VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

VYHLEDÁVÁNÍ KORESPONDENCE V STEREOSNÍMCÍCH TVÁŘÍ

STEREOSCOPIC FACE IMAGES MATCHING

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — Bc. MARTIN KLAUDÍNY
AUTHOR

VEDOUCÍ PRÁCE — doc. PAVEL ZEMČÍK
SUPERVISOR

BRNO 2008

Abstrakt

Táto diplomová práca sa zaoberá problematikou 3D skenovania ľudských tvárí. Väčšina súčasných sytémov pre skenovanie tvárí využíva projekciu vzoru. Cieľom diplomovej práce je vyvinúť 3D skenovací systém pre tváre založený na pasívnej stereoskopickej fotogrammetrii. Fotografie tváre s vysokým rozlíšením sú zachytené párom kalibrovaných fotoaparátov. Predspracovanie snímkov pozostáva z odstránenia deformácie optickým systémom, rektifikácie obrazového páru a extrakcie tvárovej časti. Dôraz je kladený na párovanie stereosnímkov. Tri rôzne prístupy k vyhľadávaniu korešpondencie v stereosnímkoch sú prezentované - lokálna technika, globálna technika používajúca rez grafom a hybridná technika spájajúca predošlé dve. Nakoniec je 3D model zrekonštruovaný z nájdenej korešpondencie. Výsledky ukazujú, že kvalitné 3D modely tváre môžu byť získané napriek tradičným ťažkostiam s párovaním stereosnímkov tváre bez vzoru. Taktiež je možné udržať výpočtové nároky na akceptovateľnej úrovni, hoci majú spracovávané snímky vysoké rozlíšenie.

Klíčová slova

počítačové videnie, spracovanie obrazu, skenovanie 3D objektov, rekonštrukcia tváre, pasívna stereoskopická fotogrammetria, stereoskopická korešpondencia

Abstract

This thesis is dedicated to the problem of 3D face capture. A majority of current face capture systems exploits a pattern projection. The goal of thesis was to develop the 3D face capture system based on the passive stereo photogrammetry. The high-resolution images of a face are captured by one pair of the calibrated digital still cameras. The image pre-processing consists of the removal of lens distortion, rectification of image pair and face extraction. The stereo image matching is accentuated. Three different approaches to the stereo correspondence search are presented - the local technique, the global technique using a graph cut and the hybrid technique merging previous two. At last, the 3D model is reconstructed according to the found correspondence. The results show that the high-quality 3D face models can be obtained despite traditional difficulties with matching pattern-free face images. Also it is possible to keep the computational demands on the acceptable level, although the high-resolution images are processed.

Keywords

computer vision, image processing, 3D object capture, face reconstruction, passive stereo photogrammetry, stereo correspondence

Citace

Martin Klaudíny: Stereoscopic face images matching, diplomová práce, Brno, FIT VUT v Brně, 2008

Stereoscopic face images matching

Prohlášení

Prehlasujem, že som túto diplomovú prácu vypracoval samostatne pod vedením doc. Dr. Ing. Pavla Zemčíka. Ďalšie informácie poskytol Prof. Adrian Hilton (University of Surrey, Veľká Británia). Uviedol som všetky literárne pramene a publikácie, z ktorých som čerpal.

. . . .
Martin Klaudíny
May 16, 2008

Poděkování

I would like to thank Prof. Hilton (University of Surrey, UK) and doc. Zemčík for their supervision in different phases of the work. I am also grateful to Dr. Jean-Yves Guillemaut and Dr. Jonathan Starck (University of Surrey, UK) for discussions about graph cuts. The construction of capture rig would be impossible without a help from Mr. Andrew Birt (University of Surrey, UK).

Contents

Chapter 1

Introduction

A computer vision belongs to the most intensive research areas in the artificial intelligence. Although the ultimate goal to give a machine an ability to see is still distant, some of the fundamental problems have been deeply researched. The developed solutions are practically useful not only in the computer vision but in many other areas. The acquisition of 3D digital models of objects is one of the active research problems. Practical outcomes of this research are deployed in wide range of applications. The reconstruction of 3D object is valuable for automated production lines, a new product design, archiving objects with historical value etc.

A common application is also a 3D capture of human face. The digital representation of human face is essential in several areas. For example, 3D models of real people and their dynamics can be used as foundations for realistic character animation in a film production or game development. The face model can be displayed as an avatar in communication interfaces. The accurate face shape representation can find a use in an anthropometry or medical applications. It is also useful for a face recognition in security systems.

The 3D capture of face has specific constraints and requirements with respect to the character of examined object. The issues connected with a safety, speed and natural behaviour of a person make the 3D face capture system more complex than general-purpose capture systems. To address these requirements the face capture systems usually reconstruct 3D model from the images of face. The majority of the systems are based on the projection of structured light pattern or the stereoscopic[1] photogrammetry. The structured light techniques capture the face illuminated by a known pattern. The detection of pattern in an image together with known spatial relationship between a camera and projector enable the reconstruction of 3D model. The disadvantage is a loss of natural appearance of face.

The principle of stereo photogrammetry is inspired by the human visual system. A face is captured by a stereo pair of cameras. A spatial point is observed in different position in each image. This difference conveys the 3D information. The fundamental task is to find the corresponding image points which are observations of same 3D points. The stereo image matching is complicated by the uniform appearance of skin. Therefore, the random pattern is often projected on a face to simplify the search for corresponding image points. It brings again the problem with an inability to record the appearance of face. But this problem has been overcome in recent years by quick switching between the shape and appearance capture or projecting pattern in the non-visible part of electromagnetic spectrum.

[1] The short term stereo is widely used in a literature instead of stereoscopic.

A research in this field has reached the point where complete facial dynamics together with a speech can be recorded in video rate and subsequently converted into 3D representation. The main limitation of current face capture systems is a complex capture environment with the controlled light conditions. The capture rig has to provide a regulation of illumination for correct pattern projection and observation. Moreover, it contains a specialised equipment when the stereo matching is shifted to the non-visible part of electromagnetic spectrum. A research challenge is to develop the capture system with same functionality and reliability which would be able to operate under natural illumination. This step would lead to the transition of 3D face capture systems from laboratory conditions to the real environment.

This thesis is addressed to this trend. The goal of thesis is to develop the 3D face capture system based on the passive stereo photogrammetry. Therefore, the pattern projection is not involved during capturing. The images of face are recorded by one pair of digital still cameras with high resolution. The focus is aimed on the stereo image matching as a fundamental task in this type of capture system. The matching algorithms suitable for a human face are developed. They are integrated into full processing pipeline of 3D capture system involving the image acquisition, image pre-processing and reconstruction of 3D model. The implemented system is evaluated in terms of the quality of 3D model and computational demands. The fact that the face images have high resolution offers an opportunity to investigate whether high image resolution can bring enough skin details to overcome traditional difficulties with matching pattern-free face images.

The thesis is structured into five chapters. Chapter 2 presents a background for the area of 3D capture. It summarises the current state of research in the field. The theoretical basics are provided for significant tasks within 3D capture system. This involves a calibration, view geometry and stereo correspondence. Chapter 3 is dedicated to the design of system. The problem is analysed at first. The identified major tasks are presented in standalone sections. The face capturing, camera calibration, image pre-processing, construction of disparity space, construction of disparity map and surface reconstruction are described. The implementation of system is explained in Chapter 4. It consists of the descriptions of capture rig, image acquisition and calibration procedure. The implementation of the developed software is briefly outlined as well. Chapter 5 presents the achieved results. The properties of the system are discussed. Finally, Chapter 6 provides the conclusions and contributions of the thesis. The possible improvements and extensions to the presented methods are outlined.

The thesis is a continuation of the Term project. The Chapters 2, 3 were partially adopted from the Term project report. A majority of the work in the thesis have been made during the study intership at the University of Surrey in United Kingdom. It was supervised by Prof. Adrian Hilton during that period.

Chapter 2

Background

The 3D object capture is a complex problem which requires a knowledge from various different fields. Mathematical methods together with information about physical phenomenons provide foundations for creating capture system. This chapter contains a brief overview of main approaches to the problem. The theoretical basics are presented for significant tasks which are performed by the 3D capture system. The described concepts are relevant for the explanation of system design.

2.1 3D capture

3D capture can be defined as the acquisition of an object's spatial representation and other characteristics of its surface. The centre of interest is the shape of object itself, hence its relationship to the environment such as position and orientation are not so important. This fact allows flexible choice of coordinate system, which the surface will be represented in. It is often important to capture object's appearance in addition to its shape. The most important property describing appearance is intuitively the surface colour.

A large number of different techniques have been developed to acquire surface shape of a particular object. All of them try to achieve the highest accuracy of reconstructed surface as possible together with reasonable speed of the whole process. From an economic point of view the technical demands and simplicity of control are significant as well. The most popular approaches are summarised in this section. The capturing techniques can be generally divided into two main groups with respect to the occurrence of direct contact with the object surface during the capturing procedure. The first group contains *contact techniques* and the second, their opposite, *non-contact techniques*.

