

# **Learning Structure from Motion**

**DRAFT**

## **Masterarbeit**

der Philosophisch-naturwissenschaftlichen Fakultät  
der Universität Bern

vorgelegt von

Adrian Wälchli

2017

Leiter der Arbeit:  
Prof. Dr. Paolo Favaro  
Institut für Informatik und angewandte Mathematik

## **Abstract**

# Contents

<b>1</b>	<b>Prior Work</b>	<b>1</b>
1.1	Feature-based methods SURF, ORB, SIFT . . . . .	1
1.2	The Eight Point Algorithm . . . . .	1
1.3	Factorization Methods . . . . .	2
1.4	Bundle Adjustment . . . . .	3
1.5	2015 - PoseNet . . . . .	3
1.6	04/2016 - 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction . . . . .	4
1.7	07/2016 - Unsupervised CNN for Single View Depth Estimation	4
1.8	04/2017 - SfM-Net . . . . .	5
1.9	04/2017 - Unsupervised Learning of Depth and Ego-Motion from Video . . . . .	6
1.10	04/2017 - DeMoN: Depth and Motion Network for Learning Monoc- ular Stereo . . . . .	6
1.11	05/2017 - Relative Camera Pose Estimation Using Convolutional Neural Networks . . . . .	7
<b>2</b>	<b>Introduction</b>	<b>9</b>
2.1	Relative Camera Pose . . . . .	9
2.2	Two-Frame Structure from Motion . . . . .	9
2.3	Multiview Structure from Motion . . . . .	9
<b>A</b>	<b>Appendix</b>	<b>10</b>
	<b>List of Tables</b>	<b>11</b>
	<b>List of Figures</b>	<b>11</b>
	<b>Bibliography</b>	<b>12</b>

# Chapter 1

## Prior Work

**Add text that introduces the chapter**  
**Find good structure/order to present the different works**  
**Better point out differences and similarities**

Possible definition of structure from motion by ?:

“The structure from motion (SfM) problem in computer vision is the problem of recovering the three-dimensional (3D) structure of a stationary scene from a set of projective measurements, represented as a collection of two-dimensional (2D) images, via estimation of motion of the cameras corresponding to these images.”

One can define the three basic steps of SfM as:

1. Feature detection, extraction and matching
2. Camera motion estimation
3. Recovery of 3D structure

This chapter presents the prior work that focuses on one or multiple of these steps.

### 1.1 Feature-based methods SURF, ORB, SIFT

### 1.2 The Eight Point Algorithm

The eight point algorithm, as introduced by ?, makes use of the epipolar geometry between a pair of views in order to compute relative translation and orientation of the cameras. The relationship between the two camera views is described by the so-called *essential matrix*. It is defined as

$$E = [t]_{\times} R, \quad (1.1)$$

where  $[t]_{\times}$  is the matrix that computes the cross-product between  $t$  and an arbitrary vector. It describes the relationship between corresponding points in the image planes of the two cameras. When  $R$  and  $t$  are unknown, it is possible

to compute the essential matrix using known corresponding points. The epipolar constraint is formulated as

$$\hat{x}_1^\top E \hat{x}_0 = 0, \quad (1.2)$$

where  $\hat{x}_0$  and  $\hat{x}_1$  are the corresponding points in image one and two respectively. The unknown elements of the essential matrix can be determined using this system of equations. It requires eight equations for eight elements of the matrix and one element has to be fixed because any scaling of the true matrix satisfies the epipolar constraint. After estimating  $E$ , the rotation and translation can be recovered from it up to some ambiguities.

In practice, there are some problems that need to be addressed. First, the corresponding points are usually not known and must be found using a feature detection and matching technique (SfM step 1). This then can lead to outliers which are wrong correspondences. In this case, RANSAC is applied to select the best eight points among many possible choices. Second, the selected points are usually not exact and contain noise that lead to a rank three estimate of  $E$  instead of rank two. The singular value decomposition is used to discard the component with the smallest singular value. **numerical issues, 5 point algorithm** Third, in some cases the calibration matrices  $K_0$  and  $K_1$  are unknown. This means that the normalized coordinates are  $\hat{x}_j = K_j^{-1}x_j$  are no longer available and only the image coordinates  $x_j$  are known. This results in the so called *fundamental matrix*  $F$  in the epipolar constraint:

$$\hat{x}_1^\top E \hat{x}_0 = x_1^\top K_1^{-\top} E K_0^{-1} x_0 = x_1^\top F x_0 = 0. \quad (1.3)$$

