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

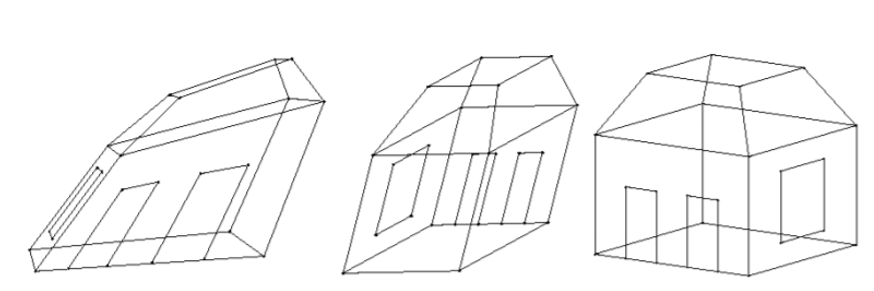
3D Reconstruction of object shapes from still images and video stream is an ongoing research topic that puzzled researches for decades. As the final aim of this paper is to improve the quality of the reconstruction and the final evaluation is also based on the reconstructed models, it is relevant to 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.

2.1.3 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., 2016a)

//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., 2016a). 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., 2016a).

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., 2016a).

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., 2016a) 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 ((Schonberger et al., 2016a).

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 modelling* of the object from a known sparse 3D cloud and known camera positions and intrinsic matrices. (Smith et al., 2016) summarises the review of MVS methods by As detailed by (Schonberger et al., 2016a) with reference to (Seitz et al., 2006), there is a wide variety of MVS algorithms, which can be classified into:

**TODO – add references**

1. Voxel-based methods which are 3D grids that are occupied to define the scene (for example, Seitz and Dyer, 1999)
2. Surface evolution-based methods that use iteratively evolved polygonal meshes (for example, Furukawa and Ponce, 2009)
3. Depth-map merging methods where individual depth maps showing the distance between the camera viewpoint to the 3D scene objects are combined into a single model (for example, Li et al. , 2010a)
4. Patch-based methods where collections of small patches or surfels represent the scene (for example, Lhuillier and Quan, 2005).

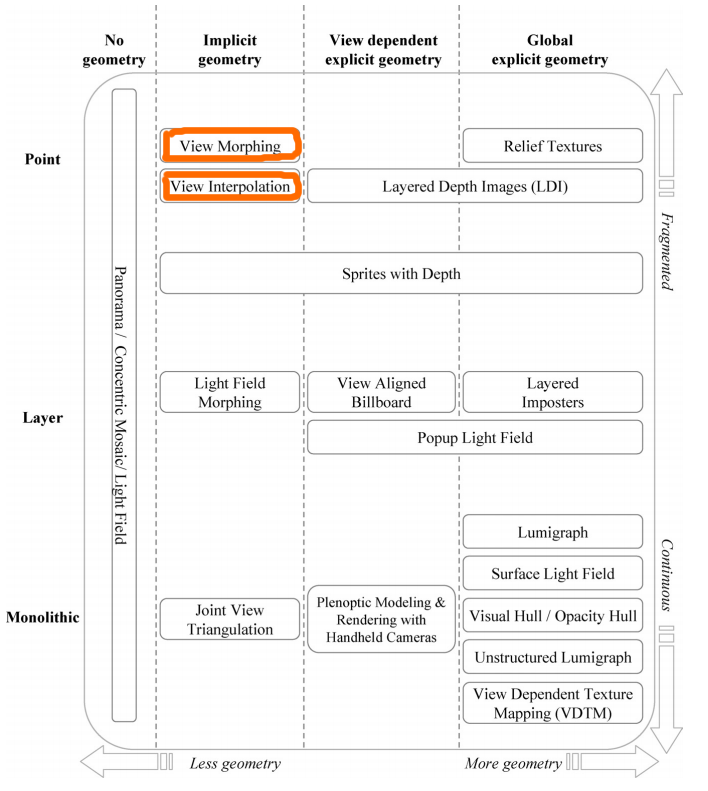
The latter method is also called PMVS and is the one utilised by COLMAP (Schonberger et al., 2016b). Schonberger suggests a new PMVS algorithm based on (Zheng et al., 2014).

The last step in Multi-view stereo is to generate a polygonal 3D mesh from dense point cloud. This can be achieved with Poisson Surface Reconstruction (PSR) (Kazhdan et al., 2006)

It can be seen that 3D reconstruction is a multi-step process, where different methodologies can be selected at every step. The particular choice of algorithms for each stage will strongly affect the accuracy of reconstruction, we have therefore specified the particular algorithms applicable to COLMAP, which will be used for 3D testing.

# 2.2 View interpolation

(Kang et al., 2007) proposes the following classification of Image-based rendering (IBR) techniques. View interpolation and view morphing are on the left of the continuum as relying on rendering with implicit geometry and acting on pixel-per-pixel basis.