In contact methods an interaction of the object with a scanning tool can be destructive or non-destructive. An example of a destructive variant is the cutting object into thin slices for their measurement. This approach is suitable in a small range of applications. Obviously non-destructive interaction is preferred, so the whole surface is scanned by touching it with a specialised tool (e.g. digitisation arm). The position of a 3D point is identical with the position of contact spike on the digitisation arm. The advantage of these techniques is high accuracy of points in coordinate space and capability to acquire the desired density of samples. Disadvantages are the need for high-precision specialised equipment, slow scanning process and applicability only on objects with insensitive and hard surface.

Non-contact methods scan an object from a distance exploiting various physical principles and phenomenons. The image of object examined is captured from the viewpoint of the sensor. These 2D data are then analysed and results of this analysis allow the determination of the positions of surface points. The capture with no contact is advantageous when an object has limited access or it is made from the material sensitive to the used physical phenomenons. Another advantage is shorter time of capturing than contact methods. However, demanding procedure of data analysis and surface reconstruction is required. Also lower precision of digital representation is achieved by many non-contact techniques in comparison to the contact approach.

However, a large number of non-contact methods, which exploit various physical principles, exists because the contact approach is not suitable in many applications. The majority of these methods capture the images of object in the visible electromagnetic spectrum and uses light phenomenons observable to the human eye. Therefore, a part of the capture equipment is always a passive component - camera. Moreover, various types of active components (e.g. projectors, special lights) can be added to the capture system to influence light conditions in an environment. The aim is to achieve particular light effect on the surface of object that provides more information about shape. The non-contact techniques may be categorised as *passive* or *active*. The active methods create special illumination conditions in a scene. In contrast, the passive methods use only passive sensors for capturing naturally illuminated scene.

The first active technique to be introduced is a *scanning with laser stripe* projected on the object. The deformed stripe on the surface of object is observed by the camera with known spatial relationship to the plane of laser light (see Figure 2.1). A curve of the stripe in the image is a projection of surface points lying on the laser plane. The 3D positions on these points are computed by a triangulation. The main problem are occlusions of the laser stripe on highly ragged surfaces. A disadvantage is also a time spent by progressive laser scanning. The methods using shadows are based on the same principle. Instead of laser stripe the edge of shadow volume is detected in the picture. But an accuracy is lower because the sharpness of shadow edge is not comparable with the laser beam.

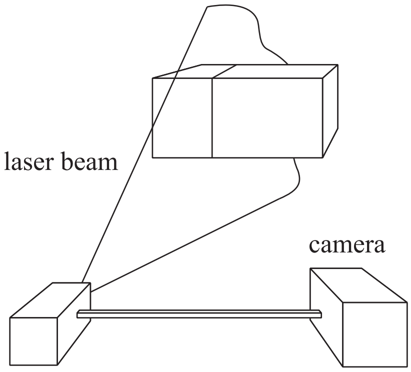


Figure 2.1: Laser scanning.

An extension of the concept used in the laser triangulation to the whole examined space is present in *structured light techniques*. A scene is illuminated by special light pattern which projection is easily recognised in the image. Usually a stripe pattern is projected because it divides a space into planes. This leads to the complex variant of laser scanning where many sets of surfaces points lie on many slightly translated parallel planes. The time of the scanning process is highly reduced because information about the shape of object is gathered simultaneously. It is also necessary to determine exact position of each point within the pattern itself from the picture. Each pixel, which is a part of the detected pattern feature in the image, must be correctly matched with the plane containing equivalent 3D point.

Various strategies are used to connect a pixel with corresponding plane. Scanned surfaces can be considered as smooth, therefore neighbouring projected stripes in the image are also adjacent in the space. This approach fails when a surface contain discontinuities. The second way is to differentiate stripes by colours. But it is not reliable when the object is textured. The last major strategy is multiplexing different patterns in a time. A pixel membership to the particular plane is coded by changing illumination in the sequence of patterns. A disadvantage is capturing several images in a time period which limits a usage on still objects.

Coding each plane in the image can be solved by a number of techniques. A pattern can simply contain sequence of n frames with one shifting stripe. In each image there is one deformed stripe which defines one plane. It is practically similar to laser scanning. Better concept are *binary Gray coded patterns*. The sequence of black and white patterns with various width of stripes is projected on the object. It is possible to code 2^n planes by n patterns. Matching a pixel to an appropriate plane is determined by its illumination state (yes/no) under different patterns. Except reduction of needed frames an advantage is also a resistance against errors in the pixel state. A property of Gray code is that successive codewords differ in one bit. A plane can be also coded by intensity of grey level. *Grey level patterns* cover object with changing grey spectrum. A simple ramp profile of grey intensity is sensitive to noise and anomalies in projected light, so sawtooth or sinusoidal profiles are used to improve localisation of pixels on planes with same intensity. The methods which combine Gray code with grey intensity exist as well.

A special case of structured light method which is able to capture slowly moving objects is described in [26]. Four black and white stripe patterns are repetitively projected into scene. The cutting planes are placed in boundaries of stripes. They are coded by the presence of illumination on both sides of boundary during a time. The movement of boundaries is tracked in sequence to join parts of codeword from each pattern.

The *moire techniques* also belong among the active techniques. They are based on an interference of two periodic patterns projected into a scene as it is depicted in Figure 2.2. A combination of patterns creates new periodic signal which cuts the volume of examined scene by planes with different intensity of illumination. It is similar to the structured light approach except for a fact that the planes are parallel to the camera. The planes create contours on the object surface. The period of signal defines depth difference between the image points on adjacent contours in the picture. This rule is violated when the surface contains rapid discontinuities.

Many passive techniques for the 3D capture have been developed but they are not so widespread as their active rivals. Main reasons are lower accuracy of object digital representation and lesser reliability. Both of them are consequences of higher complexity of 3D capture in naturally illuminated scene. The examples of passive methods are model reconstruction from texture deformation on the object, changing silhouette of rotating object, optical flow of surface points on the object in motion, etc.

The family of methods based on a *stereo photogrammetry* is influential in the area of 3D capture as well. It contains representatives from both groups - active and passive methods. A principle of stereo photogrammetry is copied from human visual system. Two images of object are captured from slightly different viewpoints. Known spatial relationship between these viewpoints and dissimilarities between the pictures provide enough information to

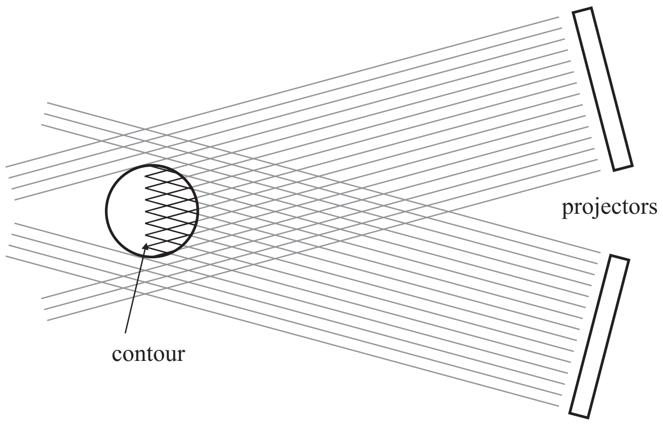


Figure 2.2: Moire technique.

resolve 3D representation of a scene. Basic situation in the stereo vision is illustrated in Figure 2.3. Same 3D point is observed by two cameras and its projections are situated in different pixels on image planes of cameras. The shift between positions of these pixels in the pictures is known as a *disparity*. It is obvious that the rays shot out from camera centres through the pixels intersect each other in the observed spatial point.

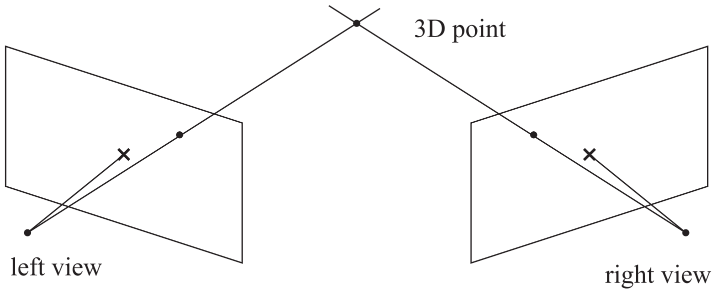


Figure 2.3: Stereo vision.

To exploit explained principle, the capture system based on the stereo photogrammetry has to perform several important steps. At first, it is necessary to calibrate the system. It means to establish a global coordinate system and determine a position and orientation of both cameras within it. In addition, the optical systems of cameras are mathematically described. It provides an opportunity to determine spatial ray shot through specified pixel to a scene. Next step is a search for pairs of pixels which are the projections of same 3D point in the images. This is a main problem in the stereo photogrammetry, usually referred as a *stereo correspondence problem*. Basic idea is that the areas around matching pixels in the pictures are similar when they are the projections of same 3D point. However, the appearance of same area on the surface of object is quite view-dependent in a practice. It is caused by a shape of the object and optical properties of its surface. Therefore, matching stereo images is difficult task. Last step is a reconstruction of spatial model of the object. The position of 3D point is computed using the pair of matching pixels and the description of system from calibration stage. Obtained surface points are organised into single 3D model.

The stereo photogrammetry may be classified as passive or active technique according to the used illumination in a capture space. The active variant simplifies the stereo correspondence problem by projecting a random pattern on the object. However, an appearance of object is contaminated by the projected pattern. The passive variant is more general because it works under usual light conditions. But the correspondence search is more difficult in this case, especially for objects without strong texture.

