometry-aware two stage approach for pairwise image matching that is both more efficient and superior in quality of correspondences. We then extend this idea to SfM pipeline and present an iterative multistage framework for coarse to fine 3D reconstruction. Finally, we suggest that a key to solving many of the reconstruction problems is to address the problem of filtering and improving the view-graph in a way that is specific to the underlying problem. To this effect, we propose a unified framework for view-graph selection and show its application to achieve multiple reconstruction objectives.

Contents

List of Figures

List of Tables

Chapter 1

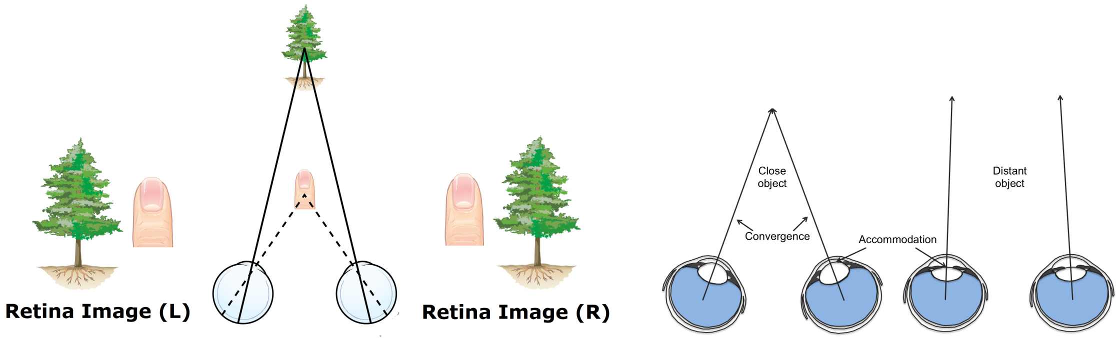
Introduction

1.1 Historical Perspective

Seeing the world in 3D Understanding depth perception was an important curiosity for early vision scientists and much work has been done to illuminate how the human visual system perceives depth [1, 3, 14]. While biological understanding of human visual system informs that stereoscopic vision is primarily responsible for 3-dimensional (3D) perception, studies in perceptual psychology improved our understanding of both monocular and binocular cues associated with 3D perception. Mainly two binocular phenomena are known to be effective for depth perception, (a) retinal disparity, and (b) convergence. These cues are depicted in Figure 1.1.

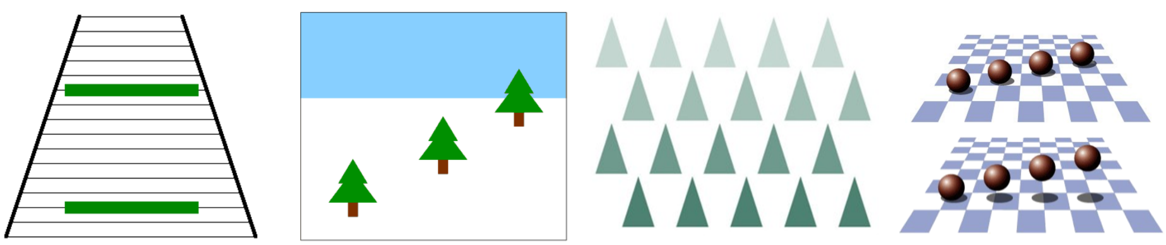
Retinal disparity or binocular disparity is the observation that both eyes form a slightly different picture of the world on the retina. The difference between positions of an object in these images (disparity) is what allows the human brain to infer the distances of the objects from each other (or relative depth) in the 3D world. Convergence is an oculomotor phenomenon related to depth sensing. To look at a near object, both eyes converge while looking at a distant object, the eyes diverge. The sensation due to the movement of extra-ocular muscles helps in absolute depth perception.

In addition to the binocular cues, through the process of habituation and learning, the human brain can also perceive depth using several monocular cues. Examples of a few monocular cues are shown in Figure 1.2. Linear perspective is one of the strongest monocular cues. We expect the parallel lines to meet/vanish at the horizon, giving us clues to the size and distance of the objects with respect to the perspective. This is illustrated by Ponzo illusion in Figure 1.2a, where the top green bar appears much wider despite being the same in size as the bottom ones. We are attuned to give importance to the horizon line and hence objects placed higher up in the composition appear farther to us (Figure 1.2b).



(a) Binocular Disparity (b) Convergence

Figure 1.1: Binocular cues for depth perception : (a) Extending a finger to mask a distant object (tree) and looking with only one eye open at a time, parallax / retinal disparity can be observed in relative placement of the finger and the tree. (b) To focus on nearby objects, eyes converge as opposed to distant objects. The activation of extra-ocular muscles provide a strong cue about object distance.



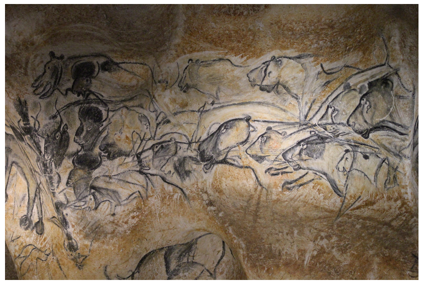
(a) Linear Perspective (b) Relative height (c) Atmospheric Perspective (d) Shadows

Figure 1.2: Monocular cues for depth perception

Atmospheric perspective makes farther objects appear hazier and bluer (Figure 1.2c). One of the most important cue for monocular depth sensing is shadows (Figure 1.2d). In addition to these, several other monocular cues such as the interposition of objects, the relative size of objects, texture gradient, accommodation, motion parallax play a vital role in depth perception. These cues are well studied in the psychology of vision [4, 90]. I refer the reader to Chapter 6 of the text by LaValle [83] for further reading on depth cues.

Early artists were aware of these cues and often used these to push realism and immersion of paintings, 2D representations the 3D world. Figure 1.3 shows some examples of paintings and artworks that leverage various depth cues for realism. The human brain is so habitual to interpret monocular cues from the context that for hundreds of years artists and creators successfully used flat representations (paintings, photographs, films) to tell stories about our 3D world.

(a) Landscape (∼ 1300)



(b) Chauvet Cave Painting (30,000 BCE)

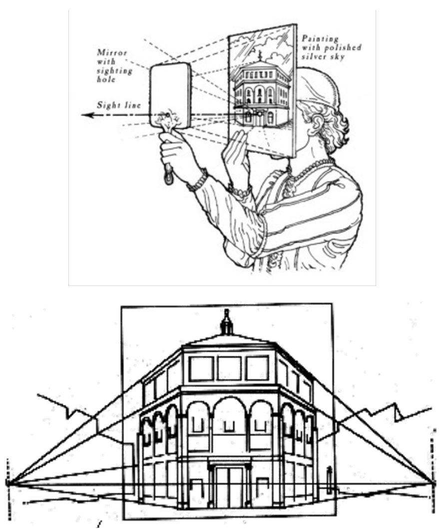
(c) 'Madonna of the Magnificat' (1481)

(d) 'Skiffs' (1877)

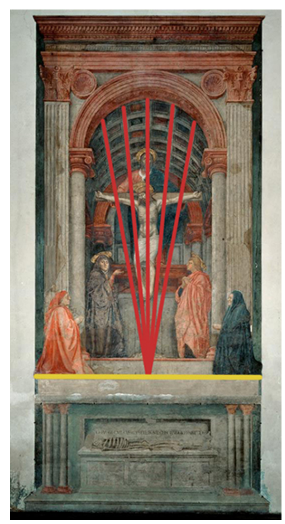
(e) 'A Rainy Day' (1877)

Figure 1.3: Use of monocular depth cues in paintings from different periods : (a) atmospheric perspective, (b–c) occlusion, interposition, & shadows, (d) relative size & height, (e) texture gradient.

(a) 'The Carrying of Cross' (1333)



(b) Brunelleschi's perspective aparatus (1415)



(c) 'Holy Trinity' (1427)

Figure 1.4: Early Renaissance paintings carried a flat look. In (a), perspective is inaccurate and the figures in the distance and the front are depicted to be of the same size. (b) shows Filippo Brunelleschi's apparatus to discover rules of drawing with perfect perspective. Brunelleschi drilled a hole into his painting of the Baptistery and observed its reflection in a mirror that was held between the painting and the Baptistery. As the mirror was moved into and out of view, the observer could see the geometric similarity between the actual view and the reflected painting. (c) shows first painting that uses Brunelleschi's discovery. The red overlay highlights perspective lines meeting in a vanishing point.

Figure 1.5: Difficulty of computer vision tasks.

Making the computers see in 3D The idea of 'making a computer see', or computer vision as we know it, would not be practical until the late 1980s as cameras were far from a common commodity and computing power was fairly limited for processing images. The early computer vision programs had simpler goals of extracting information about shapes of 3D objects from 2D images [2]. These early methods were highly influenced by the perceptual understanding of the human visual system. Computer vision was thought to be a much simpler problem, object recognition started as a summer project in the sixties. While the holy grail of computer vision remains a very challenging problem (see Figure 1.5 for humor), scientists have made tremendous progress in appearance based understanding of geometry in computer vision. The field has made rapid progress in the last decade due to advancements in robust algorithms for depth estimation, the proliferation of Internet images pushing the edge of data-driven methods, and the availability of consumer hardware for 3D sensing and capture.