# **Figure . IBR techniques classification(Kang et al., 2007)**

Motion interpolation and [view synthesis](http://apps.webofknowledge.com.elib.tcd.ie/OneClickSearch.do?product=WOS&search_mode=OneClickSearch&excludeEventConfig=ExcludeIfFromFullRecPage&colName=WOS&SID=E6lvDoNxLTSOwVpPNBU&field=TS&value=view+synthesis&uncondQuotes=true) are popular terms more recently that are also of interest.

*Single image interpolation techniques*

We will be interested here primary in the view interpolation between 2 views, but interpolation techniques are also applicable to *a single image*. These belong to a field of digital image processing and are required when images are re-sized, rotated or transformed. It is worth mentioning these interpolation techniques - as they can also be applied to the result of 2 frame interpolation. The interpolation techniques for a single image include: 2D nearest-neighbour interpolation, bilinear interpolation (2x2 neighbourhood), bicubic interpolation (4x4 neighbourhood), spline and sinc interpolation, natural neighbour interpolation using Voronoi cells, kriging based on Gaussian distribution (Wikipedia). When applied these can produce artefacts in the interpolated images: aliasing, blurring, edge halo (<https://www.cambridgeincolour.com/tutorials/image-interpolation.htm>). These can be rectified with anti-aliasing, interpolation that is “edge-aware” or “weighted edge-aware” (<http://techtidings.blogspot.com/2012/01/demosaicing-exposed-normal-edge-aware.html>).

*Interpolating between 2 views*

Coming back to the interpolation between 2 views – these belong to two broad categories:

* Spatial interpolation of image sequences, when camera position changes and the objects are static.
* Temporal interpolation of image sequences (video interpolation) – when objects in the images/ frames can move. In practice, video interpolation may combine both the moving camera and non-static objects hence incorporating special interpolation.

The simplest technique for *interpolation between 2 images* is *linear interpolation* where the intermediate pixel can be calculated at any intermediate point with:

Where and are pixel values in image 1 and 2. This method produces blurry result.

*Feature based image morphing*

An improved version of this algorithm applicable to *spatial interpolation* was proposed early on by (Chen and Williams, 1993). They worked with computer graphics images in order to improve the speed of generating CG views. Their technique first determined *pixel-by-pixel correspondences* between images and stored morph maps for further calculation. When required positions and colors of the points were linearly interpolated. As the authors worked with synthetic images - range data and the camera transformations were readily available. They were able to synthesize arbitrary intermediate points of view with bi-directional mapping. The new views only had view-independent shading.

(Seitz and Dyer, 1996) worked with natural images and assumed known camera projections. They expanded on view morphing techniques aiming to keep the shape of 3D objects. The authors first resolve the case of parallel views and prove that for parallel views with orthographic projection linear interpolation between feature points produces the correct result. For non-parallel views,   
*image reprojection* was used – this allows to move the image to a different plane using homography matrix. Their algorithm is composed of 3 steps:

1) Pre-warping – applies reverse homography camera matrices to the 2 images, to bring the images to a single plane and all 3 cameras to a single line.

2) Morph – linearly interpolate position and colors between 2 images.

3) Post-warping - apply homography of the target image camera to obtain the final view.

The work by (Seitz and Dyer, 1996) had a big influenced on subsequent research.

Before proceeding to the more recent interpolation techniques, two methods applicable to spatial view interpolation need to be described here: forward mapping and backward (inverse) mapping.

*Forward Mapping*

Forward mapping maps each pixel on the reference view(s) to the target view using some form of geometry, e.g., depth map (explicit geometry) or correspondences between views (implicit geometry). If – 2D point in the target image, – 2D point in the reference image, – point in 3D space, and – camera positions for reference and target images, and – camera projections, and – scaling factors.

The resultant pixel in the target image can be evaluated from the above equation. There is a problem with this approach – not all pixels in the target image may be populated and therefore will need to be interpolated, or at the same time as many pixels may land on the same pixel in the target image. Even after the interpolation there may still be holes due to magnification and disocclusion. (Kang et al., 2007)

Therefore, it is more traditional approach to use is the reverse of forward mapping:

*Inverse mapping*

In inverse mapping the pixel mapping in the target is found by tracing the ray from the target view back to the reference view:

Or, expressed in terms of homography :

Where defines the 2D planar perspective transformation from target screen to reference camera, is the epipole, is a scale factor and therefore defines epipolar line.

Inverse mapping ensures that there are no gaps in the target image, however, if is occluded in the reference view - the search yields no result. (Kang et al., 2007)

More recent references on *view synthesis* are able to work with un-calibrated cameras. (Fragneto et al., 2012) expands the work of (Seitz and Dyer, 1996), work with uncalibrated cameras and proposes a new algorithm based on interpolating homographies rather than pixel positions and colours.