The presented concept of stereo vision is often extended to more views. Advantage of multiple views can be exploited in two major ways. Firstly, the views are grouped in pairs and each pair is processed as standalone stereo pair. Afterwards, the surfaces obtained from all pairs are merged into single model. As a result, larger portion of object is modelled. Also the parts of surface which are incorrectly reconstructed by one stereo pair are substituted by better result from another pair. Secondly, information from multiple views can be used during the correspondence search together. This idea is exploited by volumetric techniques which partition a capture space into voxel grid. Each voxel is projected on the images from all viewpoints and a similarity of areas in the place of projections is examined. If the similarity is high across the set of views, particular voxel can be considered as a part of the surface of object.

2.2 Calibration

The initial stage of 3D capture process based on arbitrary technology is a *calibration*. The calibration of capture system consists of two main parts. The first part can be defined as a determination of the position and orientation with respect to the set world coordinate system for every part of equipment actively involved in capturing. Second part involves an acquisition of important technical parameters of particular appliances. The specification of the world coordinate system is essential because it creates a frame for building digital representation of object. But a complexity of the rest of stage varies according to the equipment used by the chosen method. In context of the stereo photogrammetry it is necessary to calibrate only cameras. Therefore, this section is focused on the camera calibration (partially adopted from [17]).

2.2.1 Camera calibration

The calibration of camera determines position and orientation of the camera in world coordinate system and computes parameters which describe camera optical system. Position and orientation are represented by *extrinsic parameters*. They define relationship between local camera coordinate system and world coordinate system. Camera coordinate system has usually origin in optical centre (also referred as pinhole). Extrinsic parameters consist of a *translation vector* and *rotation matrix* (or rotation angles). Together they define 6 degrees of freedom.

The way how scene is observed and projected on image plane in camera is characterised by *intrinsic parameters*. Graphical illustration of following terms is in Figure 2.4. First important property of optical system is a *focal length* what is distance between optical centre and virtual image plane in front of camera. Second group of parameters describes capturing the image by discrete array of sensors (CCD chip). Another set of intrinsic parameters is focused on a description of various imprecisions and abnormalities in the optical system. A skew between axes of sensor array can be present from a manufacture. The intersection between image plane and optical axis can deviate from the geometrical centre of picture. The actual position of intersection is called a *principal point*. A *radial lens distortion* is caused by a change of focal length with increasing distance from the optical axis. A *tangential lens distortion* reflects imprecise centring of lens system. Exact mathematical representation of intrinsic parameters depends on the used *camera model* and can simplify a physical reality.

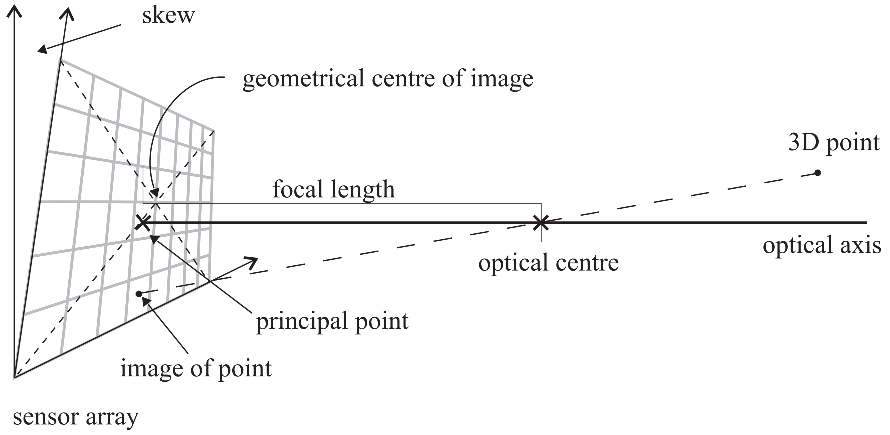


Figure 2.4: Camera optical system. The lens distortion is not illustrated.

Intrinsic and extrinsic parameters are mutually independent because the change of camera position does not influence the properties of optical system. Extrinsic parameters define a *transformation matrix* between world coordinate system and camera coordinate system. Intrinsic parameters form a *camera intrinsic matrix* which defines the conversion of 3D point coordinates in camera system to 2D pixel coordinates on image. The parameters connected with lens distortion cannot be included into this matrix because they describe non-linear operation. Composition of mentioned matrices is called a *projection matrix* and it determines complete transformation between 3D point in world coordinate system and 2D pixel coordinates of its image (without lens distortion deformation).

The existence of *calibration object* in observed scene is important for a computation of the projection matrix. A calibration object has known geometry and should be easily recognised in the image. Its geometrical features establish world coordinate system. It provides a number of reference points with easily determinable position in the world coordinate system. Because of that it is usually simple spatial primitive covered with periodic pattern. Its purpose is to provide the 3D world coordinates and image coordinates of reference points. These coordinate pairs are used for computing the projection matrix for particular view. The values of intrinsic and extrinsic parameters can be determined by decomposing the projection matrix. If calibration object is planar (all reference points are co-planar), calibration images from more viewpoints are required to determine all intrinsic parameters of the camera. Extrinsic parameters are determined from one principal view when the camera is set in its final position.

Acquired information in the calibration stage are utilised in the consequent processing steps. For instance the projection matrix is used for reverse transformation of image pixel on the ray passing trough this pixel from optical centre into a scene. Also spatial relationship between more cameras can be discovered with the assistance of their extrinsic parameters in the world coordinate system. Captured images may be undistorted by transformation defined by radial and tangential distortion coefficients. More details about the camera optical system and its properties in the context of calibration can be found in [11].

2.2.2 Camera model

The mathematical model of camera used in this project is introduced in following text. It is quite complex and comprises all types of intrinsic parameters mentioned in previous section. More theoretical background about this model can be found in [14]. The scheme of model is shown in Figure 2.5.

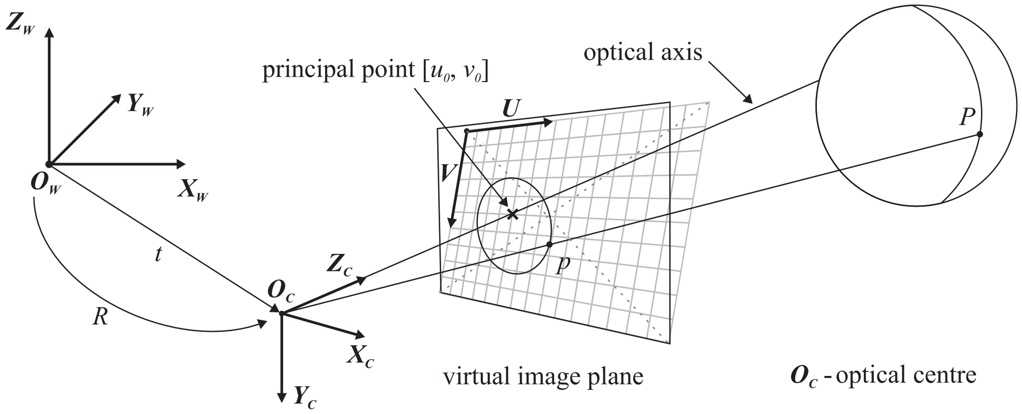


Figure 2.5: Camera model with important coordinate systems illustrated.

The extrinsic camera parameters are represented by rotation matrix R_W^C and translation vector t_W^C. According to Equation 2.1 they transform coordinates P_W of 3D point P in the world coordinate system to its coordinates P_C in the camera coordinate system.

$$P_C = R_W^C P_W + t_W^C \quad R_W^C = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \quad t_W^C = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \tag{2.1}$$

By using homogeneous coordinates (notation - tilde above vector) rotation matrix R_W^C and translation vector t_W^C can be assembled into a transformation matrix T_W^C as it is shown in Equation 2.2.

$$\tilde{P}_C = T_W^C \tilde{P}_W \quad \rightarrow \quad \begin{bmatrix} X_C \\ Y_C \\ Z_C \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} \tag{2.2}$$

Rotation matrix R_W^C has only 3 degrees of freedom and can be converted to 3 rotation angles [33].

The coordinates of spacial point in camera coordinate system are starting point for expressing actual projection through lenses. Firstly non-linear operations connected with pinhole projection and lens distortion are performed. Equation 2.3 shows the normalisation of vector P_C what is the part of basic formula for pinhole projection.

$$P_N = \begin{bmatrix} X_C/Z_C \\ Y_C/Z_C \end{bmatrix} = \begin{bmatrix} X_N \\ Y_N \end{bmatrix} \tag{2.3}$$

Both lens distortions - radial and tangential are considered. Radial distortion is specified by coefficients k_1, k_2 and k_5 and tangential distortion by k_3, k_4. Non-linear distortion transformation is defined by Equation 2.4.

$$R^2 = X_N^2 + Y_N^2$$
$$P_D = \begin{bmatrix} X_D \\ Y_D \end{bmatrix} = \begin{bmatrix} (1 + k_1R^2 + k_2R^4 + k_5R^6).X_N \\ (1 + k_1R^2 + k_2R^4 + k_5R^6).Y_N \end{bmatrix} + \begin{bmatrix} 2k_3X_NY_N + k_4(R^2 + 2X_N^2) \\ k_3(R^2 + 2Y_N^2) + 2k_4X_NY_N \end{bmatrix} \tag{2.4}$$

In Equation 2.5 normalised coordinates $\tilde{P_D}$ modified by lens distortion are multiplied by camera intrinsic matrix A. Result of this operation are final image coordinates $\tilde{p_D}$ of projected 3D point in pixels. Origin of image coordinate system is situated in upper-left corner of picture (see Figure 2.5).

$$\tilde{p_D} = A\tilde{P_D} \quad \rightarrow \quad \begin{bmatrix} U \\ V \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & \alpha f_x & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} X_D \\ Y_D \\ 1 \end{bmatrix} \tag{2.5}$$

Coefficients f_x and f_y are focal lengths for row and column direction of sensor array in pixel units. Their values can differ because sampling frequency in these directions can be different. The value of α encapsulates skew between axes of sensor array. The image coordinates of principal point consists of parameters u_0 and v_0 (in pixels).