As discussed previously, human visual perception leverages a variety of cues to infer the 3D structure of the surrounding environment. Similarly, many methods are modeled in geometric computer vision that leverage a variety of cues in 2D images to infer depth. These methods are based on mathematical modeling of the imaging process that relates the 3D world and their 2D projections. Figure 1.6 depicts the standard imaging process involving a light source, an object, and the camera capturing the object. Understanding the forward imaging process, i.e. the interaction of light, materials, and the camera model allows us to find ways to solve the inverse problem of going to the world from the image. The larger class of algorithms falls under the umbrella of shape from 'x' methods, where 'x' can be any depth cue such as shading, motion, defocus, texture, etc. The output of these methods is typically a pixel-wise depth map corresponding to the point-of-view of one of the images. Another way of obtaining 3D information is through active range sensing. While camera sensors only capture reflected light, an active sensor is equipped with a laser source that typically sends out a pulse, and the sensor then detects that pulse's reflection and records its time of flight. There are other active sensing methods also like light strips, Kinect, etc. An excellent survey of state-of-the-art depth-sensing methods and their applications

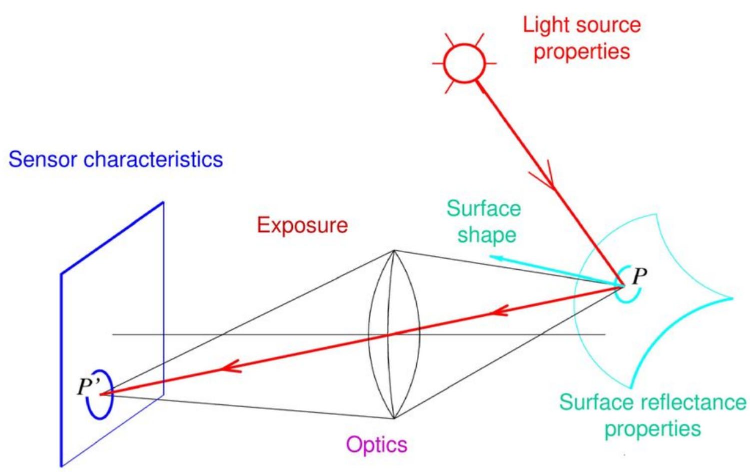


Figure 1.6: Depiction of basic imaging process : The light reflected from the surface of the object enters camera through the lens and forms an image on the sensor plane. This process is characterized by the properties and positions of the light source, the object, and the camera (optics).

can be found here [117]. The output of these methods is typically a point cloud corresponding to the point-of-view of the sensor. We refer the reader to Chapter 12 of the textbook on Computer Vision by Szeliski [47] for a broader overview of these methods.

The last decade has seen exponential progress of data-driven methods for reasoning and inference tasks in computer vision. While primarily the focus of the inference side of computer vision was more centered on tasks such as segmentation, detection, classification, recently with the rise of large datasets and deep learning networks, supervised and unsupervised depth estimation from a single image is nicely cast as an inference problem [110, 81, 103].

1.2 Image-based 3D modeling and applications

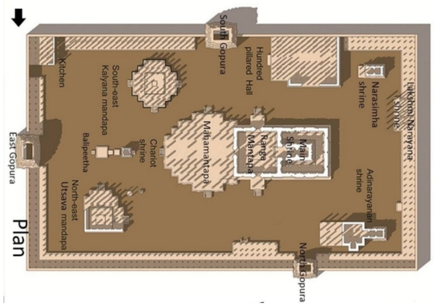
Pictures are powerful devices to capture a moment in time of the world around us. However, an image does not provide the same experience as physically being present at a scene. Being present in a 3D world allows us to look around, walk/move around, and navigate the environment, observe the physical world from different viewpoints. While capturing depth from a single view-point is one step closer to capturing reality, the goals of 3D modeling and reconstruction are to capture and represent the 3D world around us as faithfully as possible. This has always been an important problem in Photogrammetry and Computer

Figure 1.7: A collection of photos downloaded from Flickr with geotags – ‘Colosseum, Rome’ (left); SfM-based 3D reconstruction of Colosseum with point-cloud and camera positions (right).

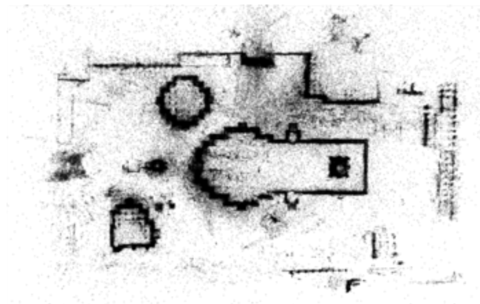
Vision and one of the most successful approaches to do this with passive sensors (cameras/images) fall under the category of Structure from Motion (SfM) algorithms.

Given a number of images of a scene captured from different viewpoints (either using a single freely moving camera or multiple cameras located at different positions and orientations), SfM aims to recover the structure of the scene as well as poses of the cameras. The recovered structure is a sparse representation of the scene, a point/line cloud based on the 2D primitives used for matching across images. Figure 1.7 shows an example of SfM-based 3D reconstruction of Colosseum in Rome, Italy. The point clouds represent features visible in each image whereas the wire-frame pyramids represent cameras that have taken the pictures. These rich 3D descriptions and camera information are useful for a variety of applications.

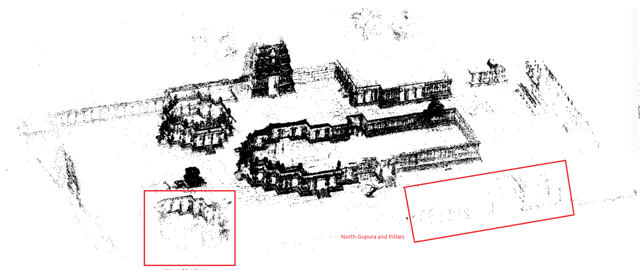
Virtual reality and Visual effects The ability to create synthetic representations (computer-generated 3D models) of the world and pushing the realism of these computer-generated imageries (CGI) has always been a fascination of computer graphics community. The advancements in CGI has transformed the way our films are made, the stories we can tell, and the holy grail of computer graphics, what early researchers called ‘virtual environments’ and what we now call ‘virtual reality’. The algorithms for image-based 3D modeling have made a huge impact on virtual reality and visual effects industry. Video-based rendering techniques have been successfully employed for media production. In visual effects (VFX) industry, match-moving is a very popular technique for inserting effects into video footage in a realistic manner. At the core of match-moving is camera tracking algorithms essentially based on SfM. Figure 1.8 shows match-moving in action.



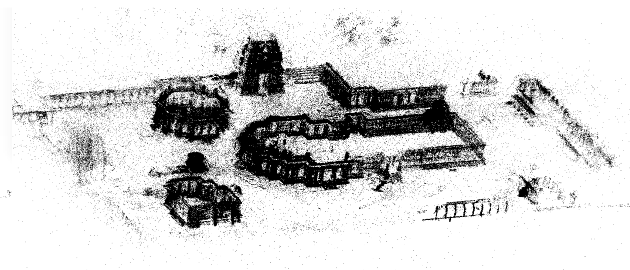
HVT floor plan



top-view of the full reconstruction

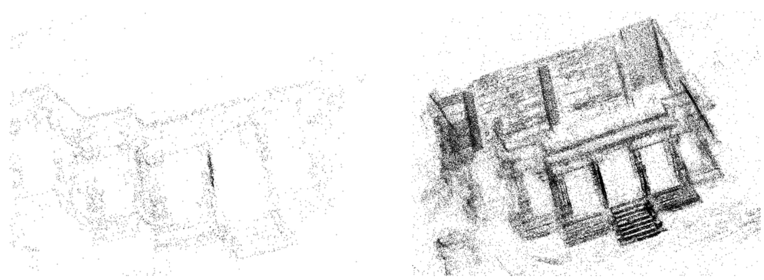


coarse reconstruction



full reconstruction (multistage pipeline)

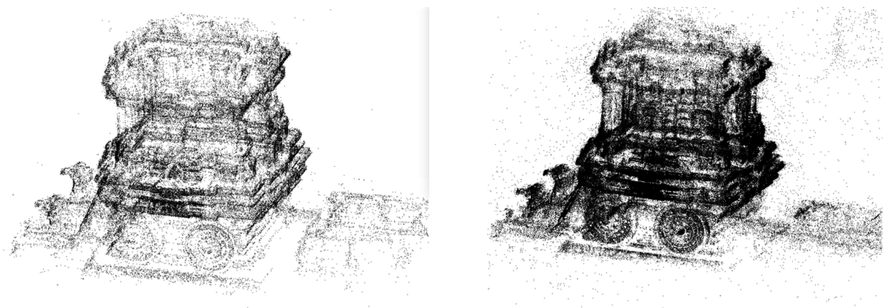
Renders of Hampi Vitthala Temple Complex (HVT) point clouds. Sparsely reconstructed parts in the coarse model are highlighted by red rectangles. Close up views of these parts are given below.



Utsava Mandapa : coarse (left), full (right)

North Gopura : coarse (left), full (right)

Sparsely reconstructed parts after coarse reconstruction are enriched after multistage pipeline



coarse (left) and full (right) reconstructions

Additional views of the full reconstruction with reference images

Selected renders of the stone chariot structure of HVT

Selected renders of the Maha Mandapa (main shrine) structure of HVT

Figure 4.6: Point clouds of Hampi Vitthala Temple Complex reconstruction

Chapter 5

A framework for Optimal View-graph Selection

5.1 Introduction

In chapter 2, we saw that the structure recovery process consists of the two essential steps, (i) view-graph construction, and (ii) structure from motion (SfM). Given a set of images capturing a particular site, view-graph construction involves putative feature matching and geometric verification by estimating two-view epipolar geometry (EG). Nodes in the view-graph represent the images and edges between the nodes represent that the pairs are related by valid epipolar geometries. View-graphs help in 'organizing' unordered image collections useful to select, (i) a core set of images for reconstruction, and (ii) identify noisy EGs that might degrade the quality of reconstruction. State-of-the-art SfM methods like incremental [25, 68, 96], hierarchical [41, 33, 89], or global [45, 50, 51, 66, 65], all rely on a view-graph based initial step for ordering and pruning images for efficient and accurate reconstruction.

5.1.1 Sources of inaccuracy in large-scale SfM and common mitigations

Large community photo collections often display point-of-view bias, as some viewing angles are more popular than others. Using the full view-graph for SfM is often computationally expensive and unnecessary due to the high redundancy in the image set. Moreover, for closely clustered images, such narrow baseline pairs increase the uncertainty of triangulation, causing large re-projection errors. While isolated erroneous EGs can be overcome by robust averaging (global SfM) and repeated bundle adjustment (incremental SfM), a large number of incorrect/ill-conditioned EGs can result in incorrect reconstruction. Hence, it is crucial to select 'good' images and more importantly 'good' pairs for accurate reconstruction.

