CZECH TECHNICAL UNIVERSITY IN PRAGUE

DOCTORAL THESIS STATEMENT
Incremental Structure from Motion for Large Ordered and Unordered Sets of Images

Ph.D. Programme: Electrical Engineering and Information Technology, P2612
Branch of study: Artificial Intelligence and Biocybernetics, 3902V035

Doctoral thesis statement for obtaining the academic title of “Doctor”, abbreviated to “Ph.D.”

Prague, June 2012
The doctoral thesis was produced in full-time manner during Ph.D. study at the Center for Machine Perception of the Department of Cybernetics of the Faculty of Electrical Engineering of the CTU in Prague.

Candidate: Mgr. Michal Havlena  
Department of Cybernetics  
Faculty of Electrical Engineering of the CTU in Prague

Supervisor: Ing. Tomáš Pajdla, Ph.D.  
Department of Cybernetics  
Faculty of Electrical Engineering of the CTU in Prague

Opponents: ............................................................................................
............................................................................................
............................................................................................

The doctoral thesis statement was distributed on ....................

The defence of the doctoral thesis will be held on ............... at ...... a.m./p.m. before the Board for the Defence of the Doctoral Thesis in the branch of study Artificial Intelligence and Biocybernetics in the meeting room No. ........ of the Faculty of Electrical Engineering of the CTU in Prague.

Those interested may get acquainted with the doctoral thesis concerned at the Dean Office of the Faculty of Electrical Engineering of the CTU in Prague, at the Department for Science and Research, Technická 2, 166 27 Prague 6.

Prof. Ing. Vladimír Mařík, DrSc.  
Chairman of the Board for the Defence of the Doctoral Thesis in the branch of study Artificial Intelligence and Biocybernetics  
Department of Cybernetics, Karlovo náměstí 13, 121 35 Prague 2
# Contents

1 Problem Formulation 2

2 Contributions 3

3 State of the Art 5

4 City Modeling from Google Street View 6

5 Omnidirectional Sequence Stabilization 7

6 Randomized Structure from Motion 9

7 Image Set Reduction and Prioritized SfM 10

8 Conclusions 12

**Keywords:** Structure from Motion, Omnidirectional Vision, City Modeling

**Note:** Full text of the thesis is available at ftp://cmp.felk.cvut.cz/pub/cmp/articles/havlena/Havlena-TR-2012-13.pdf
1 Problem Formulation

In this thesis, we deal with the problem of modeling large predominantly static scenes from ordered, i.e. sequential, and unordered image sets comprising thousands of images when there is no further a priori information about the scene captured available. To solve such a task, we build upon the foundations of multiple view geometry [14], namely Structure from Motion.

The goal of Structure from Motion (SfM) is to recover both the positions of the 3D points of the scene (structure) and the unknown poses of the cameras capturing the scene (motion, external camera calibration). Knowing camera poses, one can use multi-view stereo 3D surface reconstruction methods [9, 17] to construct compelling 3D surface models of the scene, see Figure 1.

There is no closed form solution for solving the SfM problem [30]. The problem is highly non-linear and the search for the optimal model parameters by minimizing the reprojection error is likely to get stuck in a local minimum.

The methods described in this thesis fall into the group of the “incremental” methods which use simple components such as epipolar geometry computation for a pair of images, 3D point triangulation, and camera resectioning to build the resulting model incrementally, starting from one or more seed reconstructions and connecting additional cameras until the whole scene is reconstructed [37], all interleaved with necessary non-linear optimization [22]. The art of such methods resides in the design of a computational “pipeline” which combines the aforementioned components in order to create a reconstruction procedure which is both robust and efficient for the input image data. When the size of the image set grows, exhaustive pairwise image matching commonly used for revealing the structure of the data becomes infeasible.