When distortion transformation is excluded, the projection matrix for camera may be defined. It is created by composition of camera intrinsic matrix A and transformation matrix T_W^C (3x4 variant). Equation 2.6 shows that projection matrix H converts world point coordinates to scaled vector of image coordinates.

$$\tilde{p} = (AT_W^C)\tilde{P_W} = H\tilde{P_W}$$
$$\begin{bmatrix} wU \\ wV \\ w \end{bmatrix} = \begin{bmatrix} f_x & \alpha f_x & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \end{bmatrix} \cdot \begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} \tag{2.6}$$

2.3 View geometry

The model of single camera which is situated standalone in the world coordinate system was discussed in Section 2.2.2. The addition of another cameras into this system does not bring only new set of views at the scene but also new principles which are inferred from mutual relationships between cameras. These principles are independent of observed scene and define options and limits of scene spatial reconstruction.

This section explains the geometry in the system with one pair of cameras and methods how the scene can be reconstructed. Presented principles are essential for stereo photogrammetry techniques. A comprehensive text about the geometry of binocular system and systems with multiple cameras is [13].

2.3.1 Epipolar geometry

The general stereo configuration is illustrated in Figure 2.6 (the left or the right are defined from the position behind camera towards to an object). The spatial point P is projected into point p_L on the image plane of left camera and point p_R on the image plane of right camera. Therefore, the points p_L and p_R are corresponding to each other. The optical centres of cameras are denoted as O_L and O_R. They are connected with a *baseline*. The *epipolar plane* is defined by baseline and observed point P. The *epipolar lines* l_L, l_R lie in the intersections of this plane and the image planes. The baseline intersects the image planes in points e_L and e_R called *epipoles*. The epipole e_L is the projection of optical centre of right camera (equivalently in opposite direction). Every epipolar plane contains the baseline with epipoles without regard to the choice of 3D point. This fact gives epipoles the property that every epipolar line passes through them.

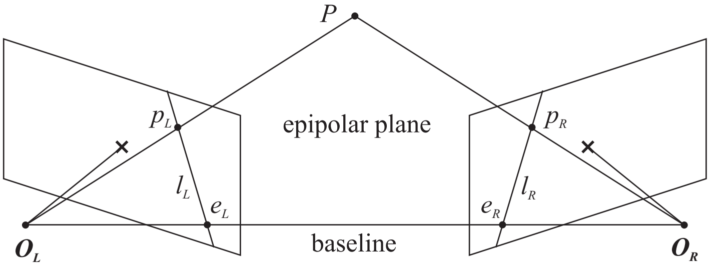


Figure 2.6: General stereo configuration.

The epipolar plane can be also established by the baseline and the ray shot from optical centre of one camera through a point on its image plane. Consider the ray from O_L through p_L for example. It can be seen that every spatial point lying on this ray is projected on the image point p_L such as point P. All these 3D points can be observed by the right camera and their projections belong to single epipolar line l_R. The l_R lies on the epipolar plane defined by the ray and the baseline as well. It means that every point in the right image potentially matching to p_L is situated on l_R (proved by correct correspondence p_R). This idea is known as *epipolar constraint* and it is independent on the observed scene or the parameters of used cameras.

2.3.2 Image rectification

A *canonical stereo configuration* is the special case of general stereo configuration. Both image planes are co-planar, align to each other and parallel to baseline as it is depicted in Figure 2.7. The images do not contain any epipoles because there are no intersections of image planes with the baseline. The epipolar lines l'_L, l'_R associated with corresponding points p'_L, p'_R are located on same line parallel to the baseline.

In the canonical stereo configuration many calculations are simpler than in general variant. Because of that it is usually worth to convert the general stereo configuration to canonical form. Practically it means to transform the stereo pair of images from general configuration to the pair which would be obtained in canonical configuration. This transformation is called a *image rectification*. It performs the projection of original image planes onto new common image plane. Only constraint put on the position of this plane is a parallelism with baseline. The direction of plane normal is usually perpendicular to the intersection line between original image planes. The most straightforward projection can be made in this case. A distance from baseline is not very important and it only influences scale of projected images. The coordinate systems of new images on common plane should

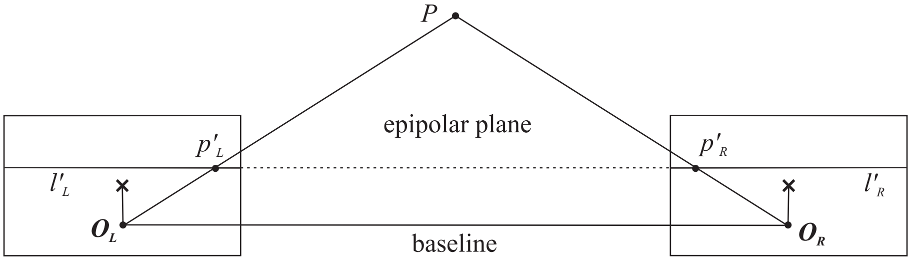


Figure 2.7: Canonical stereo configuration.

be mutually aligned and their horizontal axes parallel with baseline. The epipolar lines are then parallel to image rows. It is also useful to set horizontal and vertical effective focal lengths f_x, f_y of both new views to same values to obtain the pixel grids have same spacing in both directions.

2.3.3 3D point reconstruction

When the pair of mutually corresponding points is found, the original spatial point may be reconstructed. A necessary precondition is that both cameras are calibrated. The computation of point position in space can be done by geometrical or algebraical approach. Several major methods are presented in following text.

A basic geometrical solution computes the rays $\overrightarrow{O_L p_L}$ and $\overrightarrow{O_R p_R}$ using the results of calibration. An intersection of these rays is desired 3D point. However, the imprecisions in the calibration and stereo correspondence process may cause that rays miss each other in a space. This problem can be overcome by constructing a line segment perpendicular to both rays. Its middle point is declared as original 3D point then. Other way of solving non-intersecting rays is finding image points $\hat{p_L}$, $\hat{p_R}$ close to observed p_L, p_R which have intersecting rays. The deviation between these pair of points should be as small as possible. A mathematical expression of deviation is shown in Equation 2.7.

$$C(p_L, p_R) = |p_L - \hat{p_L}|^2 + |p_R - \hat{p_R}|^2 \tag{2.7}$$

The minimisation of function $C(p_L, p_R)$ can be done iteratively via non-linear least-squares methods. The resulting pair $\hat{p_L}$, $\hat{p_R}$ determines desired spatial point.

The algebraic methods are based on the general formula defining conversion between 3D point and its image (see Equation 2.6). The projection matrices H_L and H_R for left and right camera are known, projections p_L and p_R of point P as well. Two linear equations with $\tilde{P_W}$ as a variable are obtained by exploiting the scale factor w in Equation 2.6. Equation 2.8 shows calculations for left camera (notation - $\bar{h_{L1}}$ is first row of matrix H_L).

$$\begin{gathered} \tilde{p_L} = H_L \tilde{P_W} \quad \rightarrow \quad \begin{bmatrix} wU_L \\ wV_L \\ w \end{bmatrix} = \begin{bmatrix} \bar{h}_{L1} \\ \bar{h}_{L2} \\ \bar{h}_{L3} \end{bmatrix} \cdot \begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} \\ 0 = \bar{h_{L1}} \tilde{P_W} - U_L \bar{h_{L3}} \tilde{P_W} = (\bar{h_{L1}} - U_L \bar{h_{L3}}) \tilde{P_W} \\ 0 = \bar{h_{L2}} \tilde{P_W} - V_L \bar{h_{L3}} \tilde{P_W} = (\bar{h_{L2}} - V_L \bar{h_{L3}}) \tilde{P_W} \end{gathered} \tag{2.8}$$

Two similar equations are provided by right view as well. The composition of these 4 linear equations gives an overdetermined linear system in Equation 2.9, which can be solved by least-squares method.

$$0 = M\tilde{P}_W = \begin{bmatrix} \bar{h}_{L1} - U_L\bar{h}_{L3} \\ \bar{h}_{L2} - V_L\bar{h}_{L3} \\ \bar{h}_{R1} - U_R\bar{h}_{R3} \\ \bar{h}_{R2} - V_R\bar{h}_{R3} \end{bmatrix} \cdot \tilde{P}_W \qquad (2.9)$$

The detailed explanation of variants of presented methods and other techniques is available in [13].