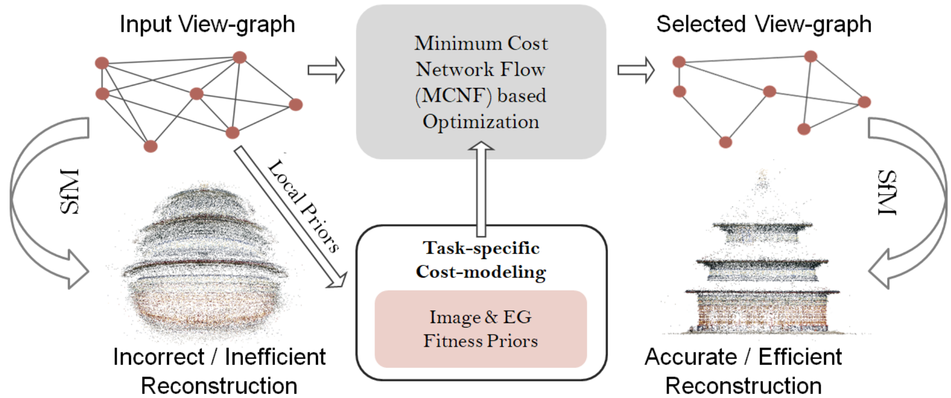


Figure 5.1: Outline of the proposed selection framework. With appropriately modeled costs, the framework can select view-graphs that meet desired reconstruction objectives.

Implicit view-graph filtering in SfM methods Pruning undesirable images and pairs from the input view-graph is an implicit part of both incremental and global SfM methods and is often done by putting thresholds on various criteria. Accuracy of incremental SfM hinges on seed pairwise reconstruction and next best view selection. Wide baseline seed pair selection is ensured using criteria based on epipolar inliers (that don't fit a homography) [25, 28], or inlier ratios (to detect pairwise motion as planar, rotational, or general) [96]. For the next best view selection, triangulation angle, inliers, or correspondence distribution are used as criteria [25, 96]. Global SfM methods first use relative rotations to estimate global rotations followed by global translations estimation [45, 50, 66, 65, 79, 88]. In [50], to reduce the state space of camera parameters for MRF based estimation, relative twist, and unusual aspect ratios are used as view filtering criteria. Methods that use lie-algebraic averaging of relative rotations [18, 24, 48, 51, 66, 65, 88] often discard images with unknown calibration and handle outlier EGs with loop consistency checks [48, 51, 66] and robust cost functions.

While these heuristics work well for a large variety of datasets, they are insufficient to identify and remove consistent noise that arises in ambiguous scenes. Ambiguity in pairwise matching arises due to the fact that man-made structures often comprise of repetitions (windows, pillars, arches), symmetries (circular structures, identical facades), and large duplicate instances (minarets, domes). While standard SfM pipelines are robust enough to handle a large number of inaccurate EGs in isolation, for scenes with high ambiguity, such wrong EGs form large consistent sets, causing ghosting errors in form of misregistered cameras, 'phantom' structures, or structures incorrectly folded along symmetry.

SfM methods for Disambiguation Some researchers have proposed specialized solutions to handle such difficult scenes with local and global steps for 'reasoning' symmetry and ambiguity. Initial methods for disambiguation focused on inferring missing correspondences [32] and using it as a prior with global objectives of camera pose inference [53] or consistent view-graph expansion [59]. Other methods consist of social-network-principles-based track refinement [67], triplet-consistent graph matching and expansion [97], and geodesic manifold based ambiguity detection and correction [109]. A post-reconstruction approach to disambiguation uses back-projected 3D points to identify conflicting observations [74, 73]. While these methods show promising results on some very challenging datasets, they do not operate within the framework of traditional SfM pipelines.

Overall, in SfM literature, the view-graph selection problem is generally not dealt with in a unified sense. Instead, algorithms are often designed to meet specific objectives for subgraph selection. We believe that many of these challenges can be addressed under the same umbrella of 'selecting' a complete, consistent and noise-free view-graph to achieve accurate and efficient reconstruction.

5.1.2 Motivation of proposed work

The main motivation of this work is to formalize the core problem of view-graph selection and introduce a method to meet different subgraph selection objectives within a unified framework. We pose this as an optimization problem with the image and pair selection costs and propose a novel network-flow based formulation for its efficient approximate solution. This abstraction allows different objectives to be achieved by plugging in task-specific costs while keeping the overall selection and reconstruction framework the same. Figure 5.1 shows an outline of this framework. We mainly focus on the application of this framework for accurate and ghost-free reconstructions of highly ambiguous datasets and show how pairwise selection cost can be modeled using local priors, and also introduce a new context-based prior. Additionally, we also show its usefulness for accurate and efficient reconstruction of general large-scale Internet landmarks datasets with image and pairwise selection costs modeled using commonly used local heuristics.

The proposed framework brings greater flexibility and generalization to standard SfM pipelines and its application is not limited to the specified use-cases. A unified framework also lands optimal view-graph selection as a learning problem if and when task-specific ground-truth data becomes available.

5.2 Posing view-graph selection as optimization

Let the input view-graph be $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the set of vertices (nodes) $\mathcal{V}$ represents the images and the set of edges $\mathcal{E}$ represents the pairwise epipolar geometries (EGs). The goal is to select a subset of nodes $\mathcal{V}'$ and a subset of edges $\mathcal{E}'$ (a subgraph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$) that meets a desired objective. We denote the indicator variables for image selection as δ_i and pair selection as δ_{ij}. Here, each δ_i corresponds to the image vertex $v_i \in \mathcal{V}$ and each δ_{ij} corresponds to the pairwise edge $e_{ij} \in \mathcal{E}$. Intuitively, the selection problem can be represented as a minimization of the following form,

$$
\begin{aligned}
\arg\min \quad & f(\delta) = \sum_i q_i \delta_i + \sum_i \sum_j q_{ij} \delta_{ij} && (5.1)\\
\text{subject to} \quad & \sum \delta_i \leq N, \ \sum \delta_{ij} \leq M && (5.2)\\
& \delta_{ij} \leq \delta_i, \ \delta_{ij} \leq \delta_j && (5.3)\\
\text{where} \quad & \forall i \in \mathcal{V}, \ \forall (i,j) \in \mathcal{E}, \delta_i, \delta_{ij} \in \{0, 1\}, \\
& N < |\mathcal{V}|, \ M < |\mathcal{E}|
\end{aligned}
$$

Activation of indicator variables δ_i and δ_{ij} imply selection of corresponding image v_i and pair e_{ij}. q_i is the cost of selecting the view v_i and q_{ij} is the cost of selecting the edge e_{ij} and these costs assumed to be negative in the minimization sense. N and M indicate thresholds on the maximum number of selected images and pairs and ensure that the subgraph selection is non-trivial (eqn. 5.2). Selection of image pair (i, j) requires that both images i and j constituting the pair also must be selected (eqn. 5.3).

Modeling these costs appropriately can express different objectives for view-graph selection. Here, we discuss cost modeling for accurate and ghost-free reconstructions using a few relevant priors. However, this can be used to achieve other objectives using a variety of other priors known in the literature or even learned costs.

This minimization problem can be formulated as a binary integer linear program (BILP), which is NP-complete. To find an exact optimal solution, stock solvers for BILP use branch-and-bound like techniques to intelligently iterate through all possible solutions. However, owing to the NP-complete nature of the problem, it is infeasible to compute the exact optimal solution for many problems of our interest. In fact, in our experiments, the computation time for a branch-and-bound based ILP solver was very high even for toy-sized problems. A standard trick to achieve efficient solutions with some approximation is to use a linear programming (LP) relaxation with rounding to obtain an integer solution.

However, in preliminary experiments, the solutions obtained using this approach seemed too inaccurate to be useful. Furthermore, this formulation does not take into account the connectivity of the selected subgraph. To tackle both these problems, we pose view-graph selection as a minimum cost network-flow (MCNF) problem that guarantees us a binary solution in polynomial time and encourages connectivity in the selected sub-graph. The proposed MCNF formulation is explained in the next section.

5.3 View-graph Selection as MCNF Problem

The network in MCNF problems is a directed graph with at least one source and one sink nodes. We denote the network as $\mathcal{T} = (\mathcal{N}, \mathcal{A})$, where $\mathcal{N}$ represents the set of nodes and $\mathcal{A}$ represents the set of directed edges/arcs. Each edge (i, j) in the network has a cost c_{ij} associated to let across one unit of flow, and the cost incurred by an edge is proportional to the flow (x_{ij}) through it. Each edge also has a lower and an upper bound (l_{ij}, u_{ij}) on the amount of flow (x_{ij}) that can pass through it, known as capacity constraints (**c.c**). The source sends a certain units of flow that the sink node must receive. At all other nodes, flow must be conserved, i.e. the total incoming and outgoing flow must be equal. Let us denote the total flow as F and the remainder flow at a node i as b_i, then $b_i = F$ when i is source, $b_i = -F$ when i is sink, and $b_i = 0$ otherwise. These constraints are known as equal flow constraints (**e.f.c**). The MCNF problem is about sending the total flow from source to sink at a minimum cost, without violating the capacity and flow constraints. This minimization with flow and capacity constraints can be described as,

$$\text{minimize} \sum_{(i,j)\in\mathcal{A}} c_{ij}x_{ij} \tag{5.5}$$

$$\text{subject to,} \sum_{\{j:(i,j)\in\mathcal{A}\}} x_{ij} - \sum_{\{j:(j,i)\in\mathcal{A}\}} x_{ji} = b_i \qquad \forall i \in \mathcal{N} \quad \textbf{e.f.c} \tag{5.6}$$

$$l_{ij} \leq x_{ij} \leq u_{ij} \qquad \forall (i,j) \in \mathcal{A} \quad \textbf{c.c} \tag{5.7}$$