![Figure 1](image1.png)

**Figure 1:** (a) Sample images from 316 images of Marseille Centre de la Vieille Charité. (b) External camera calibration and a sparse 3D point cloud obtained using method [38]. (c) 3D surface reconstruction computed by method [17].
2 Contributions

The contribution of the thesis is related to large scale Structure from Motion (SfM) from both ordered and unordered image sets. The research on sequential SfM was conducted in collaboration with my colleague, Akihiko Torii, and the research on SfM from unordered data was done by me. Specifically, my contributions are the following:

- **Use of visual indexing for image pair/triplet selection.** We avoid the most time consuming step of large scale SfM from unordered image sets, the computation of all pairwise matches and geometries, by sampling pairs of images and estimating visual overlap using the detected occurrences of visual words. The evaluation of the similarity scores by computing scalar products of so-called tf-idf vectors [36] is also quadratic in the number of images in the set but scalar product is a much simpler operation than full feature matching which leads to a significant speedup of SfM. Furthermore, we proposed to sample triplets of images instead of pairs for the seeds of the reconstruction because 3D points verified in three views are more likely to be correct. The constructed atomic 3D models are merged together to give the final large scale 3D model at later stages of the computation.

- **Image set reduction by applying a graph algorithm (CDS).** The idea of using visual indexing for SfM was further extended in order to be able to reconstruct image sets with uneven image coverage, i.e. community image sets of cities with landmarks, efficiently. A small subset from the set of input images is selected by computing the approximate minimum connected dominating set of a graph with vertices being the images and edges connecting the visually similar images by a fast polynomial algorithm [13]. This kind of reduction guarantees, to some extent, that the removed images have visual overlap with at least some images left in the set and therefore can be connected to the resulting 3D model later.

- **Task ordering using a priority queue.** We use task prioritization to avoid spending too much time in a few difficult matching problems instead of exploring other easier options. Compared to our previous work having the computation split in several stages [15], the usage of a priority queue for interleaving different “atomic 3D model construction” and “image connection” tasks facilitates obtaining reasonable reconstructions in limited time. The priorities of the individual tasks are set according to image similarity and the history of the computation.
Joint contributions follow:

- **Computation of dominant apical angle (DAA).** When performing sequential SfM by chaining pairwise epipolar geometries [14], the reconstruction usually fails when the amount of translation between consecutive cameras is not sufficient. We demonstrate that the amount of translation can be reliably measured for general as well as planar scenes by the most frequent apical angle, the angle under which the camera centers are seen from the perspective of the reconstructed scene points. By selecting only image pairs which have sufficient DAA, one is able to easily reconstruct even sequences with variable camera motion speed.

- **Sequence bridging by visual indexing.** We extend the known concept of loop closing, e.g. [16], which tries to correct the trajectory of the camera once the same place is re-visited, by searching for all the trajectory loops at once based on co-occurring visual words. Geometrically verified loop candidates are added to the model as new constraints for bundle adjustment which closes the detected loops as it enforces global consistency of camera poses and 3D structure in the sequence.

- **Image stabilization using non-central cylindrical image generation.** A new technique for omnidirectional image rectification based on stereographic image projection was introduced as an alternative to central cylindrical image generation. We show that non-central cylindrical images are suitable for people recognition with classifiers trained on perspective data, e.g. [6], once the images are stabilized w.r.t. the ground plane.

- **Using cone test instead of reprojection error.** When verifying 2D-3D matches in a RANSAC loop [7], we do not rely on the widely used reprojection error but make use of the fact that the projections of the 3D point to all the related images are known and use a “cone test” instead. Two pixels wide pyramids are cast through the corresponding pixel locations in the related cameras and an LP feasibility task is solved to decide whether the interaction of the “cones” is non-empty or not. This allows for accepting a correct match even if the currently estimated 3D point location is incorrect without modeling the probability distribution of the location of the 3D point explicitly.