2.4 Stereo correspondence

This section is dedicated to the most difficult task in the chain of creating 3D model from two views. Stereo correspondence algorithms try to find a matching between pixels in captured images which are projections of same spatial point. One of the images is usually selected as a *reference image*. The corresponding pixels to the pixels in the reference image are searched in a *matching image*. Basic idea behind the correspondence search is the fact that a 3D point is projected on the images together with its surroundings. The surroundings should look similar in both views. Thus, the pixels in different images which have very similar neighbourhoods are probably projections of same 3D point. According to this assumption the stereo correspondence methods compare the region around a pixel in the reference image with the matching image. From the theoretical point of view the neighbourhood around every pixel in the matching image should be checked for one pixel in reference image. However, this 'brute-force' approach is hardly feasible even for small pictures. Several constraints, which are discussed later, are exploited in the practice and can significantly reduce the complexity of this task. Despite these facts a stereo correspondence problem is computationally expensive and time-consuming.

Many difficulties for correspondence algorithm are caused by the properties of observed scene. It is problematic to find correct matching for the object which has uniform appearance in the image. The lack of texture on the object brings high number of pixels with similar neighbourhood and it is difficult to distinguish which one corresponds to the examined pixel. The view-depended appearance of same surface patch complicates correct matching as well because the matching relies on the similarity of patch projections. Different appearance of patch is often caused by its orientation in a space. When it is more oriented towards one of the cameras, its projection covers larger area in the image of that camera than in the image of another camera. The source of view-dependent appearance is also the structure of surface. The patch with fine spatial details such as hair looks differently from every viewpoint. The same problem can be caused also by illumination effects on the object. The reflections or shadows change positions when the observer is moving. It is impossible to find correct matching for the parts of scene which are occluded in one of the views or both of them.

2.4.1 Commonly used constraints

Several constraints are exploited in the context of matching stereo images [34, 11]. The epipolar constraint mentioned in Section 2.3.1 is essential. It significantly reduces a search space during the matching input images. The arbitrary pixel in the reference image has its corresponding pixel situated on one epipolar line in the matching image.

Another simplifications for stereo correspondence search are brought by the rectification of image pair. The epipolar line corresponding to the pixel in the reference image simply lies on the same row in the matching image. The alignment of epipolar lines with rows brings easy pass along them during the correspondence search. The relationship between matching points is defined by simple mathematical formula as well (Equation 2.10).

An *ordering constraint* assumes same ordering of matching pixels in both images. The pairs of matching pixels are situated on epipolar lines determined by particular epipolar plane. The order of pixels along epipolar line in the reference image is same as the order of their correspondences in the matching image. This assumption is correct when no small object is placed in front of another object in a scene.

A *neighbourhood consistency constraint* is suitable for continuous surfaces without rapid changes. It assumes than neighbouring pixels in the picture are projections of adjacent spatial points. One application of this constraint is the estimation of correct matching for a pixel from existing matching in its neighbourhood in ambiguous situations. Another use is an assumption that arbitrary region in the image is a projection of compact surface patch.

During the correspondence search are standardly compared two regions with same shape and size from different images. A *window similarity constraint* is implicitly used in such comparison (the regions have traditionally window shape). It means that both regions are the projection of same surface patch. This is completely true only when the patch is parallel with the baseline. If the patch is oriented more towards one of the cameras, its projections in the images have different shape and size.

2.4.2 Concept of disparity

Many techniques have been developed to solve the stereo correspondence problem. In spite of differences between them they work with the quantity of disparity. The disparity is a measurement of difference between stereo pair of images. It is usual precondition that the stereo image pair is rectified because the disparity can be easily expressed for each pixel afterwards. The image from left viewpoint in the canonical configuration is selected as the reference image in this work. The disparity d between two corresponding points p'_L, p'_R in the reference and matching image is described by Equation 2.10.

$$d = U'_L - U'_R \quad V'_L = V'_R \tag{2.10}$$

It is a difference between horizontal pixel coordinates of these points. It is important to note that each coordinates are connected to the coordinate system of their own rectified image. The formula is valid in this form only when principal point coordinates are same in both rectified images.

Because the disparity describes the relationship between two corresponding pixels which define 3D point, the connection between disparity and depth exists. The disparity can be considered as inverse depth with respect to the baseline. It is obvious from geometrical reconstruction of 3D point in the canonical stereo configuration as it is shown in Figure 2.7. When the disparity d is bigger, the rays through image points p'_L and p'_R cross each other closer to the common image plane and vice versa. The mathematical expression of the depth of 3D point from the baseline in the direction perpendicular to the common image plane is in Equation 2.11. The value of disparity d_N between p'_L and p'_R is expressed in the normalised image coordinates to convert from pixels to spatial units. The length of baseline is denoted B.

$$Z_C = B/d_N \tag{2.11}$$

The spatial position of point P is computed by Equation 2.12. It defines the transformation of normalised image coordinate vector $\tilde{P_N}$ to spatial coordinate vector P_C in left camera coordinate system. The vector $\tilde{P_N}$ represents the position of image point p'_L in a space (image plane can be situated in a space).

$$P_C = (B/d_N)\tilde{P_N} \quad \rightarrow \quad \begin{bmatrix} X_C \\ Y_C \\ Z_C \end{bmatrix} = (B/d_N) \cdot \begin{bmatrix} X_N \\ Y_N \\ 1 \end{bmatrix} \tag{2.12}$$

The vector P_C is then transformed to the world coordinate system according to Equation 2.2.

Every pixel in the reference image has a set of candidates for matching pixel along corresponding epipolar line in the matching image. The relationship with any of them can be expressed in terms of disparity (Equation 2.10). Therefore, it is possible to define a *disparity space* [27]. The disparity space is three-dimensional system with axes U, V, D. The axes U and V are aligned with rows and columns of the reference image. The axis D represents the disparity and is perpendicular to U and V. The point $[U'_L, V'_L, d]$ in the disparity space represents one pair of pixels - $[U'_L, V'_L]$ in the reference image and $[U'_R, V'_R]$ in the matching image as it shown in Figure 2.8. The disparity space has a discrete nature for purposes of this work. Therefore, the integer coordinates $[U'_L, V'_L]$ address a pixel in the reference image and the disparity d is measured in whole pixels as well. The disparity space is infinite from the definition but its used volume is limited by the boundaries of the images in practice. The work range on the axes U and V is set by the resolution of reference image. The range of possible disparity values can be derived from widths of images. The points on same disparity level in the disparity space define the set of 3D points which lie on same plane in 3D space. The reason is that the depth depends only on the disparity, not on the position of pixel pair (see Equation 2.11). The range of values on discrete disparity axis represents the sequence of parallel depth planes. A distance between them is increasing with decreasing disparity (but uniform spacing between disparity values). According to these facts the image of disparity space in 3D space is truncated pyramid with the reference image as smaller base.

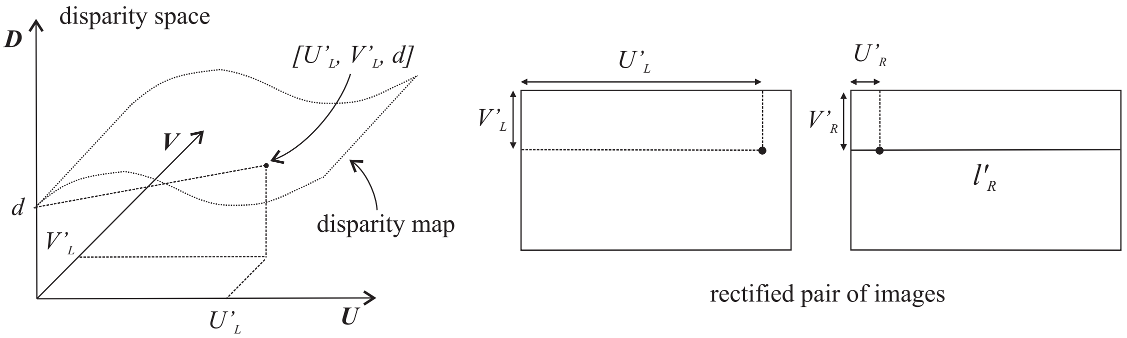


Figure 2.8: The relationship between rectified pair of images and disparity space. The disparity map is also illustrated.

A *disparity space image* (DSI) is a function over the disparity space. It defines a measure of confidence that a pair of pixels is corresponding to each other. An assumption that the neighbourhoods of matching pixels are similar is exploited to define this measure.

The similarity of neighbourhoods can be expressed by various measures. According to properties of the chosen measure the value of DSI is seen as matching cost (lower value means higher probability that the pixels are corresponding) or matching score (higher value means higher probability that the pixels are corresponding). A *correlation function* is a bore through DSI for a pixel in the reference image describing its matching along epipolar line.

The matching between the reference and matching image is defined by a *disparity map* (DM). It is a 2D function which assigns to a pixel in the reference image particular disparity value, so it defines its corresponding pixel in the matching image. The DM can be visualised as a surface floating over the reference image inside the disparity space (see Figure 2.8). The construction of DM depends on the DSI. Except the fact that DM should pick for every pixel the disparity to the best possible correspondence in the matching image it has to often satisfy other criteria. These criteria describe required shape of DM. The main task of the stereo correspondence search is to find optimal DM with respect to the specified properties it should have.