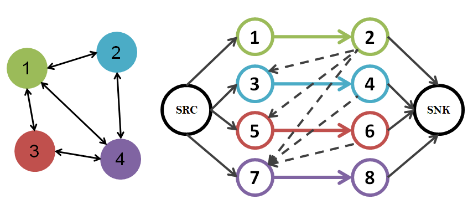
5.3.1 Network construction

To pose view-graph selection as an MCNF problem, we construct the network as follows. All indicator variables corresponding to selection of views $\{\delta_i\}$, and pairwise geometries $\{\delta_{ij}\}$ are represented using arcs/ edges $(i, j) \in \mathcal{A}$ in the network, source and sink nodes are auxiliary. Since view selection variables are represented as arcs, each vertex i in the view-graph corresponds to two nodes, $(2i - 1, 2i)$ in the network. Each odd node $(2i-1)$ corresponding to the vertex i in the view-graph is connected to the source node and similarly each even node $(2i)$ is connected to the sink node. The arcs corresponding to the pairwise selection variables $\{\delta_{ij}\}$, join the even node of the lower index image with the odd node of the higher index image. This choice prevents cycle formation in the network. Summarizing, the network consists of $|\mathcal{N}| = 2|\mathcal{V}| + 2$ nodes and $|\mathcal{A}| = |3\mathcal{V}| + |\mathcal{E}|$ arcs. These arc connections in the network are summarized in Figure 5.2 along with a pictorial example.

5.3.2 Equal flow and capacity constraints

At source and sink the sent and received flow is equal to the total flow. Equal flow constraints require that in and outflow at every other node remain equal. In our formulation, capacity constraints - lower and upper bounds on flow through an arc are specified based on the edge type as mentioned in Figure 5.2.

To understand the choice of these capacities, consider the vertex v_1 in the depicted view-graph. The arc corresponding to v_1's selection variable δ_1 in the network is $a(1,2)$. Corresponding to v_1's degree in the view-graph, the node n_2 in the network has three outgoing arcs $a(2,3)$, $a(2,5)$, and $a(2,7)$ for pairwise selection variables δ_{12}, δ_{13}, and δ_{14}. The flow starting from source node, after passing through $a(1,2)$, should plausibly be able to pass through all three outgoing arcs. Since flow can only be divided in integer units, the minimum capacity of image selection arc $a(1,2)$ has to be at least 3 $(deg(v_i))$. Under minimum cost solution, the flow at any node will continue to take the path of least resistance (cost). Now, suppose that the cost assigned to $a(2,3)$ is the least amongst the three outgoing arcs and its max. capacity is 3 units (or any value $1 < k < deg(v_i)$). In this scenario, $a(2,3)$ being the lowest



(a) View-graph & Network

Edge Type	Connected Nodes		Capacity
Source:	$(0, 2i-1)$	$\forall i \in \mathcal{V}, 0$: source	$[0, F]$
Sink:	$(2i, \lvert\mathcal{V}\rvert+1)$	$\forall i \in \mathcal{V}, \lvert\mathcal{V}\rvert+1$: sink	$[0, F]$
Image:	$(2i-1, 2i)$	$\forall i \in \mathcal{V}$	$[0, \text{degree}(i)]$
Pairwise:	$(2i, 2j-1)$,	$\forall e(i,j) \in \mathcal{E}, i < j$	$[0, 1]$

(b) Arc connections and capacities

Figure 5.2: (b) shows network construction for the sample view-graph of (a). Image nodes and the corresponding selection arcs are color-coded to match vertices in the view-graph. Pairwise selection arcs are depicted by black dashed lines. (c) describes the arc connections and their capacities.

cost arcs will pull all 3 units of flow from $a(1,2)$, starving the other arcs of any flow and preventing the corresponding view-graph edges from ever getting selected. To avoid this, we restrict the maximum flow through pairwise selection arcs to 1.

5.3.3 Effect of cost normalization and flow on solution

In minimization sense, negative costs provide encouragement for flow to pass through an arc, whereas positive costs provide discouragement. Suppose, all costs are negative, then total flow F of 1 unit will select the lowest cost chain in the network (often the longest). As we increase the value of total flow, more paths get explored, and when $F = |\mathcal{E}|$ all images and EGs get selected. When costs are both encouraging and discouraging, many positive cost arcs will act as barriers for the flow. As a result, at some value of total flow F the selection will (nearly) saturate and may never select the full view-graph. For the proposed applications, we use only encouraging costs ($-1 \leq c_{ij} \leq 0$), while flow remains the only free parameter.

5.3.4 Tuning flow parameter

To systematically study the effect of flow parameter on sub-graph selection, we solve the proposed MCNF problem for a large number of synthetic graphs with a varying number of vertices, varying degrees of connectivity simulated using Gaussian assumptions on the neighborhood, and randomly assigned (negative) cost. We construct the synthetic graphs to simulate connectivity of real-world unorganized image collections using the below process,

1. Begin with the number of vertices (N) in the graph and average degree (d_{avg}) as a fraction of N.
2. The degree of each vertex is sampled from a Gaussian distribution with $\mu = d_{avg}$ and varying σ.
3. The adjacency list of each vertex is also created by sampling a Gaussian distribution. For a vertex i, the d_i connected vertices are sampled from a Gaussian distribution with $\mu = i$ and σ as a function of average degree (d_{avg}). When vertex i is sampled, we randomly choose between vertices $i-1$ and $i+1$ for ensuring validity of edges.
4. The vertices are assumed to be arranged, in a linear fashion (e.g, capturing a facade or a street), and in a loop (e.g, capturing around a building or a block). For loopy configuration, we use modulo for defining vertex adjacency.

5. After defining the connections, each edge is randomly assigned a cost between -1 and 0.

6. Finally, the vertices are shuffled randomly to simulate the unorganized nature of SfM view-graphs.

We vary the number of vertices from 100 to 5000 in multiples of 2, the average degree from 5% to 40% in multiples of 5, and loopy/linear arrangement to obtain synthetic graphs with a wide range of scale and connectivity. In most Internet collection view-graphs, many images come from popular viewpoints; these images have a higher degree in the graph and are more likely connected to other adjacent images. While images corresponding to less popular viewpoints also exist in the datasets with fewer connections to popular viewpoint images. This scenario is well captured by our Gaussian sampling and neighbourhood assumptions in our simulation. Hence, the synthetic view-graphs are good approximations of connectivity in real-world unorganized community photo collections.

We performed sub-graph selection on ∼100 such synthetic view-graphs using the proposed MCNF approach for increasing values of total flow and observed a consistent pattern in selection behavior. Figure 5.3 shows the plot for fraction of total edges selected vs. normalized total flow. It can be seen that for negative costs, the relation between total flow and selected vertices and edges is logarithmic in nature. This relation is also observed on real-world (Internet) datasets with well-connected view-graphs (see Figure 5.9). Depending on the fraction of the total vertices and edges we want to be selected, the logarithmic dependence allows us to find the desired sub-graph by a binary search of the flow parameter over the $[0, 1]$ interval with a logarithmic scale. Such an iterative search is particularly effective in view of the extremely low computational time of the MCNF algorithm. In fact, the combined processing time of the whole search procedure is still an insignificant fraction of the total time required for SfM, and is thus far from being the computational bottleneck.

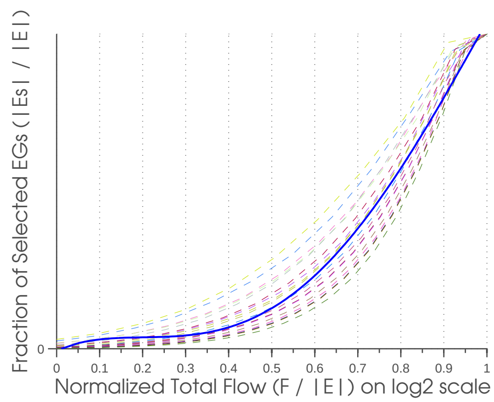


Figure 5.3: Effect of total flow on edge selection: selected edges increase monotonically (logarithmic) with the flow.

Running time A crucial advantage of this formulation is that it can be solved very efficiently. Constructing the network and solving for MCNF takes less than a second even for graphs with ∼1000 nodes and ∼100K edges.

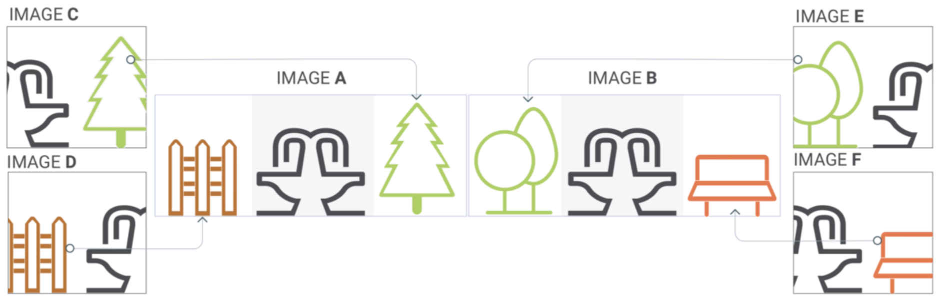


Figure 5.4: Image pair (A,B) captures a scene with duplicate elements. Though duplicate elements in the images yield many matches, elements in context regions tend to find matches with non-intersecting sets of images.

5.4 Applications and Cost Modeling

To show how the proposed framework can be used to address different concerns using task-specific costs, we tackle two use-cases. We first discuss cost modeling for the reconstruction of highly ambiguous datasets using local disambiguation priors (contribution of this work). Later, we discuss cost modeling for the general use-case of accurate and efficient reconstruction of large-scale Internet datasets using common SfM priors.