It is worth mentioning that the methods were implemented to work with the general central camera model, covering the most common special cases including (i) perspective cameras, (ii) fish-eye lenses, (iii) equirectangular panoramas, and (iv) cameras calibrated by the polynomial omnidirectional model.
3 State of the Art

The complexity of large scale SfM computation is quite different for ordered and unordered image sets as image order gives a clue which pairs of images should have an overlapping field of view and are therefore suitable for processing. On the other hand, the methods for fast selection of promising image pairs can be used also for ordered image sets to improve the consistency of the resulting models via loop closing.

Ordered Image Set Processing

Short baseline SfM using simple image features [5], which performs real-time detection and matching, recovers camera poses and trajectory sufficiently well when all camera motions between consecutive frames in the sequence are small. On the other hand, wide baseline SfM based methods, which use richer features such as MSER [25], Laplacian-Affine, Hessian-Affine [27], SIFT [23], and SURF [2], are capable of producing feasible tentative matches under large changes of visual appearance between images induced by rapid changes of camera pose and illumination. Work [10] presented SfM based on wide baseline matching of SIFT features using a single omnidirectional camera and demonstrated the performance on indoor environments.

The state of the art technique for finding relative camera poses from image matches first establishes tentative matches by pairing image points with mutually similar features and then uses RANSAC [7, 14, 4] to look for a large subset of the set of tentative matches which is, within a predefined threshold $\epsilon$, consistent with an epipolar geometry (EG) [14]. Unfortunately, this strategy does not always recover the epipolar geometry generated by the actual camera motion, which has been observed in [20, 29, 41]. Often, there are more models which are supported by a large number of matches. Then, the chance that the correct model, even if it has the largest support, will be found by running a single RANSAC is small. Work [20] suggested to generate models by randomized sampling as in RANSAC but to use soft (kernel) voting for a physical parameter instead of looking for the maximal support.

Work [26] demonstrated 3D modeling from perspective images exported from Google Street View images using piecewise planar structure constraints. Another recent related work [39] demonstrated the performance of SfM which employs guided matching by using epipolar geometries computed in previous frames, and robust camera trajectory estimation by computing camera orientations and positions individually for the calibrated perspective images acquired by Point Grey Ladybug Spherical Digital Video Camera System [33].
In [18], loop closing capable of removing the drift error of sequential SfM is achieved by merging partial reconstructions of overlapping sequences which are extracted using an image similarity matrix [36, 19]. Work [34] finds loop endpoints by using the image similarity matrix and verifies the loops by computing the rotation transform between the pairs of origins and endpoints under the assumption that the positions of the origin and the endpoint of each loop coincide. Furthermore, they constraint the camera motions on a plane to reduce the number of parameters in bundle adjustment.

**Unordered Image Set Processing**

Most of the state of the art techniques for 3D reconstruction from unordered image sets [35, 3, 42, 24] start the computation by performing exhaustive pairwise image matching in order to reveal the structure and connectivity of the data.

Bundler [37] uses exhaustive pairwise image feature matching and epipolar geometry computation to create an image graph which is later used to lead the reconstruction. By finding the skeletal set [38] of this graph, the reconstruction time improves significantly but the time spent on image matching remains the same. Recent advancement of the aforementioned technique [1] abandons exhaustive pairwise image matching by using shared occurrences of visual words [31, 36] to match only the ten most promising images per each input image. A similar approach was used in Google Maps to construct the models of several popular landmarks from all around the world using user-contributed Picasa and Panoramio photos [12].

Another possible approach to reducing the number of necessary pairwise image feature matchings lies in reducing the number of images to be processed because the input image set may be highly redundant. The approach presented in [21] clusters the input images using the GIST [32] descriptor giving raise to “iconic images”. These images and the pairwise geometric relationships between them define an “iconic scene graph” that captures all the important aspects of the original image set. In [8], the method has been re-implemented in order to be highly parallel and therefore suitable for GPU computing.

