Literature review:

*Literature search strategy:*

* Research 3D reconstruction techniques, in particular - MVS and SfM. SfS is also of interest.
* Research traditional IBR / video interpolation techniques – as this are combined with deep-learning
* Research deep-learning architectures, in particular – convolutional networks architectures
* Research data and data formats
* Research specific to video interpolation using deep-learning – study in detail articles that have done similar research

Chapter 2 State of the Art

# 2.1 3D Reconstruction

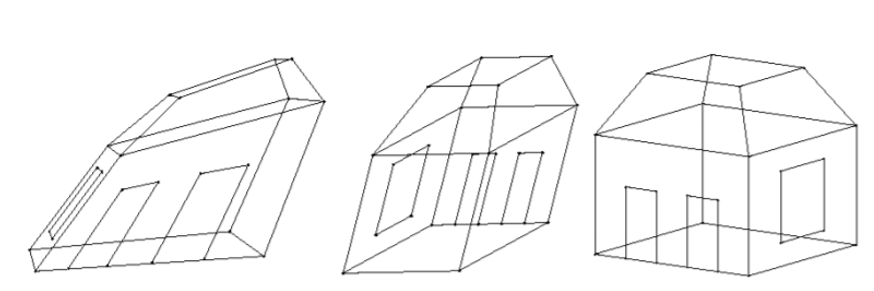
3D Reconstruction of object shapes from still images and video stream is an ongoing research topic that puzzled researches for decades. As the final aim of this paper is to improve the quality of the reconstruction and the final evaluation is also based on the reconstructed models, it is relevant to describe here the state-of-the-art of 3D Reconstruction methodologies. This will also justify the choices for the particular 3D reconstruction software used for the evaluation.

# 2.1.1 Orthographic projection

As early as 1992 (C.Tomasi and T.Kanade, 1992) obtained good results from a stream of images using orthographic, rather than perspective, projection and introduced *Factorization method*. The orthographic projection simplified processing, removing the depth dimension[[1]](#footnote-1). Examples of projective, affine and Euclidean projections are given in Figure 1. (Tomasi and Kanade, 1992) worked with affine and orthographic projections only. They decomposed the measurement matrix W (F frames, P tracked points forms 2\*F\*P matrix in 2D) into 2 matrices – R and S – representing the camera rotation and the object shape respectively plus the projection of the camera translation t along the image plane.

They were able to process input with *noisy measurements* by introducing 3x3 matrix Q (, ) and metric constraints to solve for Q.

They are also able to cope with *occlusions* by recovering position of feature points from 3 other positions of the feature (??Ulmann).



**Figure 1. Examples of Perspective, Affine and Orthographic projections** [Source: (Belongie, 2009, chapter 9)]

# 2.1.2 Perspective projection – epipolar geometry

When working in *perspective projection* several terms are useful:

*Calibrated cameras:*

First, a special case of calibrated cameras is described.

The *essential matrix E* for correspondence between 2 images was introduced by (Longuet-Higgins, 1981):

So, the same point in 2 images correspond as:

where is the position of the point in the second image, is the position of the same point in the first image and is the essential matrix – forming *epipolar constraint*. The essential matrix maps a point in the first image into a line in the second image - , which is called *epipole* (Szeliski, 2010).

Both translation and rotation of the 2nd camera – and traditionally, any subsequent cameras in a sequence of images – are taken with reference to the camera position of the first image in the sequence, i.e. the camera of the first image is the origin of the world co-ordinates and its orientation equals identity matrix.

If more than one feature point is available between the 2 images, the Essential Matrix can be determined from a series of equations: ⊗ , where ⊗ denotes point-wise multiplication, and i- is the index of the feature. The series of equations can be resolved with SVD (singular-value decomposition) algorithm. It has been shown by several researchers ((Hartley); (Torr and Murray, 1997); (Hartley and Zisserman, 2004)) that *7 point correspondences* (i.e. features) is sufficient to find the elements of the essential matrix (Szeliski, 2010).

In addition, (Hartley, 1995) suggests that the point co-ordinates should be translated and scaled to the centre of the object, so that the sum of x- and y- co-ordinates is 0 and the squared sum of both co-ordinates equals twice the number of points (, )

*Uncalibrated cameras*

The above equations describe *an ideal case*, where cameras are *perfectly* *calibrated*. The assumption of un-calibrated cameras adds an additional complexity of the calibration matrix K. The essential matrix becomes the fundamental matrix F:

Where K – is the camera calibration matrix. Or, , where e – is the *focus of expansion* and matrix is one of many possible *homographies* ((Hartley and Zisserman, 2004), (Faugeras, 1992)).