**explain what F is used for**

### 1.3 Factorization Methods

The factorization method can be used to recover camera motion and 3D shape simultaneously (step 2 and 3). A first method was introduced by ? that works with an orthographic camera model. It uses the so called *measurement matrix*  $W \in \mathbb{R}^{2n \times m}$  which is defined as

$$W = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \\ y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{bmatrix}. \quad (1.4)$$

It contains the 2D coordinates of the  $m$  orthographically projected 3D points for each of the  $n$  cameras. Under the assumption that the origin of global coordinate system is located at the centroid of the point cloud, the matrix  $W$  can be factorized into

$$W = MS, \quad (1.5)$$

where  $M \in \mathbb{R}^{2n \times 3}$  contains the  $2 \times 3$  projection matrices of each camera and  $S \in \mathbb{R}^{3 \times m}$  are the points of the observed 3D shape. As described by ?, this factorization can be achieved using the singular value decomposition (SVD) as

$W = U\Sigma V^\top$  and deriving  $M$  and  $S$  from  $U$ ,  $V$  and  $\Sigma$ . **Add more details?**  
**Less details?**

This factorization method has some disadvantages. First, it can not be applied to the commonly used perspective projection model, although it is approximated by the orthographic model for distant objects. The works of ? and ? have extended the factorization method for perspective projection using an iterative process. Second, in order for the factorization method to work, all points of the shape must be visible in every frame, i.e. to obtain the measurement matrix one needs to find the correspondences across all frames.

## 1.4 Bundle Adjustment

Another method to solve multi-view SfM is the method of *bundle adjustment*. It is the most general technique for simultaneously recovering 3D shape and camera pose, but it is also computationally expensive. The bundle adjustment technique aims to minimize the re-projection error of the unknown 3D points and camera matrices. The re-projection error is formulated as the euclidean distance between the known image coordinates and the projection of the unknown points using the unknown camera matrices.

More formally, let  $\{C_j\}_{j=1}^m$  be the (unknown) camera matrices that encode location, rotation and intrinsic parameters, let  $\{p_i\}_{i=1}^n$  denote the (unknown) 3D points and let  $x_{ij}$  be the (known) observed 2D coordinates of the point  $p_i$  in camera  $j$ . Then, the objective of bundle adjustment is defined as

$$\min_{\{p_i\}, \{C_j\}} \sum_{i=1}^n \sum_{j=1}^m v(i, j) \|P(C_j, p_i) - x_{ij}\|_2, \quad (1.6)$$

where  $v(i, j)$  denotes the visibility of point  $i$  in camera  $j$  and  $P$  is the projection operation with homogeneous division. This optimization problem is non-convex and as explained by ? naïve optimization algorithms achieve only a poor local minimum. One approach to find a better local minimum is to initialize using an other SfM algorithm such as the factorization method.

## 1.5 2015 - PoseNet

The PoseNet, as proposed by ?, is a CNN that performs regression for the location and orientation of the camera given a single image as input. It outputs a 7D vector that describes the pose  $p = [x, q]^\top$  as the camera location  $x$  (3D) and orientation  $q$  (4D quaternion). They simultaneously learn location and orientation with the euclidean loss

$$\mathcal{L}(I) = \|\hat{x} - x\|_2 + \beta \left\| \hat{q} - \frac{q}{\|q\|} \right\|_2, \quad (1.7)$$

where  $\beta$  is a parameter to balance the expected error of pose and orientation. The architecture is a modified *GoogLeNet* with 23 layers which was trained on a classification task. They replace the last layers to perform regression instead of classification.

Because CNNs require a large amount of training data, the authors apply transfer learning by pretraining the network on a different task with large

datasets such as *ImageNet* and *Places*. Pose regression is performed using the pre-trained network and their own dataset called *Cambridge Landmarks* with five scenes of camera motion in large scale outdoor areas. For this limited dataset, ground truth pose is available. A challenge that comes with this dataset is the clutter in form of moving pedestrians and cars. Also, due to the long trajectories, weather and lighting conditions change a lot.

In their experiments they find that the system is robust to large spatial distance between camera samples. Also, from the visualization of the features they observe that the network does not only produce high response outputs for high level features, but is also sensitive to large textureless regions where SIFT-based approaches typically fail.