**4 City Modeling from Google Street View**

When constructing 3D models of large city areas, it is beneficial to use 360° field of view images, see Figure 2, as it increases robustness against occlusions. The main contribution of the presented method lies in demonstrating that one can achieve SfM from a single sparse omnidirectional sequence with only an
Figure 2: Camera trajectory computed by SfM. (a) Camera positions (red circles) exported into Google Earth [11]. (b) The 3D model representing 4,799 camera positions (red circles) and 123,035 3D points (color dots).

approximate knowledge of calibration as opposed to [5, 39] where the models are computed from dense sequences and with precisely calibrated cameras.

The proposed SfM pipeline is an extension of work [40] which demonstrated the performance of the recovery of camera poses and trajectory on the image sequence acquired by a single fish-eye lens camera. Loop closing is facilitated by visual indexing. SURF [2] descriptors of each image are quantized into visual words and term frequency–inverse document frequency (tf-idf) vectors [36, 19] are computed. Image similarity matrix $M$ is constructed by computing the cosines of angles between normalized tf-idf vectors, i.e. their scalar products, for all pairs of images.

The 1st to 50th diagonals of $M$ are zeroed in order to exclude very small loops and for each image $I_i$ in the sequence, a candidate $I_j$ of the endpoint of the loop which starts from $I_i$ is selected as the one having the highest similarity score in the $i$-th row of $M$. Next, the candidate image $I_j$ is verified by solving the camera resectioning [28]. If the inlier ratio is higher than 70%, camera resectioning is considered successful and the candidate image $I_j$ is accepted as the endpoint of the loop. The validated candidates are used to give additional constraints on the final bundle adjustment [22].

5 Omnidirectional Sequence Stabilization

In order to make the wide baseline SfM pipeline capable to recover camera poses and trajectories even from sequences that have large differences in the
Figure 3: The apical angle $\tau$ at the point $X$ reconstructed from the correspondence $(x, x')$ relatively depends on the length of the camera translation $t$ and on the distances of $X$ from the camera centers $C, C'$.

amount of camera translation between consecutive frames, a keyframe selection method based on DAA computation is proposed. Secondly, stabilized non-central cylindrical images are generated to facilitate pedestrian detection using detectors trained on perspective images using the known ground plane position assumption.

Measuring the Amount of Camera Translation by DAA. Having $m$ matches $(x_i, x'_i)$ and the essential matrix $E$ computed from them, we can reconstruct 3D points $X_i$. Figure 3 shows a point $X$ reconstructed from an image match $(x, x')$. For each point $X$, the apical angle $\tau$, which measures the length of the camera translation from the perspective of the point $X$, is computed. If the cameras are related by pure rotation, all angles $\tau$ are equal to zero. The larger is the camera translation, the larger are the angles $\tau$.

For a given $E$ and $m$ matches $(x_i, x'_i)$, one can select the decomposition of $E$ to $R$ and $t$, which reconstructs the largest number of 3D points in front of the cameras. The apical angle $\tau_i$, corresponding to the match $(x_i, x'_i)$, is computed by solving a set of linear equations for the relative distances $\alpha_i, \alpha'_i$

$$\alpha'_i x'_i = \alpha_i Rx_i - t$$

(1)

in the least square sense and by using the law of cosines

$$2 \alpha_i \alpha'_i \cos(\tau_i) = \alpha_i^2 + \alpha'_i^2 - \|t\|^2.$$  

(2)

For a small translation w.r.t. the distance to the scene points, approximation $\alpha_i = \alpha'_i$ can be used and the apical angle $\tau_i$ becomes a linear function of $\|t\|$. 

8
Figure 4: Image stabilization and transformation. (a) The camera poses and the world 3D points reconstructed by our SfM visualized from a bird’s eye view. (b) Original images. (c) Non-stabilized images. (d) Stabilized images.