5.4.1 Cost modeling for ambiguous datasets

We propose that even for highly ambiguous datasets, with a conservative selection of an input view-graph such that it consists of a higher fraction of 'true' EGs, correct reconstruction can be recovered without requiring any change in the reconstruction pipeline. To achieve this using the proposed framework, we use three pairwise measures that act as strong priors for disambiguation and express pair selection cost as a linear combination of these three priors (with uniform weights). These priors are based on local geometry and can be easily computed at the time of initial view-graph construction without significant overhead. Note that $c^k(e_{ij})$ denotes the contribution of prior k to the total cost c for the pair e_{ij}, and g denotes a normalization function that distributes raw prior value to the desired cost range.

Context Similarity Missing correspondences (matches in a pair, not matching the third image in a triplet) are a useful prior for disambiguation. However, by itself, this prior is not very effective for disambiguation and it is commonly used in an inference framework with other priors or with a global objective [32, 53, 59]. Track covisibility statistics [67] are also useful for disambiguation, but it is not straightforward to apply this prior to describe pairwise fitness. In a post-reconstruction disambiguation approach [73], reconstructed 3D points are back-projected into image pairs and conflicting observations in unmatched regions (context) are identified. However, this measure is also not directly useful as our approach is a pre-process to SfM reconstruction. Motivated by these priors, we propose a new, context-based pairwise prior that is suitable to our framework.

For image pair $(\mathrm{I_i}, \mathrm{I_j})$, the sets of all matched features (matched with any image) are S_i, S_j and the sets of features that match between $(\mathrm{I_i}, \mathrm{I_j})$ are M_i, M_j. The difference sets $U_i = S_i \setminus M_i$, $U_j = S_j \setminus M_j$ consist of unique features in $\mathrm{I_i}$ and $\mathrm{I_j}$. Suppose images I_i and I_j are looking at a scene with duplicate instances of a structure. The features in the match sets M_i, M_j will most likely lie on the duplicate elements. We consider the unique feature sets U_i, U_j to belong to the context regions. If two images are truly looking at the same instance, the context features of both images would have matched similar set of images. On the other hand, if two images are looking at duplicate instances, the context features would be distributed over different sets of images (see Figure 5.4). We find distribution of features in U_i and U_j over all N images in the collection based on their matches and make an N dimensional description of the context space. Context feature of image I_i w.r.t. image I_j can be described as,

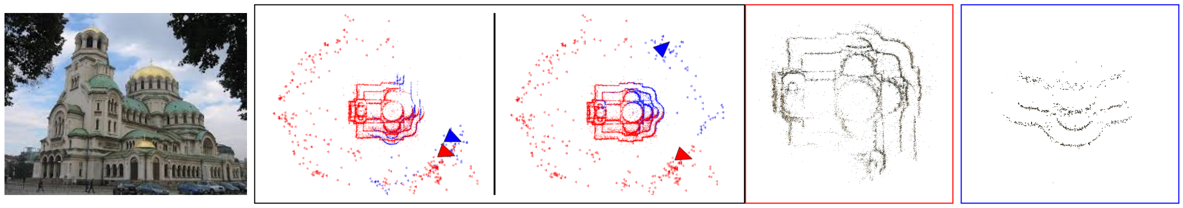
$$\mathbf{w}_{ij} = [w_i^1, w_i^2, \ldots, w_i^N] \quad w_i^j = 0 \tag{5.8}$$

$$w_i^k = |\{u \in U_i \mid u \leftrightarrow u', u' \in S_k, k \neq j\}| \tag{5.9}$$

This measure is slightly biased against pairs with very low visual overlap, however, combined with the other two measures, it works effectively for selection. We compared the context features of a pair using cosine similarity or hamming distance (after binarization) and found them to be working similarly. The context similarity can be defined as, $c^k(e_{ij}) = g(\mathbf{w}_{ij}^\intercal \mathbf{w}_{ji})$.

Loop consistency Loop consistency suggests that rotations in an EG triplet when chained should yield identity [17]. We find all triplets in the view-graph and label them as consistent or inconsistent. For each EG, we count the total number of consistent EGs it participates in and use this as a measure of its fitness. Though it is not same as explicitly enforcing loop consistency, it is a convenient way to incorporate a non-local (pairwise) geometric cue as a pairwise prior. $c^k(e_{ij}) = g(\frac{\text{\# consistent triplets on } (i,j)}{median(\text{\# consistent triplets on any}(i,j))})$.

Multiple motions We model this prior to disambiguate image pairs capturing repetitive or duplicate instances of some scene elements. We remove the correspondences that satisfy the estimated EG from



(a) (b)

Figure 5.11: Reconstruction of 'Alexander Nevsky Cathedral' dataset: (a) Result of [73] : (left) wrong reconstruction with full view-graph, (right) correctly split and merged models. Takes 16 minutes to post-process the wrong reconstruction; (b) Reconstruction with selected VG (ours) results into two models. Takes ~2 seconds for MCNF based selection with flow search criteria. Our splits are less clean than [73], notice the ghosting in red split. The selected subgraph has some consistent outlier EGs that the standard SfM process could not filter out. The reconstruction with smaller value of flow recovers just 1 partial (but clean) model.

To understand the effect of total flow, we plot the selection and reconstruction statistics for the increasing value of this parameter in Figure 5.9. It can be observed that a fraction of reconstructed vertices remains close to the fraction of selected vertices throughout, indicating that the selection scheme naturally endorses connectivity. This is a result of using only encouraging costs for selection that prefers to select longer chains over disjoint segments. Most vertices are reconstructed and the reconstruction errors become stable at a much lower value of flow (and selected pairs) than full VG selection. It is also worth noting that the pair selection vs. flow curve behaves very similarly to the synthetic view-graphs for all three datasets, validating our choice of using synthetic view-graph based selection statistics to model the flow parameter search strategy in practice.

5.6 Conclusions and Future Work

We presented a novel and efficient, unified framework for selecting subgraphs from initial view-graphs that can achieve different selection objectives with appropriately modeled image and pairwise selection costs. This mechanism provides an interesting way to separate dataset and task-specific challenges from the standard SfM pipeline, thereby improving its generality. We demonstrated the utility and potential of this framework by achieving satisfactory results for two different objectives with task-specific cost modeling. One interesting way to achieve even further abstraction to this problem would be to replace hand-designed costs by a weighted combination ($c_i = \sum_k \alpha_k f_k(i)$) of a number of known and designed priors. Cost formulation of this form would be expressive enough to cater to a wide variety of selection objectives. The problem of modeling costs to meet the desired objective then translates to that of devising new priors to add to the combination and finding the right weights for prior combination. In future, we wish to explore this direction for extending our framework. While this is non-trivial, it can lead to interesting directions of research for searching/learning new priors and the combination weights.

Sr.	Dataset	$\|\mathcal{V}\|$	$\|\mathcal{E}\|$	F	$\|\mathcal{V}_s\|$	$\|\mathcal{E}_s\|$	t_{sel}
1	Cereal	25	228	4	25	49	0.0137
2	Cup	64	1217	16	64	265	0.0268
3	Oats	24	220	4	24	49	0.0137
4	Street	19	95	2	19	26	0.0089
5	Books	21	161	4	21	41	0.0135
6	Desk	31	261	4	31	64	0.0138
7	ANC	448	5037	512	416	2497	0.1288
8	ADT	381	3627	512	340	2148	0.0940
9	RDC	271	3378	128	258	1498	0.0717
10	BG	161	2003	128	129	848	0.0522
11	CSB	277	5191	128	233	1380	0.0743
12	ToH	341	50332	32	341	1990	0.2372

(a) Datasets and selection details

		1	2	3	4	5	6
Other Methods	Roberts et al. [53]*	✗	✗	✗	✗	✗	✗
	Jiang et al. [59]	✓	✗	✓	✓	✓	✓
	Wilson et al. [67]	✗	✗	✗	✓	✗	✗
	Heinly et al. [73]	✓	–	–	–	–	✓
	Shen et al. [97]	–	✓	✓	✓	–	✗
	Yan et al. [109]	–	✓	–	–	✓	–
Ours with different priors	Baseline	✗	✗	✗	✗	✓	✗
	General+Disamb.	✗	✗	✓	✓	✓	✓
	Context	✓	✓	✓	✓	✓	✓
	Loop	✗	✓	✓	✓	✓	✓
	MM	✗	✗	✗	✗	✓	✓
	Context+Loop	✓	✓	✓	✓	✓	✓
	Context+MM	✓	✓	✓	✓	✓	✓
	Loop+MM	✓	✓	✓	✓	✓	✓
	Context+Loop+MM	✓	✓	✓	✓	✓	✓

(b) Comparison and Ablation

Table 5.2: (a) shows details of ambiguous datasets and selection statistics. t_{sel} shows the combined running time for all search iterations of MCNF solver. Abbreviated labels correspond to these datasets : ANC – Alexander Nevsky Cathedral, ADT – Arc de Triomphe, BG – Brandenburg Gate, CSB – Church on Splilled Blood, RDC – Radcliff Camera, TOH – Temple of Heaven; (b) top rows show results reported by other methods on small-scale ambiguous datasets (1 to 6) where – implies results not reported, *[53] succeeds on these sets only when time-stamp info is used. Bottom rows show results of our selection method with various prior based costs for given value of flow. Please see Figure 5.6 for ablation across flow values.