2.4.3 Overview of techniques

Large number of different stereo correspondence techniques have been developed. But the majority of them consist of 4 main stages as it is stated in the comprehensive overview in [27]. However, the stages are not always clearly divided or present at all. In the first stage the matching cost is computed for every desired point in the disparity space. It is usually simple computation from the intensities of two examined pixels such as squared difference or absolute difference. The result is initial DSI. The second stage is dedicated to a cost aggregation. The matching costs within some region around a point in the DSI are weighted and summed by particular formula. The result is assigned to the central point then. The region can be two-dimensional in the plane of constant disparity or fully three-dimensional. In the case of 2D region it means that the correspondence of two points is evaluated according to a similarity between the areas around them in the images. After the aggregation step the DSI has final form. The task of third stage is computing DM. The DM is extracted from DSI. The complexity of this stage depends on the properties imposed on the DM. The refinement of DM is performed in the last stage. It may consist of few different processes. The disparity space is traditionally discrete, therefore the values in DM are integers. The precision can be increased by refining them on sub-pixel level. The occlusions are detected by checking the DM from left view with DM from right view. The holes created by occlusions can be filled by the estimates of disparity from adjacent known values.

The stereo correspondence techniques may be divided into two main groups - *local techniques* and *global techniques*. The local algorithms, also called window-based, determine a disparity for given pixel according to its correlation function in the DSI [34]. The core of local algorithms are a computation of matching cost and its aggregation. Initial cost between pixels can be expressed as absolute or squared difference of intensities or binary result of exact intensity match. An aggregation is usually made by summing or averaging matching costs within a window on one disparity level (e.g. SSD, SAD). The cost computation and aggregation in this form represent a comparison of two windows with same size and shape (one in the reference image, one in the matching image). Thus, it favours the window similarity constraint. The window correlation measures such as a normalised cross-correlation join together first two stages and directly determine final value for the processed point in a disparity space. The computation of DM is straightforward. The disparity value associated with the minimal matching cost is recorded to DM for every pixel.

The global algorithms try to find a DM by an iterative optimisation where whole DSI is processed simultaneously. The matching cost between pixels is determined similarly as in the local methods. But the aggregation step is often skipped. The core of algorithms is the construction of DM. The energy function which describes a quality of current DM is defined. It consists of data term and smoothness term. The data term aggregates the overall matching cost of DM. The smoothness term incorporates explicit smoothness assumptions about the shape of DM. It is usually restricted to the local scale to make the computation feasible. The energy function is iteratively minimised by the refinement of DM. Main difference between global algorithms lies in the choice of minimisation technique. The choice can be narrowed by the class of function to be minimised. Traditional approaches are a simulated annealing, probabilistic diffusion or continuation. More recent methods are for instance a belief propagation [19, 30], tree-reweighted message passing [19, 30] or *graph cuts* [24, 23, 16, 18, 19, 30]. A group of global methods based on a dynamic programming is also important [29]. They approach the minimisation of energy function differently. They search for optimal solution of disparity for a pair of corresponding epipolar lines. A slice of DM is solved as a minimum cost path through the matrix of matching costs between pixels on epipolar lines.

2.4.4 Normalised cross-correlation

A *normalised cross-correlation* n_{CC} is a representative of matching cost measures [11]. The similarity of two square windows with same dimensions placed around examined pixels can expressed as a combination of covariance and variances. The mathematical formulation of n_{CC} is shown in Equations 2.13, 2.14, 2.15 and 2.16 [29].

$$n_{CC} = \frac{cov}{\sqrt{var(I'_L)}\sqrt{var(I'_R)}} \tag{2.13}$$

$$cov = \frac{1}{N}\sum_{i=-w}^{w}\sum_{j=-w}^{w}(I'_L[U'_L+i,V'_L+j]-\bar{I}'_L[U'_L,V'_L])(I'_R[U'_R+i,V'_R+j]-\bar{I}'_R[U'_R,V'_R]) \tag{2.14}$$

$$var(I'_L) = \frac{1}{N}\sum_{i=-w}^{w}\sum_{j=-w}^{w}(I'_L[U'_L+i,V'_L+j]-\bar{I}'_L[U'_L,V'_L])^2 \tag{2.15}$$

$$var(I'_R) = \frac{1}{N}\sum_{i=-w}^{w}\sum_{j=-w}^{w}(I'_R[U'_R+i,V'_R+j]-\bar{I}'_R[U'_R,V'_R])^2 \tag{2.16}$$

The square window with area $N = (2w+1)\times(2w+1)$ is centered around the pixel $[U'_L, V'_L]$ in the reference image (equivalently for matching image). The function $I'_L[U'_L+i, V'_L+j]$ represents an intensity in the pixel $[U'_L+i, V'_L+j]$ of reference image and $\bar{I}'_L[U'_L, V'_L]$ is an average intensity over the matching window.

The range of n_{CC} is from -1 to 1. The positive values mean strong correlation between the intensity values in compared windows. The negative values mean that the intensity values are opposite with respect to the means in the windows. Therefore, bigger positive value reflects stronger similarity between the windows and higher confidence that the centres of windows are matching pixels. This is a difference from other measures like SSD or SAD where lower value represents stronger correlation. The n_{CC} is seen as matching score rather than cost.

The interesting property of n_{CC} is an independence on affine transformation between intensities in the windows. The difference between overall intensity level in the windows is suppressed by comparing only the variation of values with respect to average intensity in each window. It brings a robustness in the cases where the appearance of surface patch projected into the windows is different in each view.

2.4.5 Graph theory

The stereo correspondence problem can be solved by the methods based on the graph cuts. This section explains basic theoretical information necessary for understanding these methods. The terminology and notation is adopted from [22].

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{E}$. The edges connect the nodes and have the orientation - $(a, b) \in \mathcal{E}$ where $a, b \in \mathcal{V}$. Each edge has a capacity $w(a, b)$ which is nonnegative number. The capacity of edge (a, b) can differ from the capacity of reversed edge (b, a). The set $\mathcal{V}$ contains two special *terminal* nodes - a *source* s and a *sink* t. The rest of nodes are non-terminal nodes $\mathcal{V}_0$ ($\mathcal{V} = \{s, t\} \cup \mathcal{V}_0$). The edge linking non-terminal node with one of terminal nodes is called a *t-link*. The edge linking two non-terminal nodes is called a *n-link*.

A *s/t cut* (in following text only a cut) divides the nodes into two disjoint subsets $\mathcal{S}$ and $\mathcal{T}$ ($\mathcal{S} \cup \mathcal{T} = \mathcal{V} \wedge \mathcal{S} \cap \mathcal{T} = \emptyset$). The condition is that the terminal nodes are included into different subsets ($s \in \mathcal{S} \wedge t \in \mathcal{T}$). The cut severs the set of edges $\mathcal{E}_C$ which interconnect the nodes in the subset $\mathcal{S}$ with the nodes in the subset $\mathcal{T}$ ($(a, b) \in \mathcal{E}_C$ if $a \in \mathcal{S} \wedge b \in \mathcal{T}$). The orientation of edge is important because the reverse edge (b, a) does not belong to the set $\mathcal{E}_C$. A cost of cut is a sum of capacities of the edges in the set $\mathcal{E}_C$ which corresponds to the cut. A *minimum cut* is a cut with minimum cost among all possible cuts on the $\mathcal{G}$.

The minimum cut problem can be solved by the maximum flow computation. The maximum flow is informally the maximum amount of 'liquid' which can be passed through the edges of the graph from the source to the sink. The minimum cut - maximum flow theorem states that the maximum flow is equal to the cost of minimum cut. Therefore, the set of edges saturated by the maximum flow corresponds to the minimum cut.

Chapter 3

System design

The design of 3D face capture system is introduced in this chapter. At first, the problem is analysed and divided into individual tasks. Each major task is then explained in detail in separate section.

3.1 Problem analysis

Primary objective of 3D face capture is the acquisition of a precise face model. However, the face colour appearance is usually demanded as well. The systems which have ambition to complexly cover a problem record a face dynamics in addition.

The 3D face capture systems are based on same principles as general-purpose capture systems but they address a few specific needs connected with the specialisation on a human face. At first, a health of scanned person must be protected under all circumstances. This fact excludes number of techniques which use an equipment potentially dangerous for human (e.g. contact methods or methods using strong laser beam). Also the procedure should be reasonably comfortable. The time of capture has to be very short because it is impossible hold a head in completely stable position. In the case of the systems recording face dynamics a capture has to be a continuous process in the video rate. Whole procedure should be reasonably quick and reliable to avoid long waiting or repeated capture of a person.