Omnidirectional Image Stabilization. The recovered camera poses and trajectory can be used to rectify the original images to the stabilized ones. Image stabilization is beneficial e.g. for facilitating visual object recognition where (i) objects can be detected in canonical orientations and (ii) ground plane position can further restrict feasible object locations. When the sequence is captured by walking or driving on the roads, the images can be stabilized w.r.t. the ground plane with a natural assumption that the motion direction is parallel to the ground plane. If there exists no constraint on camera motion in the sequence, the simplest way of stabilization is to rectify images w.r.t. the up vector in the coordinate system of the first camera, see Figure 4.

6 Randomized Structure from Motion

The computation of the randomized SfM pipeline for unordered image sets consists of four consecutive steps which are executed one after another: (i) computing image similarity matrix, (ii) constructing atomic 3D models from camera
triplets, (iii) merging partial reconstructions, and (iv) gluing single cameras to the best partial reconstruction, see Figure 5.

Image similarity matrix $M$ is used to select triplets of cameras suitable for constructing atomic 3D models. The maximum score in the matrix gives a pair of cameras $C_i$ and $C_j$. Then, three “third camera” candidates are found and atomic 3D models are constructed for each of the candidates. The resulting models are ranked by the quality score, which checks (i) whether there is a sufficient number of 3D points with large apical angles and (ii) the uniformity of image coverage by the projections of reconstructed 3D points, and the model with the highest quality score is selected. Denoting the third camera corresponding to the selected atomic 3D model as $C_k$, cameras $C_i$, $C_j$, and $C_k$ are removed from future selections by zeroing rows and columns $i$, $j$, and $k$ of $M$.

When the atomic 3D model construction step finishes, each accepted atomic 3D model gives raise to a partial reconstruction and merging guided by a similarity matrix containing scores between selected atomic 3D models is performed. During the merging step, the partial reconstructions are connected together forming larger partial reconstructions containing the union of cameras and 3D points of the connected reconstructions. Finally, in the gluing step, the best partial reconstruction is selected as the one containing the highest number of cameras and the poses of the cameras which are not contained in it are estimated [28]. Cone test is used to evaluate the support of individual RANSAC samples during merging and gluing.

7 Image Set Reduction and Prioritized SfM

Unstructured web collections often contain a large number of very similar images of landmarks while, on the other hand, image sequences often have a very limited overlap between images. To speed up the reconstruction, it is desirable
Figure 6: Schematic visualization of the computation. The task retrieved from the head of the priority queue can be either an atomic 3D model construction task (dark gray) or an image connection task (light gray). Unsuccessful atomic 3D model construction (–) inserts another atomic 3D model construction task with the priority key doubled into the queue, a successful one (+) inserts five image connection tasks. Unsuccessful image connection (–) inserts the same task again with the priority key doubled, a successful one (+) inserts a new image connection task. Merging of overlapping 3D models is called implicitly after every successful image connection if the overlap is sufficient.

to select a subset of the input images in such a way that all the remaining images have a significant visual overlap with at least one image from the selected ones, so the connectivity of the resulting model should not be damaged. For selecting such a subset of input images, the approximate minimum connected dominating set can be computed by a fast polynomial algorithm [13] on the graph constructed according to the estimated visual overlap. This is closely related to the maximum leaf spanning tree algorithm employed in [38] but the composition of the graph is less computationally demanding in our case. Compared to [21] using GIST, our method is more robust to viewpoint changes.

The proposed SfM pipeline, reconstructing only the selected subset of images, still uses the atomic 3D models constructed from camera triplets as the basic elements of the reconstruction but the strict division of the computation into steps is relaxed by introducing a priority queue which interleaves different reconstruction tasks, see Figure 6, in order to get a good scene covering recon-
Figure 7: The largest partial 3D models reconstructed from the reduced image sets after 6 hours of computation. (a) Omnidirectional image set CASTLE (1,063 images). (b) Perspective image set VIENNA (1,008 images).

struction in limited time. Our aim here is to avoid spending too much time in a few difficult matching problems by exploring other easier options which lead to a comparable resulting 3D model in shorter computational time compared to the previous method, see Figure 7 for the obtained 3D models.