Dataset	VG	$\|\mathcal{V}\|$	$\|\mathcal{E}\|$	Incremental SfM R_{err}	Incremental SfM T_{err}	Incremental SfM r_{err}	Global SfM R_{err}	Global SfM T_{err}	Global SfM r_{err}
Castle	S	30	49	2.44	0.15	0.38	2.21	1.29	1.01
	F		118	2.22	0.22	0.34	2.17	7.49	1.14
Fountain	S	11	21	2.90	0.01	0.29	2.82	0.29	0.35
	F		25	2.90	0.01	0.72	2.82	0.27	0.59
Herzjesu	S	25	55	2.36	0.03	0.50	2.38	0.75	1.18
	F		128	2.38	0.02	0.43	2.39	0.56	1.71

(a) MVS dataset statistics

Dataset	VG	$\|\mathcal{V}\|$	$\|\mathcal{E}\|$	t_{sel}	N_c	r_{err}	R_{err}	T_{err}	t_{sfm}
Notre Dame	S	659	16970	1.744	628	1.41	0.072	0.195	1151
	F	714	46746	–	682	1.53	0.089	0.217	1760
Pantheon	S	761	15975	3.721	754	1.06	0.098	0.310	1785
	F	781	139630	–	775	1.31	0.125	0.309	3601
St. Peters	S	1132	39640	2.864	1095	1.341	0.037	0.517	1147
	F	1155	119977	–	1111	1.458	0.028	0.496	1367

(b) Internet landmarks datasets statistics

Table 5.3: VG selection and reconstruction statistics for MVS benchmark and Internet landmarks datasets. Labels 'S' and 'F' indicate selected and full VGs. $|\mathcal{V}|$, $|\mathcal{E}|$ indicate the vertices and edges in associated (selected or full) VGs. N_c is the number of reconstructed cameras and R_{err}, T_{err}, and r_{err} indicate median rotation, translation, and reprojection errors. t_{sel} and t_{sfm} indicate the time taken for MCNF based VG selection (cumulative) and SfM reconstruction respectively.

Chapter 6

Discussion on Future Directions

Large-scale structure from motion has matured significantly as a field to go from academic labs to the technology industry. Mapping environments at large-scale from structured and unstructured data sources opens up many new possibilities for two key industries, Augmented Reality and Autonomous Driving. While the autonomous driving industry operates closer to robotics and SLAM setting, enabling crowd-sourced AR requires mapping the environments from heterogeneous data sources and lies much closer to the large-scale SfM problem. While the core problems in large-scale SfM have good solutions, numerous challenges and corner-cases still remain in making world-scale mapping a reality. In this section, we will discuss two directions that I see of importance for future research in this domain.

6.1 Role of deep learning in large-scale SfM

In the last decade, deep learning has taken over as the default way of solving many traditional problems in computer vision such as retrieval, recognition, segmentation. SfM pipelines still largely follow standard computer vision paradigms. There are a few good reasons for this.

Parametrization The structure and parameterization of SfM problems do not fit naturally to deep learning pipelines. Most deep learning architectures such as CNNs, RNNs, autoencoders are made for processing Euclidean data. While image-to-pose regression methods are gaining popularity, end-to-end SfM problem is still not amenable to deep learning. Recent advancements in Graph Convolution Networks (GCN) show a lot of promise. As GCNs mature and become more mainstream, it will become possible to find a representation for core SfM data-structures like view-graph and covisibility graph, opening up possibilities for learning.

Feature learning mechanism Deep learning is good at finding statistical patterns in data and its true power lies in replacing hand-engineered features. While SIFT is still very popular in SfM pipelines and continues to outperform learned features, I believe this to change in the future. Key challenges where I see learned features to be of a greater advantage are matching data across the domain, matching seasonal variations, such as matching aerial to ground, matching heterogeneous sources, etc.

Lack of labeled ground-truth To gather large-scale ground-truth for SfM problems is nontrivial. This has lead to poor evaluation mechanisms where results are compared with a 'gold-standard' method (currently incremental SfM) rather than ground-truth. This issue becomes a bottleneck to making significant progress with learning-based methods since any learned representations inherit the strengths and weaknesses of best-known engineered solutions. For stereo problem, self-supervised methods with left-right consistency loss have shown great promise. I believe there is a need to investigate similar self-consistency metrics in SfM that lift the need for large ground-truth.

6.2 Continuous large-scale SfM

For AR applications to be able to use large-scale 3D models, it is important to consider that most urban environments change in appearance and often also in geometry with time. While most SfM methods deal with recovering the structure from given data at a time, there are no good mechanisms for updating recovered structures with new data. There are three sub-problems of importance in this context.

Model updates and model compression for localization Image-based localization systems aim to provide 6-DoF visual positioning by querying an image into an SfM model (using either 2D-3D feature matching or deep pose regression). It becomes essential to keep the model/index up-to-date with new observations of the environment if the positioning has to succeed with change in time. Another relevant challenge is model compression. With continuous updates to the map with new observations, it becomes imperative to reduce or maintain the model size while still keeping the map relevant for localization.

Model Merging With large-scale SfM, strategies that divide the reconstruction problem into multiple overlapping areas and combines the individually built models later are fairly common. Considering the fact that in a continuous reconstruction scenario, individual models might represent environments under widely different conditions, there will be a need for robust mechanisms for model merging.

Publications

Publications from IIIT Hyderabad

Journal Papers and Book Chapters

- Recovering the 3D Geometry of Heritage Monuments from Image Collections, R Shah, A Deshpande, A M Namboodiri, P J Narayanan, Digital Hampi: Preserving Indian Cultural Heritage, 2017.

- Interactive Video Manipulation Using Object Trajectories and Scene Backgrounds, R Shah and P J Narayanan, IEEE Transactions on Circuits and Systems for Video Technology, vol. 23, no. 9, 2013.

Refereed Conference Papers

- Unsupervised Image Style Embeddings for Retrieval and Recognition Tasks, S Gairola, R Shah and P J Narayanan, IEEE Winter Conference on Applications of Computer Vision (WACV), 2020.

- View-graph Selection Framework for SfM, R Shah, V Chari, P J Narayanan, European Conference on Computer Vision (ECCV), 2018.

- Find Me a Sky: A Data-Driven Method for Color-Consistent Sky Search and Replacement, S Rawat, S Gairola, R Shah, P J Narayanan, International Conference on Multimedia Modeling (MMM), 2018.

- SynCam: Capturing sub-frame synchronous media using smartphones, I Mehta, P Sakurikar, R Shah, P J Narayanan, IEEE International Conference on Multimedia and Expo (ICME), 2017

- From Traditional to Modern: Domain Adaptation for Action Classification in Short Social Video Clips, A Singh, S Saini, R Shah, P J Narayanan, German Conference on Pattern Recognition (GCPR), 2016.

- Learning to hash-tag videos with Tag2Vec, A Singh, S Saini, R Shah, P J Narayanan, Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP), 2016.

- Geometry-aware feature matching for structure from motion applications, R Shah, V Srivastava, P J Narayanan, IEEE Winter Conference on Applications of Computer Vision (WACV), 2015.

- Multistage SFM: Revisiting Incremental Structure from Motion, R Shah, A Deshpande, P J Narayanan, International Conference on 3D Vision (3DV), 2014.

- Trajectory based video object manipulation, IEEE International Conference on Multimedia and Expo (ICME), 2011.

- GPU-Accelerated Genetic Algorithms, R Shah, K Kothapalli, P J Narayanan, Workshop on Parallel Architectures for Bio-inspired Algorithms (PACT Workshop), 2010.

Other Publications

- Facial expression editing in images based on collections of images V Kwatra and R Shah, US Patent App. 13/693,701, 2014

- Video collections in panoramic contexts, James Tompkin, Fabrizio Pece, Rajvi Shah, Shahram Izadi, Jan Kautz, and Christian Theobalt, ACM symposium on User interface software and technology (UIST), 2013

- R Shah and V Kwatra, All smiles: automatic photo enhancement by facial expression analysis, ACM European Conference on Visual Media Production (CVMP), 2012

References

[1] R. D. Walk and E. J. Gibson. "A comparative and analytical study of visual depth perception." In: *Psychological Monographs: General and Applied* 75.15 (1961) (cit. on p. 1).

[2] L. G. Roberts. *Machine Perception of Three-Dimensional Solids*. Outstanding Dissertations in the Computer Sciences. Garland Publishing, 1963 (cit. on p. 4).

[3] D. Marr, T. Poggio, E. C. Hildreth, and W. E. L. Grimson. "A computational theory of human stereo vision". In: *Royal Society of London. Series B, Biological Sciences* 204.1156 (1979) (cit. on p. 1).

[4] E. B. Goldstein. *Sensation and perception*. Wadsworth Pub. Co, 1980 (cit. on p. 2).

[5] M. A. Fischler and R. C. Bolles. "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography". In: *Communications of the ACM* 24.6 (1981) (cit. on p. 37).

[6] J. J. Koenderink. "The structure of images". In: *Biological Cybernetics* 5 (1984) (cit. on p. 12).

[7] C. Taylor, D. Kriegman, and P. Anandan. "Structure and motion in two dimensions from multiple images: a least squares approach". In: *IEEE Workshop on Visual Motion*. 1991 (cit. on p. 21).

[8] C. Tomasi and T. Kanade. "Shape and motion from image streams under orthography: a factorization method". In: *International Journal of Computer Vision* 9.2 (1992) (cit. on p. 20).

[9] R. Szeliski and S. B. Kang. "Recovering 3D shape and motion from image streams using non-linear least squares". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 1993 (cit. on p. 21).

[10] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Springer International Series in Engineering and Computer Science, 1994 (cit. on p. 12).

[11] T. Lindeberg. “Scale-Space Theory: A Basic Tool for Analysing Structures at Different Scales”. In: *Journal of Applied Statistics* 21 (1994) (cit. on p. 12).

[12] P. F. Sturm and B. Triggs. “A Factorization Based Algorithm for Multi-Image Projective Structure and Motion”. In: *The European Conference on Computer Vision (ECCV)*. 1996 (cit. on p. 20).

[13] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu. “An Optimal Algorithm for Approximate Nearest Neighbor Searching Fixed Dimensions”. In: *Journal of the ACM* 45.6 (1998) (cit. on pp. 34, 35, 43, 55).

[14] S. Palmer. *Vision Science: Photons to Phenomenology*. MIT Press, 1999 (cit. on p. 1).

[15] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. Fitzgibbon. “Bundle Adjustment — A Modern Synthesis”. In: *Vision Algorithms: Theory and Practice*. Vol. 1883. 2000 (cit. on p. 21).

[16] V. M. Govindu. “Combining two-view constraints for motion estimation”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2001 (cit. on p. 21).