The paper demonstrates that transfer-learning can be used for pose estimation when a large labeled dataset for training is not available and that the pre-trained network can learn pose information despite being forced to produce pose-invariant outputs. **connection to other work?**

## 1.6 04/2016 - 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction

?

## 1.7 07/2016 - Unsupervised CNN for Single View Depth Estimation

The work of ? implements a autoencoder on stereo pair images to predict depth from a single image. Their architecture consist of the following parts.

- **Encoder**

Takes a single image as input. At training time, this is the left image of a stereo pair. The output is the predicted disparity map (scaled inverse depth).

- **Decoder**

The decoder is only used at training time. It takes two inputs: The predicted disparities from the encoder and the right image of the stereo pair. The right image is warped using the displacements and compared to the encoder input (left image) using the color constancy error (photometric error).

Since the stereo pair is rectified, the disparity is a displacement along the scan-line of the images. Thus the decoder implements a simple geometric transformation that does not need to be learned. At test time, only the encoder network with a single image as input is used.

As noted by the author, there are standard stereo algorithms that produce disparity maps. However, these methods can not deal with distortions such as lens flare, motion blur, shadows, etc. The idea is that a neural network could learn to deal with such problems that occur in natural images.

The dataset they use is KITTI ? which they augment by random crops, color channel scaling, and flipping the images.

## 1.8 04/2017 - SfM-Net

? implement a deep learning approach to structure from motion. Their architecture consists of two subnetworks:

- **Structure**

Learns per-frame depth. The input is a single frame. The output of the CNN is a cloud of 3D points, one for each pixel value in the input image.

- **Motion**

The input is a pair of frames. The CNN computes a set of  $K$  segmentation masks for moving objects. Using these masks, each pixel is assigned to an object specific transformation given by a rotation and a translation. In addition, camera rotation and translation are computed using the features of the inner layers.

Given a pair of images, the forward operation of the network works as follows:

1. The structure network computes the point cloud for the first frame.
2. The motion network computes the object transformations as well as the camera transformation using both frames.
3. The point cloud is transformed using the learned object transformations and masks.
4. The transformed 3D points are re-projected to 2D using the learned camera transformation between the two frames.

The transformed point cloud corresponds to the depth of the second frame. Optical flow can be computed directly from the re-projected points.

The authors of the paper propose various modes of supervision to evaluate the architecture and to handle ambiguities in the reconstruction due to the ill-posed problem:

- **Self-supervision**

No ground truth is given. The loss is defined by the brightness constancy constraint of the second frame warped to the first frame using the predicted optical flow. For this mode, they use the KITTI 2012/15 datasets.

- **Depth**

Ground truth is given in the form of depth for each pixel. This can be acquired for example by a Kinect sensor. They use the RGB-D SLAM dataset for ground truth depth. This helps to improve camera motion estimation.

- **Camera motion**

Camera motion is given as ground truth in form of a rotation and translation matrix. The relative transformation between predicted and ground truth transformation is



- **Optical flow and object motion**

They use this type of supervision with the MoSeg dataset which contains ground truth segmentation for each frame. This dataset contains more non-rigid body transformation. They evaluate the quality of the object motion mask by intersection over union (IoU).

## 1.9 04/2017 - Unsupervised Learning of Depth and Ego-Motion from Video

This work from ? is concurrent to the SfM-Net by ?. Both works share a similar concept. However, ? focus on unsupervised training and do not explicitly model scene dynamics. Similar to SfM-Net, their architecture consists of two jointly trained networks, one for learning depth from a single image (Depth-CNN), and another for learning camera motion from two frames (Pose-CNN).

The key difference to SfM-Net is that they train with an image sequence  $I_1, \dots, I_N$  with one image being the target image  $I_t$ . The task for the motion network is to synthesize the target image from two nearby views  $I_{t-1}$  and  $I_{t+1}$ .

- **Depth-CNN**

Takes the target image  $I_t$  as input and predicts a per-pixel depth map  $\hat{D}_t$ .

- **Pose-CNN**

Uses two source images  $I_{t-1}$  and  $I_{t+1}$  as input and predicts relative camera poses  $\hat{T}_{t \rightarrow t-1}$  and  $\hat{T}_{t \rightarrow t+1}$ . These transformations are then used to synthesize the target image.

Using both the estimated depth and pose matrix as well as the intrinsic matrix of the camera, the coordinate frame of the source images is warped to the coordinate frame of the target image. To transfer the color information from one frame to the other, bilinear interpolation is used. The general term for this is *differentiable image warping*.