8 Conclusions

The thesis contributes to improving scalability and efficiency of Structure from Motion computation from both ordered and unordered image sets. In particular, the usefulness of visual indexing for estimating visual overlap between images was shown. When visual indexing is used, exhaustive pairwise image matching can be avoided, the sizes of redundant image sets can be significantly reduced, and the SfM computation can be properly prioritized.

The reconstruction pipelines are accessible to the registered users through our web-based interface, http://ptak.felk.cvut.cz/sfmservice, and were successfully used to reconstruct many challenging image sets.

References


Disertace se zabývá výpočtem tvaru z pohybu (Structure from Motion) z rozsáhlých uspořádaných, tj. sekvenčních, a neuspořádaných množin obrazů. Navrhujeme nahradit nejvíce časově náročnou část výpočtu z neuspořádaných dat, hledání vizuálních korespondencí mezi všemi páry obrazů, rychlým výpočtem odhadu vizuálního překryvu dvou obrazů na základě detekovaných výskytů vizuálních slov. Hledání korespondencí pouze mezi páry obrazů s dostatečným odhadnutým překryvem vede k významnému zrychlení výpočtu. Efektivita rekonstrukce z redundantních množin obrazů, například z obrazů městských památek stažených z Internetu, však může být dále zlepšena použitím navržené redukce velikosti vstupní množiny obrazů rychlým grafovým algoritmem.

Ke ztrátě efektivity výpočtu také dochází, když použitá metoda tráví příliš mnoho času řešení několika obtížných rekonstrukčních podúloh místo hledání jiných, často lehčích, cest výpočtu. Navrhujeme využít prioritní frontu k prokládání jednotlivých podúloh, což umožňuje získat rozumné 3D modely v omezeném čase. Priority jednotlivých úkolů jsou nastavovány s ohledem na odhadnuté vizuální překryvy, ale zároveň jsou ovlivněny i historií výpočtu.

Vizuální překryvy odhadnuté z opakujících se výskytů vizuálních slov se osvědčily také při zpracování uspořádaných množin obrazů. Geometricky ověřené návraty do dříve navštívených částí scény, tzv. vizuální smyčky, přidáváme do konstruovaného 3D modelu jako nová omezení pro nelineární metodu vyrovnání svazků paprsků. Ta detekované smyčky uzavírá tím, že vynucuje globální konzistenci 3D modelu pro celou sekvenci.

Byla navržena také řada technických vylepšení výpočtu: (i) trojice obrazů jsou použity místo párů jako zárodky rekonstrukcí, protože 3D body ověřené ve třech pohledech jsou mělké i chybné. (ii) Velikost posunu kamery vzhledem ke scéně, ukazatel spolehlivosti výpočtu relativního posunu kamery, je měřena dominantním apikálním úhlem. Nalezením páru obrazů s dostatečně velkými apikálními úhly vybíráme klíčové snímky ze vstupních sekvencí nebo kvalitní zárodky rekonstrukcí z neuspořádaných množin obrazů. (iii) Test průniku kuželů je použit jako měření tradiční zpracovávání částečně súhlasících větví v 2D-3D korespondencích v RANSACu. Takto mohou být správné korespondence přijaty, i pokud jsou aktuálně odhadnuté polohy 3D bodů nesprávné.

Funkčnost navržených metod je demonstrována množstvím experimentů z uspořádaných i neuspořádaných množin sejmutých z tisíců obrazů. Modely části měst jsou zkonstruovány jak z obrazů pořízených rybím okem, tak ze všesměrových panoramat. Obrazy vygenerované navrženou necentrální válcovou projekcí se sekvencí stabilizovaných vzhledem k vodorovné rovině jsou úspěšně použity k detekci chodců.
Author’s Publications

Publications related to the thesis

Impacted journal articles


Publications excerpted by WOS


Other conference/congress publications


Additional publications

Impacted journal articles


Publications excerpted by WOS


Other conference/congress publications


Citations of Author’s Work