[17] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003 (cit. on pp. 16, 19, 20, 37, 83, 85).

[18] V. M. Govindu. “Lie-algebraic averaging for globally consistent motion estimation”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2004 (cit. on pp. 21, 75).

[19] D. G. Lowe. “Distinctive Image Features from Scale-Invariant Keypoints”. In: *International Journal of Computer Vision* 60.2 (2004) (cit. on pp. 11–13, 21, 33, 34, 41).

[20] D. Nistér. “An Efficient Solution to the Five-Point Relative Pose Problem”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2004) (cit. on p. 20).

[21] A. Whitehead and G. Roth. “Estimating Intrinsic Camera Parameters from the Fundamental Matrix Using an Evolutionary Approach”. In: *EURASIP Journal on Advances in Signal Processing* 8 (2004) (cit. on p. 16).

[22] M. Brown and D. Lowe. “Unsupervised 3D object recognition and reconstruction in unordered datasets”. In: *International Conference on 3-D Digital Imaging and Modeling (3d DIM*. 2005 (cit. on pp. 8, 21, 23, 27).

[23] S. Chopra, R. Hadsell, and Y. LeCun. “Learning a similarity metric discriminatively, with application to face verification”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 1. 2005 (cit. on p. 29).

[24] V. M. Govindu. “Robustness in Motion Averaging”. In: *Asian Conference on Computer Vision (ACCV)*. 2006 (cit. on pp. 26, 75).

[25] N. Snavely, S. M. Seitz, and R. Szeliski. “Photo Tourism: Exploring Photo Collections in 3D”. In: *ACM Trans. Graph.* 25.3 (2006) (cit. on pp. 8, 21, 23, 26, 27, 30, 33, 42, 58, 61, 74, 75, 90).

[26] W. Zhang and J. Kosecka. “Generalized RANSAC Framework for Relaxed Correspondence Problems”. In: *International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*. 2006 (cit. on p. 35).

[27] C. Wu. *SiftGPU: A GPU Implementation of Scale Invariant Feature Transform (SIFT)*. http://cs.unc.edu/~ccwu/siftgpu. 2007 (cit. on pp. 41, 43, 55, 61).

[28] N. Snavely, S. M. Seitz, and R. Szeliski. “Modeling the World from Internet Photo Collections”. In: *International Journal of Computer Vision* 80.2 (2008) (cit. on pp. 23, 25, 26, 52, 53, 75).

[29] N. Snavely, S. M. Seitz, and R. Szeliski. “Skeletal graphs for efficient structure from motion”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2008 (cit. on pp. 23, 26).

[30] C. Strecha, W. von Hansen, L. J. V. Gool, P. Fua, and U. Thoennessen. “On benchmarking camera calibration and multi-view stereo for high resolution imagery”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2008 (cit. on p. 90).

[31] T. Tuytelaars and K. Mikolajczyk. “Local Invariant Feature Detectors: A Survey”. In: *Foundations and Trends in Computer Graphics and Vision* 3.3 (2008) (cit. on p. 13).

[32] C. Zach, A. Irschara, and H. Bischof. “What can missing correspondences tell us about 3D structure and motion?” In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2008 (cit. on pp. 76, 82).

[33] M. Havlena, A. Torii, J. Knopp, and T. Pajdla. “Randomized structure from motion based on atomic 3D models from camera triplets”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009 (cit. on pp. 23, 27, 28, 74).

[34] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. "From structure-from-motion point clouds to fast location recognition". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009 (cit. on pp. 52, 58).

[35] V. Lepetit, F. Moreno-Noguer, and P. Fua. "EPnP: An Accurate O(N) Solution to the PnP Problem". In: *International Journal of Computer Vision* 81.2 (2009) (cit. on p. 17).

[36] S. Agarwal, K. Mierle, et al. *Ceres Solver*. `http://ceres-solver.org`. 2010 (cit. on p. 61).

[37] S. Agarwal, N. Snavely, S. M. Seitz, and R. Szeliski. "Bundle Adjustment in the Large". In: *The European Conference on Computer Vision (ECCV)*. 2010 (cit. on p. 27).

[38] M. Byröd and K. Åström. "Conjugate Gradient Bundle Adjustment". In: *The European Conference on Computer Vision (ECCV)*. 2010 (cit. on p. 27).

[39] O. Chum and J. Matas. "Large-Scale Discovery of Spatially Related Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.2 (2010) (cit. on pp. 23, 24).

[40] 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". In: *The European Conference on Computer Vision (ECCV)*. 2010 (cit. on pp. 8, 23–25, 28, 33, 34, 60).

[41] R. Gherardi, M. Farenzena, and A. Fusiello. "Improving the efficiency of hierarchical structure-and-motion". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2010 (cit. on pp. 23, 27, 28, 74).

[42] M. Havlena, A. Torii, and T. Pajdla. "Efficient Structure from Motion by Graph Optimization". In: *The European Conference on Computer Vision (ECCV)*. 2010 (cit. on pp. 23, 26).

[43] J. Hoberock and N. Bell. *Thrust: A Parallel Template Library*. Version 1.7.0. 2010. URL: `http://thrust.github.io/` (cit. on p. 41).

[44] Y. Li, N. Snavely, and D. P. Huttenlocher. "Location Recognition Using Prioritized Feature Matching". In: *The European Conference on Computer Vision (ECCV)*. 2010 (cit. on pp. 52, 54, 58, 59, 90).

[45] S. Sinha, D. Steedly, and R. Szeliski. "A Multi-stage Linear Approach to Structure from Motion". In: *The European Conference on Computer Vision (ECCV) RMLE Workshop*. 2010 (cit. on pp. 23, 27, 74, 75).

[46] J. Soman, K. Kothapalli, and P. J. Narayanan. “Some GPU Algorithms For Graph Connected Components and Spanning Tree”. In: *Parallel Processing Letters* 20.04 (2010) (cit. on p. 68).

[47] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag, 2010 (cit. on p. 5).

[48] C. Zach, M. Klopschitz, and M. Pollefeys. “Disambiguating visual relations using loop constraints”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2010 (cit. on pp. 26, 75).

[49] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. “Building Rome in a Day”. In: *Communications of the ACM* 54.10 (2011) (cit. on pp. 8, 23–25, 27, 30, 33, 34).

[50] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher. “Discrete-continuous Optimization for Large-scale Structure from Motion”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011 (cit. on pp. 23–25, 27, 54, 60, 70, 74, 75).

[51] C. Olsson and O. Enqvist. “Stable structure from motion for unordered image collections”. In: *Scandinavian conference on Image analysis (SCIA)*. 2011 (cit. on pp. 23, 26, 27, 74, 75).

[52] R. Raguram, C. Wu, J.-M. Frahm, and S. Lazebnik. “Modeling and Recognition of Landmark Image Collections Using Iconic Scene Graphs”. In: *International Journal of Computer Vision* 95.3 (2011) (cit. on pp. 26, 28).

[53] R. Roberts, S. Sinha, R. Szeliski, D. Steedly, and R. Szeliski. “Structure from motion for scenes with large duplicate structures”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011 (cit. on pp. 76, 82, 87, 95).

[54] T. Sattler, B. Leibe, and L. Kobbelt. “Fast image-based localization using direct 2D-to-3D matching”. In: *The IEEE International Conference on Computer Vision (ICCV)*. 2011 (cit. on pp. 52, 58).

[55] F. Sur, N. Noury, and M.-O. Berger. *Image point correspondences and repeated patterns*. Rapport de recherche RR-7693. INRIA, 2011 (cit. on p. 35).

[56] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. “Multicore Bundle Adjustment”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011 (cit. on pp. 27, 61).

[57] S. Cao and N. Snavely. “Learning to Match Images in Large-scale Collections”. In: *The European Conference on Computer Vision (ECCV) Workshop*. 2012 (cit. on pp. 23, 24, 34).

[58] S. Choudhary and P. Narayanan. “Visibility Probability Structure from SfM Datasets and Applications”. In: *The European Conference on Computer Vision (ECCV)*. 2012 (cit. on pp. 52, 58).

[59] N. Jiang, P. Tan, and L. F. Cheong. “Seeing double without confusion: Structure-from-motion in highly ambiguous scenes”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012 (cit. on pp. 76, 82, 95).

[60] M. Kushnir and I. Shimshoni. “Epipolar Geometry Estimation for Urban Scenes with Repetitive Structures”. In: *Asian Conference on Computer Vision (ACCV)*. 2012 (cit. on p. 35).

[61] Y. Lou, N. Snavely, and J. Gehrke. “MatchMiner: Efficient Spanning Structure Mining in Large Image Collections”. In: *The European Conference on Computer Vision (ECCV)*. 2012 (cit. on pp. 23, 24, 34).

[62] T. Sattler, B. Leibe, and L. Kobbelt. “Improving Image-based Localization by Active Correspondence Search”. In: *The European Conference on Computer Vision (ECCV)*. 2012 (cit. on pp. 58, 59).

[63] R. Shah. “Object-centric Video Navigation and Manipulation”. MS by Research Thesis. IIIT Hyderabad, India, 2012 (cit. on p. 30).

[64] Y. Bengio, A. Courville, and P. Vincent. “Representation Learning: A Review and New Perspectives”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013) (cit. on p. 28).

[65] A. Chatterjee and V. M. Govindu. “Efficient and Robust Large-Scale Rotation Averaging”. In: *The IEEE International Conference on Computer Vision (ICCV)*. 2013 (cit. on pp. 21, 23, 27, 61, 74, 75, 87).

[66] P. Moulon, P. Monasse, and R. Marlet. “Global Fusion of Relative Motions for Robust, Accurate and Scalable Structure from Motion”. In: *The IEEE International Conference on Computer Vision (ICCV)*. 2013 (cit. on pp. 23, 26, 74, 75).

[67] K. Wilson and N. Snavely. “Network Principles for SfM: Disambiguating Repeated Structures with Local Context”. In: *The IEEE International Conference on Computer Vision (ICCV)*. 2013 (cit. on pp. 76, 82, 95).