*Temporal interpolation*

For temporal interpolation the most common approach is to use *optical flow*.

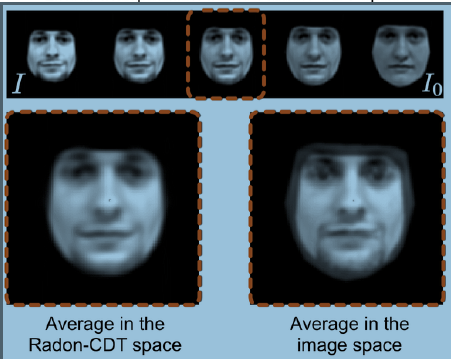
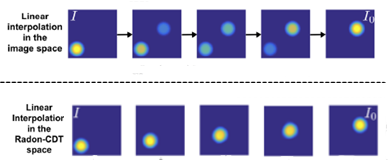
?? filing the holes after?

(Gurdan et al., 2014)

Navigating between camera views

dense image interpolation.

Methods of *non-linear interpolation* that do not include deep-learning have advanced recently. For example, (Kolouri et al., 2016) proposes linear interpolation in Radon Cumulative Distribution Transform space, where the interpolated image is still linearly separable into the 2 original images. The pixel location information is encoded in transport flows (*optimal transport metric*), so each pixel and neighbourhood are considered from ‘Lagrangian’ point of view. The transform captures translation and scaling, as well as more complicated transformations – see example of capturing movement and face interpolation below.



**Fig. Comparing linear interpolation results with non-linear interpolation using Radon-CDT space (linear interpolation images at top in the left example and bottom-right in the right example (Kolouri et al., 2016)**

?? PCA for image interpolation

An alternative approach to PCA (Principal Component Analysis) for image interpolation, is Isomap, or *dimensionality reduction* problem. This approach works well for repetitive motion, i.e. camera rotating around a rigid object, person waving hand etc (TODO – add reference ROBASZKIEWICZ Interpolating images between video frames). The method finds feature point correspondences between the interpolated images and interpolating a curve between the data points in the feature space, before fitting the intermediate images to the curve. The method works well for preserving the original shape of the object.

video interpolation with dense optical flow (1993)

based on motion estimation

Frame interpolation for video is a special case of image-based rendering (IBR) where middle frames are interpolated from *temporally* neighbouring frames (Niklaus et al., 2017).

Important properties of the required interpolated frame:

* Keeping edges
* Sharp
* No ghost artifacts

# Edge-Aware Interpolation - People notice edges…..exaggerated!

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

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

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

- 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

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.

CHEN, S. E. & WILLIAMS, L. 1993. View Interpolation for Image Synthesis. Proceedings of the 20th annual conference on Computer graphics and interactive techniques: SIGGRAPH

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

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

FRAGNETO, P., FUSIELLO, A. & SOC, I. C. 2012. Uncalibrated View Synthesis with Homography Interpolation. *Second Joint 3dim/3dpvt Conference: 3d Imaging, Modeling, Processing, Visualization & Transmission (3dimpvt 2012)***,** 270-277.

GURDAN, T., OSWALD, M. R., GURDAN, D. & CREMERS, D. 2014. Spatial and Temporal Interpolation of Multi-view Image Sequences. *Pattern Recognition, Gcpr 2014,* 8753**,** 305-316.

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.

KANG, S. B., XIN TONG, Y. L. & SHUM, H.-Y. 2007. Image-Based Rendering.

KAZHDAN, M., BOLITHO, M. & HOPPE, H. 2006. Poisson Surface Reconstruction. Eurographics Symposium on Geometry Processing.

KOLOURI, S., PARK, S. R. & ROHDE, G. K. 2016. The Radon Cumulative Distribution Transform and Its Application to Image Classification. *Ieee Transactions on Image Processing,* 25**,** 920-934.

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).

NIKLAUS, S., MAI, L. & LIU, F. 2017. Video Frame Interpolation via Adaptive Convolution. *CoRR,* abs/1703.07514.

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 2016a Seattle, WA. NEW YORK: Ieee, 4104-4113.

SCHONBERGER, J. L., ZHENG, E. L., FRAHM, J. M. & POLLEFEYS, M. 2016b. Pixelwise View Selection for Unstructured Multi-View Stereo. *Computer Vision - Eccv 2016, Pt Iii,* 9907**,** 501-518.

SEITZ, S. M., CURLESS, B., DIEBEL, J., SCHARSTEIN, D. & SZELISKI, R. 2006. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

SEITZ, S. M. & DYER, C. R. 1996. View Morphing. *SIGGRAPH,* Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques.

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.

ZHENG, E., DUNN, E., JOJIC, V. & FRAHM, J. M. 2014. Patchmatch based joint view selection and depthmap estimatio.

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)