*Calibration Matrix*

While it is possible under certain constraints to convert projective reconstruction into a metric one, i.e. recover calibration matrices associated with each image – *self-calibration* (Hartley and Zisserman, 2004), most 3D reconstructions assume *pre-calibrated cameras* or images taken with a single camera with fixed intrinsic parameters.

where F is focal length, and – size of the camera sensor per pixel, and – translation of the camera centre with regard to the image. (Source: lectures CS7GV4 by Aljosa Smolic)

*Bundle adjustment*

Two images with 7 point correspondences is sufficient to estimate the Fundamental matrix, so each new image or each new point-correspondence overdetermines the system. A cost function can be introduced that aims to minimize the re-projection error. The system of equations can be solved with a non-linear method. There are two options for bundle adjustment: this can be done incrementally – as each new image is added - or at the end of the process with all images.

*Triangulation*

The last topic to discuss in the basics chapter is triangulation. This is a method to estimate depth to the object after the Fundamental matrix is known. …

The fundamental matrix and epipolar correspondence lie at the heart of SfM algorithm, as any estimation will start from finding the matrix correspondence between the 2 images.

2.1.1 Structure from Motion

The following two sections describe SfM (Structure from Motion) and MVS (Multi-view stereo) algorithms. These correspond to obtaining sparse 3D point cloud reconstruction and camera positions (SfM) and dense 3D surface reconstruction (MVS). The project will use COLMAP for 3D reconstruction, so particular implementation specific to COLMAP is specified.

The typical workflow of a SfM algorithm is as follows (Schonberger et al., 2016)

//from (Smith et al., 2016):

1. *Feature extraction and feature descriptors*. Correspondence between images is found based on distinctive points, so the first step of SfM is to identify feature points and their descriptors for each image in the stream.

One of the early methods for finding the interesting points is - Autocorrelation function (ACF) (Szeliski, 2010), which finds if the point is unique in its surroundings. The following authors further expanded on ACF - (Harris and Stephens, 1988), (Lucas and Kanade, 1981), (Shi and Tomasi, 1994) etc.

Suitable features are then described in terms of their neighbourhood – this is to ensure rotation, scaling, perspective distortions, lighting changes etc invariance.

The proposed algorithms create a description of the point’s neighbourhood:

* SIFT (Lowe, 1999, 2001, 2004)
* SURF (Bay et al., 2008)
* BRIEF (Calonder et al., 2010)
* ASIFT (Morel and Yu, 2009)
* LDAHash (Strecha et al., 2012)

COLMAP uses RootSIFT for its feature extraction, which is an improved version of SIFT algorithm.

For this work, which intends to produce additional images based on deep-learning, it is possible to have blurriness in generated images and ghost artefacts, which may make feature detection more difficult or place features incorrectly.

1. *Feature Matching*

The above features are matched. The simplest approach is to test every image pair and for every feature in the first image to find the most similar feature in the second image (*similarity metric*).

This approach has computational complexity O(N2IMAGESN2FEATURES) and is prohibitive for large image collections (Schonberger et al., 2016). There is research to improve the efficiency of the matching, for example, k-dimensional trees and ANN (Approximate Nearest Neighbour) (Smith et al., 2016).

1. *Identifying geometrically consistent matches*

Some feature matches may be excluded when they are checked for possible geometric transformations (homography, fundamental and essential matrices). If a valid transformation maps a sufficient number of features between the images, they are considered geometrically verified. RANSAC algorithm is usually used for the outlier detection (Schonberger et al., 2016).

1. *Initialisation before reconstruction and image registration*

SfM chooses the appropriate initial pair of cameras that would represent the origin of the world co-ordinates. Typically, these will have many common features and a wide baseline (Smith et al., 2016).

Also, the order in which the images will be added is important. New images can be registered to the current model by solving the Perspective-n-Point (PnP) problem. The PnP problem involves estimating the pose of the camera for the new image and, for uncalibrated cameras, camera’s intrinsic parameters. Every new image provides additional 2D-3D correspondences. (Schonberger et al., 2016).

1. *Triangulation*

Triangulation method is used to compute 3D space point from feature point correspondence (. Again, several different methods are proposed. (Hartley and Zisserman, 2004) describe triangulation suitable for different types of transformations (affine, projective etc). They discuss the differences between linear triangulation method (DLT, inhomogeneous) , error minimisation, Sampson approximation and solving a 6-degree polynomial. COLMAP developers (Schonberger et al., 2016) propose their own version of the triangulation method.