The described requirements narrow a choice of suitable techniques. The techniques using the structured light or stereo photogrammetry are mostly applied for the 3D face capture (see Section 2.1). The stereo photogrammetry is used in both variants - active and passive. The structured light methods which project several different patterns during some time period cannot be used unless they track the projected patterns. The reason are small movements of head during capturing which would make the captured data inconsistent. The main problem for many active techniques (structured light or photogrammetry) is an inability to acquire a shape and an appearance of face simultaneously because of the projected pattern on the face. The solutions of this problem are miscellaneous. The simplest way is to capture a shape (with pattern) and an appearance (without pattern) separately but immediately one after another. It is demanding on a speed of the used cameras. The second variant is the reconstruction of appearance from the images with the projected pattern. However, this approach is computationally expensive and a result has lower quality than direct acquisition of appearance. The possibility is also a projection of pattern in the non-visible part of electromagnetic spectrum (e.g. IR light). The pattern necessary for shape reconstruction is captured by IR cameras and additional classic cameras provide the capture of appearance [34]. However, this implies complex capture rig with specialised equipment. Another limitation is a strict regulation of scene illumination to avoid a contamination of the pattern. The passive stereo photogrammetry has not this type of difficulties. The stereo matching is performed on the usual images of face without any projected pattern. Therefore, the capture rig is simple and does not contain any specialised appliances. Moreover, the constraints of an illumination in the capture volume are far less restrictive. The drawback is more difficult stereo correspondence search. It is problematic to achieve accurate and reliable matching especially in the case of human face. It is a consequence of uniform appearance of face in the images. A human skin has a weak texture which does not provide much information for the matching. Because of that active stereo photogrammetry techniques achieve higher accuracy of the reconstructed model. Thus, they are more used for the 3D face capture despite more expensive capture rig. However, a research challenge is to develop a capture system without active lighting components which has same accuracy and efficiency as the systems using pattern projection.

The stereo correspondence search, which is crucial task in the stereo photogrammetry, can be solved by large number of methods. They are considered as local or global according to an extent of data used for matching one pixel [27]. The local techniques [34] are fast but inaccurate for the materials with weak texture such as a human skin. The global techniques treat the construction of disparity map as an optimisation problem. Therefore, they achieve better quality of disparity map but computational demands are high. The techniques based on the graph cuts generally achieve better results than other global methods such as simulated annealing, probabilistic diffusion or dynamic programming [27]. They can be divided into two groups according to the way how the graph cut is used for the energy minimisation. The first group use the α expansion or $\alpha-\beta$ swap algorithm which iteratively applies the graph cut on the 2D grid graph to minimise an energy function [18, 31]. The energy function can define complex smoothness term but the algorithm does not quarantee finding a global minimum. However, the found local minimum is within a known factor of the global minimum. The techniques from the second group such as [23, 16] compute optimal disparity map corresponding to global energy minimum by single cut on the graph with 3D grid topology. But the energy function can contain only linear smoothness term. The size of graph is also larger than in the α expansion or $\alpha - \beta$ swap algorithm. To cope with high computational demands of the mentioned graph cut techniques, several approaches were proposed [25, 32]. They showed promising directions of the reduction of search space during matching the images. The global optimisation of disparity map is then feasible even for the high-resolution images.

The aim of this work is to develop the 3D face capture system based on the passive stereo photogrammetry. It follows a trend of lowering the requirements and constraints imposed on the capturing process. They limit the deployment of such systems in the real environment. The developed system processes the high-resolution images of a face. It brings a chance to explore whether such resolution records enough skin details for accurate matching. It would overcome traditional difficulties of passive stereo photogrammetry methods to produce high-quality 3D models. However, the processing high-resolution images increases the computational demands on the level unfeasible by standard computers. Therefore, the computational time and memory consumption were taken into account except the quality of 3D model during a development. The model quality and computational demands are mainly influenced by the correspondence search in the image pair. This work is focused on this most important part of 3D capture. Three different approaches are explored and compared in terms of the mentioned characteristics. The presented local technique represents computationally inexpensive approach. It is included to demonstrate what quality can be achieved with low computational demands. It also shows the way how basic local technique can be extended to be more reliable. The second technique is global optimisation of disparity map using graph cuts. The graph cuts were chosen because of their top results in a precision of disparity map. Specifically, the single cut on 3D grid graph is used because it leads to optimal solution. This is important for ambiguous disparity space image which created from face images. This method proves that high-quality stereo matching can be computationally very expensive for high-resolution images. The last technique merges previous two to demonstrate possible optimisation of global technique. The global approach is then reasonably demanding even for high image resolution.

The system captures the shape and appearance of face simultaneously in particular moment of time (no face dynamics is recorded). The capturing is done by classic digital still cameras and pattern projection is not involved. Thus, the process is quick, safe and comfortable for scanned person. The reconstruction is performed from one stereo image pair. A result is textured triangle mesh. The 3D model covers only a face with part of neck (hair are not included).

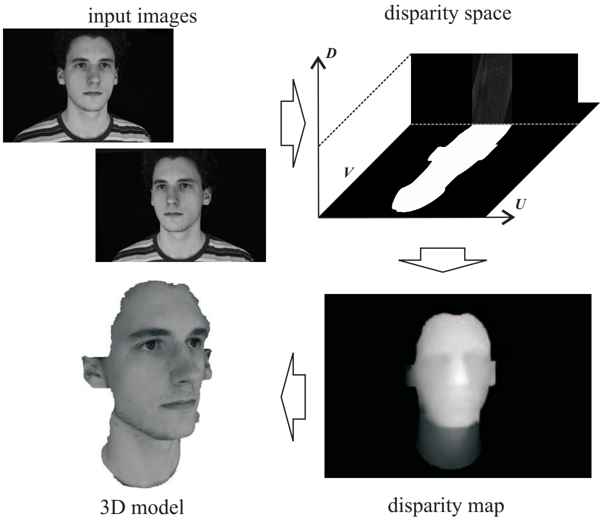


Figure 3.1: The scheme of the 3D face capture.

Figure 3.1 schematically illustrates a processing in the 3D capture. The processing pipeline has same stages regardless of captured object. However, the algorithms can exploit a priori knowledge that the application is a human face. The input images obtained by the capture rig which consists of two digital still cameras and the system of lights. The capture is done under uniform illumination of whole face. The calibration of cameras to set world coordinate system is necessary to describe the way how they capture scene. The world coordinate system is essential for 3D model reconstruction. According to the calibration information the stereo image pair is pre-processed for the correspondence search stage. The pre-processing consists of image undistortion, the rectification of image pair and face extraction. These procedures simplify the following stereo matching. The correspondence search is divided into two main phases - a construction of disparity space and a construction of disparity map. In the first phase the information about all possible variants of matching between the images are gathered inside disparity space. A matching score for a pair of pixels is a similarity between square-shaped windows centered around them in each image. The matching scores are recorded into disparity space image. The second phase extracts the matching between images from the disparity space image in a form of disparity map. The disparity map can be seen as a range image for the face region in the reference view. The construction of disparity map is done by the mentioned different methods. It can be enhanced by the post-processing. It involves the hole-filling, sub-pixel disparity refinement and Gaussian smoothing. Finally, the 3D model is reconstructed from the disparity map. The triangle mesh is created by the triangulation of computed 3D points. It is textured according to the reference image.

3.2 Face capturing

The face capture rig is not very complex when the system is based on passive stereo photogrammetry. The rig used for this work contains two digital still cameras and the system of lights.

The distances between individual cameras and captured person are important for the usability of stereo image pair in consequent processing. The pair of cameras needs to have short baseline to capture images similar enough to allow reliable matching. The risk of occlusions is reduced as well. On the other hand, the pair with longer baseline covers larger portion of face. Also the computation of 3D point from its projections in both images is more precise in the case of longer baseline. Thus, it is necessary to find a compromise in the practice. To improve the quality of matching, it is needed to record as much skin details as possible. But the distance between camera pair and captured person is limited by required capture volume. The capture volume created by camera fields of view has to contain whole head with neck. Therefore, the maximal resolution provided by cameras is exploited to capture detailed skin texture. The texture is present on very small scale, so the cameras have to be well focused on the surface of face to avoid a blur in the images. To allow correct construction of face model, the movement of face is not permitted between recording all pictures. Because of this fact both cameras are triggered simultaneously and expose in short time. The images are stored in RAW format to preserve as much information as possible and give a flexibility in later manipulation. Their resolution is 3504×2336 pixels and this resolution is used in following processing.

Although no active components for creating light effects on a face are used during capturing, a proper light setup is required. It is beneficial for stereo matching to acquire high-quality pictures without view-dependent illumination effects. More technical details about the capture rig and the image acquisition can be found in Sections 4.1, 4.2.

3.3 Camera calibration

The method used for camera calibration is partially inspired by [35]. This technique assumes planar calibration object captured by examined camera from more viewpoints in order to compute all intrinsic parameters involved in the camera model presented in Section 2.2.2. At the beginning the analytical solution is computed and it is refined by iterative optimisation afterwards. The extrinsic parameters are determined for all views. But only the set of extrinsic parameters connected to the principal view used during capturing becomes a part of final camera projection matrix.

The reference points are detected in all images of calibration object. Every view provides several pairs of spatial and image coordinates connected to individual reference points. The overdetermined system of linear equations is assembled from these coordinate pairs in the analytical method. The solution of this system is the projection matrix of particular view. The lens distortion coefficients are not included inside the projection matrix because they describe non-linear transformation. Each obtained projection matrix sets constraints on the camera intrinsic matrix which is identical for all views. These constraints can be expressed in the form of linear equations and composed in the system of equations. The extraction of individual intrinsic parameters from the solution of this system is explained in [8] (differs from [35]). After determining camera intrinsic matrix the rotation matrices and translation vectors for all views can be obtained.

The described analytical approach is fast but its disadvantages are an absence of lens distortion coefficients and a noise sensitivity. Because of that the analytical method is followed by iterative optimisation of its solution. The optimisation includes also a computation of radial and tangential distortion coefficients because it is not important for optimisation procedure whether an operation is non-linear.