[68] C. Wu. “Towards Linear-Time Incremental Structure from Motion”. In: *International Conference on 3D Vision (3DV)*. 2013 (cit. on pp. 23, 24, 27, 34, 37, 61, 74, 91).

[69] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. “Neural Codes for Image Retrieval”. In: *The European Conference on Computer Vision (ECCV)*. 2014 (cit. on p. 29).

[70] B. Bhowmick, S. Patra, A. Chatterjee, V. Govindu, and S. Banerjee. “Divide and Conquer: Efficient Large-Scale Structure from Motion Using Graph Partitioning”. In: *Asian Conference on Computer Vision (ACCV)*. 2014 (cit. on pp. 23–25, 28, 60, 67).

[71] W. Hartmann, M. Havlena, and K. Schindler. “Predicting Matchability”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014 (cit. on pp. 23, 25, 54).

[72] M. Havlena and K. Schindler. “VocMatch: Efficient Multiview Correspondence for Structure from Motion”. In: *The European Conference on Computer Vision (ECCV)*. 2014 (cit. on pp. 23, 25).

[73] J. Heinly, E. Dunn, and J.-M. Frahm. “Correcting for Duplicate Scene Structure in Sparse 3D Reconstruction”. In: *The European Conference on Computer Vision (ECCV)*. 2014 (cit. on pp. 76, 83, 87, 88, 90, 93, 95).

[74] J. Heinly, E. Dunn, and J.-M. Frahm. “Recovering Correct Reconstructions from Indistinguishable Geometry”. In: *International Conference on 3D Vision (3DV)*. 2014 (cit. on p. 76).

[75] C. Jian, L. Cong, W. Jiaxiang, C. Hainan, and L. Hanqing. “Fast and Accurate Image Matching with Cascade Hashing for 3D Reconstruction”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014 (cit. on pp. 23, 25, 43).

[76] S. Krig. “Interest Point Detector and Feature Descriptor Survey”. In: *Computer Vision Metrics: Survey, Taxonomy, and Analysis*. 2014 (cit. on p. 13).

[77] M. Muja and D. G. Lowe. “Scalable Nearest Neighbor Algorithms for High Dimensional Data”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36 (2014) (cit. on pp. 34, 35, 38, 41, 43).

[78] R. Shah, A. Deshpande, and P. J. Narayanan. "Multistage SFM: Revisiting Incremental Structure from Motion". In: *International Conference on 3D Vision (3DV), 2014*. Vol. 1. 2014 (cit. on pp. 23, 27, 58).

[79] K. Wilson and N. Snavely. "Robust Global Translations with 1DSfM". In: *The European Conference on Computer Vision (ECCV)*. 2014 (cit. on pp. 23, 27, 61, 75, 87).

[80] A. Cohen, T. Sattler, and M. Pollefeys. "Merging the Unmatchable: Stitching Visually Disconnected SfM Models". In: *The IEEE International Conference on Computer Vision (ICCV)*. 2015 (cit. on p. 28).

[81] D. Eigen and R. Fergus. "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture". In: *IEEE International Conference on Computer Vision (ICCV)*. 2015 (cit. on p. 5).

[82] A. Kendall, M. Grimes, and R. Cipolla. "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015 (cit. on pp. 17, 30).

[83] S. M. LaValle. *Virtual Reality*. Cambridge University Press, 2015. URL: http://vr.cs.uiuc.edu/book.html (cit. on p. 2).

[84] K. Lin, H. Yang, J. Hsiao, and C. Chen. "Deep learning of binary hash codes for fast image retrieval". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015 (cit. on p. 29).

[85] J. L. Schönberger, A. C. Berg, and J.-M. Frahm. "PAIGE: PAirwise Image Geometry Encoding for Improved Efficiency in Structure-From-Motion". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015 (cit. on pp. 23, 24).

[86] R. Shah, V. Srivastava, and P. J. Narayanan. "Geometry-Aware Feature Matching for Structure from Motion Applications". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2015 (cit. on pp. 23, 25, 46).

[87] C. Sweeney. *Theia Multiview Geometry Library: Tutorial & Reference*. University of California Santa Barbara. 2015 (cit. on pp. 61, 91).

[88] C. Sweeney, T. Sattler, M. Turk, T. Hollerer, and M. Pollefeys. “Optimizing the Viewing Graph for Structure-from-Motion”. In: *The IEEE International Conference on Computer Vision (ICCV)*. 2015 (cit. on pp. 23, 26, 27, 75).

[89] R. Toldo, R. Gherardi, M. Farenzena, and A. Fusiello. “Hierarchical structure-and-motion recovery from uncalibrated images”. In: *Computer Vision and Image Understanding* 140 (2015) (cit. on pp. 26, 74).

[90] J. M. Wolfe, K. R. Kluender, and D. M. Levi. *Sensation and perception: An integrated approach*. Sinauer, 2015 (cit. on p. 2).

[91] S. Zagoruyko and N. Komodakis. “Learning to compare image patches via convolutional neural networks”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015 (cit. on p. 29).

[92] H. Altwaijry, A. Veit, and S. J. Belongie. “Learning to Detect and Match Keypoints with Deep Architectures”. In: *British Machine Vision Conference (BMVC)*. 2016 (cit. on p. 29).

[93] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker. “Universal Correspondence Network”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2016 (cit. on p. 29).

[94] D. DeTone, T. Malisiewicz, and A. Rabinovich. “Deep Image Homography Estimation”. In: *arXiv preprint* arXiv:1606.03798 (2016) (cit. on p. 29).

[95] I. Melekhov, J. Kannala, and E. Rahtu. “Siamese network features for image matching”. In: *International Conference on Pattern Recognition (ICPR)*. 2016 (cit. on p. 29).

[96] J. L. Schönberger and J.-M. Frahm. “Structure-from-Motion Revisited”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 (cit. on pp. 26, 74, 75).

[97] T. Shen, S. Zhu, T. Fang, R. Zhang, and L. Quan. “Graph-Based Consistent Matching for Structure-from-Motion”. In: *The European Conference on Computer Vision (ECCV)*. 2016 (cit. on pp. 23, 26, 76, 87, 95).

[98] K. Wilson, D. Bindel, and N. Snavely. “When is Rotations Averaging Hard?” In: *The European Conference on Computer Vision (ECCV)*. 2016 (cit. on p. 26).

[99] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. “LIFT: Learned Invariant Feature Transform”. In: *The European Conference on Computer Vision (ECCV)*. 2016 (cit. on p. 29).

[100] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk. “HPatches: A Benchmark and Evaluation of Handcrafted and Learned Local Descriptors”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 (cit. on p. 13).

[101] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T.-D. Nguyen, and M.-M. Cheng. “GMS: Grid-based Motion Statistics for Fast, Ultra-Robust Feature Correspondence”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 (cit. on p. 29).

[102] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother. “DSAC — Differentiable RANSAC for Camera Localization”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 (cit. on p. 30).

[103] C. Godard, O. Mac Aodha, and G. J. Brostow. “Unsupervised Monocular Depth Estimation with Left-Right Consistency”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 (cit. on p. 5).

[104] A. Kendall and R. Cipolla. “Geometric loss functions for camera pose regression with deep learning”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 (cit. on pp. 17, 30).

[105] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu. “Relative camera pose estimation using convolutional neural networks”. In: *International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS)*. 2017 (cit. on p. 29).

[106] J. L. Schonberger, H. Hardmeier, T. Sattler, and M. Pollefeys. “Comparative Evaluation of Hand-Crafted and Learned Local Features”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 (cit. on p. 29).

[107] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki. “SfM-Net: Learning of Structure and Motion from Video”. In: *arXiv preprint* arXiv:1704.07804 (2017) (2017) (cit. on p. 29).

[108] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, and D. Cremers. “Image-Based Localization Using LSTMs for Structured Feature Correlation”. In: *IEEE International Conference on Computer Vision (ICCV)*. 2017 (cit. on p. 30).

[109] Q. Yan, L. Yang, L. Zhang, and C. Xiao. "Distinguishing the Indistinguishable: Exploring Structural Ambiguities via Geodesic Context". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 (cit. on pp. 76, 87, 88, 95).

[110] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. "Unsupervised Learning of Depth and Ego-Motion from Video". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 (cit. on p. 5).

[111] D. DeTone, T. Malisiewicz, and A. Rabinovich. "SuperPoint: Self-Supervised Interest Point Detection and Description". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2018 (cit. on p. 29).

[112] A. Locher, M. Havlena, and L. Van Gool. "Progressive Structure from Motion". In: *The European Conference on Computer Vision (ECCV)*. 2018 (cit. on p. 23).

[113] T. Nguyen, S. W. Chen, S. S. Shivakumar, C. J. Taylor, and V. Kumar. "Unsupervised Deep Homography: A Fast and Robust Homography Estimation Model". In: *IEEE Robotics and Automation Letters* 3.3 (2018) (cit. on p. 29).

[114] R. Ranftl and V. Koltun. "Deep Fundamental Matrix Estimation". In: *The European Conference on Computer Vision (ECCV)*. 2018 (cit. on p. 29).

[115] Y. Jin, D. Mishkin, A. Mishchuk, J. Matas, P. Fua, K. M. Yi, and E. Trulls. "Image Matching across Wide Baselines: From Paper to Practice". In: *arXiv preprint* arXiv:2003.01587 (2020) (cit. on p. 13).

[116] *Camera Calibration Documentation, Matlab 2018b*. https://in.mathworks.com/help/vision/ug/camera-calibration.html (cit. on p. 16).

[117] *Technical Blog on Depth Censors*. https://blog.cometlabs.io/depth-sensors-are-the-key-to-unlocking-next-level-computer-vision-applications-3499533d3246 (cit. on p. 5).

[118] A. P. Witkin. "Scale-space Filtering". In: *Readings in Computer Vision* (cit. on p. 12).