1. *Bundle Adjustment*

Bundle adjustment minimizes the reprojection error as more 2D-3D correspondences are added to the system. It performs a joint non-linear refinement of parameters Pc (Camera position and intrinsic parameters) and point positions – X. COLMAP uses the following formulae, where E – reprojection error, xj – co-ordinates of the point in image j, – loss function to down-weight the outliers and symbolises the function that converts scene points into image space.

The output of the SfM stage is a *sparse, unscaled 3D point cloud in arbitrary units along with camera models and poses*. This can be resolved into metric reconstruction if camera calibrations are known, or if metric parameters of some of the points are known (for example, ground-control points in case of georeferencing - (Smith et al., 2016)) .

2.1.1 Multi-view stereo

Multi-view stereo (MVS) provides a complete 3D reconstruction or *dense modeling* of the object from a known sparse cloud and known camera positions and intrinsic matrices.

--------------------Remove

Previous dissertations can be found at- <https://www.scss.tcd.ie/publications/theses/diss/>

---------------------------

MVS (Multi-view stereo)

SfM (Structure from motion)

- Structure from Motion

- Multi-View Stereo

- Poisson Surface Reconstruction

The 3D reconstruction technology based on multi-view is composed of techniques, such as

* feature point extraction and matching
* camera calibration
* sparse point cloud reconstruction
* dense point cloud reconstruction
* Poisson surface reconstruction
* texture mapping

Tomasi C, Kanade T. Kanade, T.: Shape and motion from image streams under orthography: A factorization method. Int. J. Comput. Vis.9(2), 137-154[J]. International Journal of Computer Vision, 1992, 9(2):137-154.

Marc PollefeysPollefeys M. Self-calibration and metric 3d reconstruction from uncalibrated image sequences[J]. Thesis K.u.leuven Departement Esat Afdeling Psi.phd.thesis, 1999.

Orthography – orthographic projection

An image stream can be represented by the 2F×P measurement matrix of the image coordinates of P points tracked through F frames. We show that under orthographic projection this matrix is of rank 3. [Tomasi and Kanade ]. singular-value decomposition technique to factor the measurement matrix into two matrices which represent object shape and camera rotation respectively

3D reconstruction begins with feature point extraction and matching (Suppose that we have tracked P feature points over F frames in an image stream. [T&K])

FAUGERAS, O. D. 1992. What can be seen in three dimensions with an uncalibrated stereo

rig? : Second European Conference on Computer Vision (ECCV’92).

HARRIS, C. & STEPHENS, M. J. 1988. A combined corner and edge detector. *Alvey*

*Vision Conference.*

HARTLEY, R. & ZISSERMAN, A. 2004. *Multiple View Geometry in Computer Vision*.

HARTLEY, R. I. In defence of the 8-point algorithm. Proceedings of IEEE International Conference on Computer Vision. 1064-1070.

HARTLEY, R. I. In defence of the 8-point algorithm. Proceedings of IEEE International Conference on Computer Vision, 1995. 1064-1070.

LONGUET-HIGGINS, H. C. 1981. A computer algorithm for reconstructing a scene from two

projections. *Nature.*

LUCAS, B. D. & KANADE, T. 1981. An iterative image registration technique with an

application in stereo vision. Seventh International Joint Conference on Artificial

Intelligence (IJCAI-81).

SCHONBERGER, J. L., FRAHM, J. M. & IEEE. 2016. Structure-from-Motion Revisited. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 27-30 2016 Seattle, WA. NEW YORK: Ieee, 4104-4113.

SHI, J. & TOMASI, C. 1994. Good features to track. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

SMITH, M. W., CARRIVICK, J. L. & QUINCEY, D. J. 2016. Structure from motion photogrammetry in physical geography. *Progress in Physical Geography,* 40**,** 247-275.

SZELISKI, R. 2010. *Computer Vision:*

*Algorithms and Applications*, Springer.

TOMASI, C. & KANADE, T. 1992. SHAPE AND MOTION FROM IMAGE STREAMS UNDER ORTHOGRAPHY - A FACTORIZATION METHOD. *International Journal of Computer Vision,* 9**,** 137-154.

TORR, P. H. S. & MURRAY, D. 1997. The development and comparison of robust methods for estimating the fundamental matrix. International Journal of Computer Vision.

1. Orthographic projection is applicable when the distance from the object to the camera (Zavg) is more than 10 times the object’s width davg: Zavg ≥ 10 \* davg (Belongie, 2009, chapter 9) [↑](#footnote-ref-1)