When working with realistic data, it is important to make the model robust to non-static objects in the scenes, occlusions and non-Lambertian surfaces. To do this, the network is forced to additionally learn a per-pixel mask for each target-source pair. They add a regularization term to force the mask to have non-zero values.

## 1.10 04/2017 - DeMoN: Depth and Motion Network for Learning Monocular Stereo

? propose a network architecture to learn depth, motion as well as optical flow and surface normals from two images. The network architecture consists of three subnetworks in sequence:

- **Bootstrap network**

Consists of a pair of encoder-decoder networks. The first encoder-decoder predicts optical flow and its confidence. The second encoder-decoder takes the previously computed optical flow and confidence, as well as the original image pair and a warp of the second image using the estimated flow.

In addition, there are three fully connected layers predicting motion. By taking the image pair as input, the network can make use of motion parallax.

- **Iterative network**

The architecture is identical to the bootstrap network, but takes additional inputs. It is applied iteratively with the idea to improve previous estimates of depth, normals and motion. The network is trained using four iterations.

- **Refinement network**

Up-samples the low resolution predictions to the full image resolution.

Motion is parameterized using an axis-angle representation and depth is estimated by its inverse value  $1/z$  in order to represent points at infinity. For surface normals, motion and optical flow they use the euclidean loss and for depth L1. In addition to the point wise loss for depth, they define a scale invariant gradient loss that penalizes relative depth errors between neighboring pixels.

They evaluate the performance of the network on various datasets: SUN3D, RGB-D SLAM, MVS, Scenes11, and NYUv2. The results show that estimating normals and flow in addition to depth and motion slightly improves the performance of the system. Compared to works for depth from a single image, their results demonstrate that the network learns to use motion parallax from the stereo images to generalize better to scenes that heavily differ from the examples seen during training.

## 1.11 05/2017 - Relative Camera Pose Estimation Using Convolutional Neural Networks

? propose a CNN architecture to estimate the relative pose between two cameras. Similar to PoseNet ?, they estimate the 7D vector of pose by a quaternion and a translation vector. Their network is composed of two main parts.

- **Representation part**

The inputs are the two camera images. This part has two identical branches with shared weights of five convolutional layers with ReLUs and spatial pyramid pooling at the end. Each branch processes one of the input images.

- **Regression part**

Regression is performed on the output of the representation part using two fully-connected layers to estimate the relative pose.

The weights for the Siamese network are initialized from a pre-trained AlexNet on the *ImageNet* and *Places* datasets. In order to overcome the limitation in the input size due to the fully-connected layers, ? apply spatial pyramid pooling (SPP) in the representation part. This allows for a larger size of input images for the network to be able to extract more structural information that helps reconstructing the pose.

The dataset used for training is a crowd-sourced image collection with ground truth pose from a structure from motion pipeline based on local feature matching. For testing and comparison with other works they use the DTU dataset

where the exact pose is obtained from a robotic arm that moves the camera.  
**incomplete summary**

## Chapter 2

# Introduction

### 2.1 Relative Camera Pose

Equations to transform pose in world coordinates to relative pose with respect to the coordinate frame of the first camera.

$$R_{2 \rightarrow 1} = R_{w \rightarrow 1} R_{2 \rightarrow w} = R_{1 \rightarrow w}^{-1} R_{2 \rightarrow w} = R_{1 \rightarrow w}^{\top} R_{2 \rightarrow w} \quad (2.1)$$

$$t_{2 \rightarrow 1} = R_{1 \rightarrow w}^{\top} (t_2 - t_1) \quad (2.2)$$

### 2.2 Two-Frame Structure from Motion

### 2.3 Multiview Structure from Motion

Appendix A

Appendix

# List of Tables

# List of Figures

check spelling, capitalization and dates of references

# **Erklärung**

gemäss Art. 28 Abs. 2 RSL 05

Name/Vorname: .....

Matrikelnummer: .....

Studiengang: .....

Bachelor ☐      Master ☐      Dissertation ☐

Titel der Arbeit: .....

.....

.....

LeiterIn der Arbeit: .....

.....

Ich erkläre hiermit, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls der Senat gemäss Artikel 36 Absatz 1 Buchstabe o des Gesetzes vom 5. September 1996 über die Universität zum Entzug des auf Grund dieser Arbeit verliehenen Titels berechtigt ist.

.....

Ort/Datum

.....

Unterschrift