The quality of camera model solution is evaluated by a *total re-projection error* in the optimisation method. This error sums deviations between the observed positions of all reference points and their positions computed through current camera model parameters. The formal expression is shown in Equation 3.1, where m is number of processed images and n is number of reference points on particular image.

$$E = \sum_{j=1}^{m} \sum_{i=1}^{n} \left| p_{ij} - p'_{ij}(f_x, f_y, \alpha, u_0, v_0, k_1, k_2, k_3, k_4, k_5, R^C_{Wj}, t^C_{Wj}) \right|^2 \tag{3.1}$$

p_{ij} denotes observed coordinates of reference point i on the image j and p'_{ij} represents projected coordinates through intrinsic and extrinsic parameters connected to image j. The matrix R^C_{Wj} is defined by 3 rotation angles in calculations.

The optimisation procedure tries to refine parameters in the way which leads to the minimisation of the total re-projection error. The error is determined after every iteration in order to set new direction of refinement. The process stops when the change of error between two following iterations falls under certain limit. Correct convergence of iterative process is usually guaranteed by exploiting analytical solution as an initial estimate. Initial values of distortion coefficients are set to zero. This simplification is usually sufficient because modern compound lenses do not deform much captured images. A *Levenberg-Marquardt method* is used for optimisation due to fast convergence and relatively small computational demands [11].

3.4 Image pre-processing

The stereo image pair is processed in several stages after capturing. The main aim of this pre-processing is the reduction of computational complexity of following correspondence search and surface reconstruction. Firstly, it is convenient to eliminate the effect of camera lens distortion in both pictures. Secondly, the image pair is transformed to form as it was captured in a canonical stereo configuration. Finally, face areas are extracted from the images.

3.4.1 Image undistortion

A lens distortion complicates the mathematical expression of perspective projection of a 3D point on the image plane (see Section 2.2.2). The transformation is non-linear, thus the relationship between spatial point and image pixel cannot be expressed by the projection matrix. The principles followed from the binocular view geometry cannot be applied as well. Because of these facts, it is suitable to transform the captured images to the form as they were taken by the camera without lens distortion.

A new virtual camera is described by same intrinsic and extrinsic parameters such as real calibrated camera but the radial and tangential distortion coefficients are zero. The correct colour of the pixel p on the image captured by virtual camera is obtained by mapping this pixel on its distorted variant p_D. The dimensions of original and new image are same. Every pixel p on new image can be converted to the normalised image coordinates P_N through inverse camera intrinsic matrix A^{-1}. The P_N represents the raw perspective projection of spatial point P before the transformation by the properties of optical system. Thus, it is same for real and virtual camera. At last, P_N is transformed through the optical system with lens distortion to acquire image point p_D (see Equations 2.4 and 2.5). Whole mapping is formulated in Equation 3.2 .

$$\tilde{P_N} = A^{-1}\tilde{p} \qquad \tilde{P_N} \rightarrow \tilde{P_D}(Eq.\ 2.4) \qquad \tilde{p_D} = A\tilde{P_D} \tag{3.2}$$

The pixel p is mapped to general image point p_D among existing pixels in the distorted image (assume that a colour is the value of sample in the centre of pixel area). Because of this fact the new colour value for p is computed by bilinear interpolation of colours from the 4 closest pixels to p_D. If the pixel p is mapped outside the borders of original image, it is set to background colour. After this transformation the relationship between image point and its 3D point is completely described by the projection matrix (see Equation 2.6).

3.4.2 Rectification of image pair

The aim of the stereo image pair rectification is to simplify correspondence search as it is explained in Section 2.4.1. The general information about rectification are summarised in Section 2.3.2. The precondition of this process is an existence of the projection matrices which fully describe capturing original images. Therefore, the lens distortion elimination has to precede. The result are new projection matrices for both cameras as they were in the canonical stereo configuration. The transformation between original images and their equivalents on new common image plane is determined as well.

The technique presented in [12] is adapted for this work. The new common image plane is parallel to the baseline. Its normal aims to the examined face and lies in the plane defined by the baseline and old Z_{CL} axis of left camera coordinate system. It is a difference from general approach mentioned in Section 2.3.2 but it is suitable for the stereo pair aiming on same object in a space. The orientation of new image plane is described by the rotation matrix R'^C_W. This matrix is a part of rectified extrinsic parameters of both cameras. But the optical centres of cameras remain on their original positions. The rows of R'^C_W are the unit vectors of new camera coordinate system. Thus, the construction of R'^C_W requires to compute directions of new axes in the world coordinate system (see Equation 3.3).

$$R'^C_W = \begin{bmatrix} \bar{r}'_1 \\ \bar{r}'_2 \\ \bar{r}'_3 \end{bmatrix} \qquad \rightarrow \qquad \begin{aligned} \bar{r}'^T_1 &= (t^W_{CR} - t^W_{CL})/|t^W_{CR} - t^W_{CL}| \\ \bar{r}'^T_2 &= \bar{r}^T_{3L} \times \bar{r}'^T_1 \\ \bar{r}'^T_3 &= \bar{r}'^T_1 \times \bar{r}'^T_2 \end{aligned} \tag{3.3}$$

The vector t^W_{CL} is a position of the optical centre of left camera in the world coordinate system (equivalently t^W_{CR}). The vector $\bar{r}^T_{3L}$ is last row in old rotation matrix of the left camera which describes Z_{CL} axis.

The rectified camera pair should have same description of optical system because the pixel grids for the left and right rectified camera are expected to be aligned and have same spacing between pixels. New common intrinsic matrix A' is established by averaging old camera intrinsic matrices. The skew coefficient is eliminated by setting to zero value. The position of principal point is not considered at this point of algorithm, so both coordinates u'_0, v'_0 are zero. New rectified projection matrices are computed according to Equation 3.4.

$$H'_L = A'[R'^C_W| - R'^C_W t^W_{CL}] \quad H'_R = A'[R'^C_W| - R'^C_W t^W_{CR}] \tag{3.4}$$

At last, the transformation between the image point p_L on the original image and the image point p'_L on the rectified image is found (equivalently for the right camera). It basically describes the perspective projection of original image to new image plane. It is derived from the original projection matrix H_L and the rectified projection matrix H'_L. The transformation is defined by the matrix Q_L which is inferred in Equation 3.5.

$$\begin{gathered}
\begin{array}{lcl}
\tilde{p}_L = H_L \tilde{P}_W & & P_W = t^W_{CL} + (R^C_{WL})^{-1} A_L^{-1} \tilde{p}_L \\
& \rightarrow & \\
\tilde{p}'_L = H'_L \tilde{P}_W & & P_W = t^W_{CL} + (R'^C_W)^{-1} A'^{-1} \tilde{p}'_L
\end{array} \\
\tilde{p}'_L = A' R'^C_W (A_L R^C_{WL})^{-1} \tilde{p}_L = Q_L \tilde{p}_L
\end{gathered} \tag{3.5}$$

Resulting transformation matrices for left and right camera are in Equation 3.6.

$$\tilde{p}'_L = Q_L \tilde{p}_L \quad \tilde{p}'_R = Q_R \tilde{p}_R \tag{3.6}$$

The H'_L and Q_L computed by original technique provide the transformations to the left rectified image which has the origin of image coordinate system in its principal point ($u'_0 = 0, v'_0 = 0$). However, in the practice the rectified image is stored in the pixel array with the origin of image coordinate system in the upper-left corner. If $\tilde{p}'_L$ will be defined with respect to this origin, the matrices Q_L and H'_L have to be modified. At first, it is necessary to detect the area on new image plane covered by projected original image to set the boundary of rectified image. It is achieved by the projection of four corners of the original image. The rectangular boundary of rectified image should be closely placed around projected polygon to store only useful information. But the bounding rectangles for left and right rectified images have to have same height and be in same position in vertical direction on common image plane. The reason is obtaining same row coordinate for a pixel and its corresponding epipolar line in the rectified pair of images. Additional transformation sets new position of principal point with respect to the upper-left corner of chosen bounding rectangle. It is different for each view and is expressed by the matrices K_L and K_R. Other matrices are changed by them in Equation 3.7.

$$\begin{array}{lcl}
K_L = \begin{bmatrix} 1 & 0 & u'_{0L} \\ 0 & 1 & v'_0 \\ 0 & 0 & 1 \end{bmatrix} & \rightarrow & \begin{array}{l} H'_L := K_L H'_L \\ Q_L := K_L Q_L \end{array} \\
K_R = \begin{bmatrix} 1 & 0 & u'_{0R} \\ 0 & 1 & v'_0 \\ 0 & 0 & 1 \end{bmatrix} & \rightarrow & \begin{array}{l} H'_R := K_R H'_R \\ Q_R := K_R Q_R \end{array}
\end{array} \tag{3.7}$$

Because the position of principal point is different in each image, the computation of disparity presented in Equation 2.10 has to be slightly modified to Equation 3.8.

$$d = U'_L - U'_R + (u'_{0R} - u'_{0L}) \quad V'_L = V'_R \tag{3.8}$$

Actual image conversion works similarly to the undistortion procedure. Each pixel p'_L in the rectified image is mapped on the image point p_L in the original image through Q_L^{-1}. Correct colour is computed by bilinear interpolation or set to background colour if the mapping leads outside of image. The example of rectification is shown in Figure 3.2.

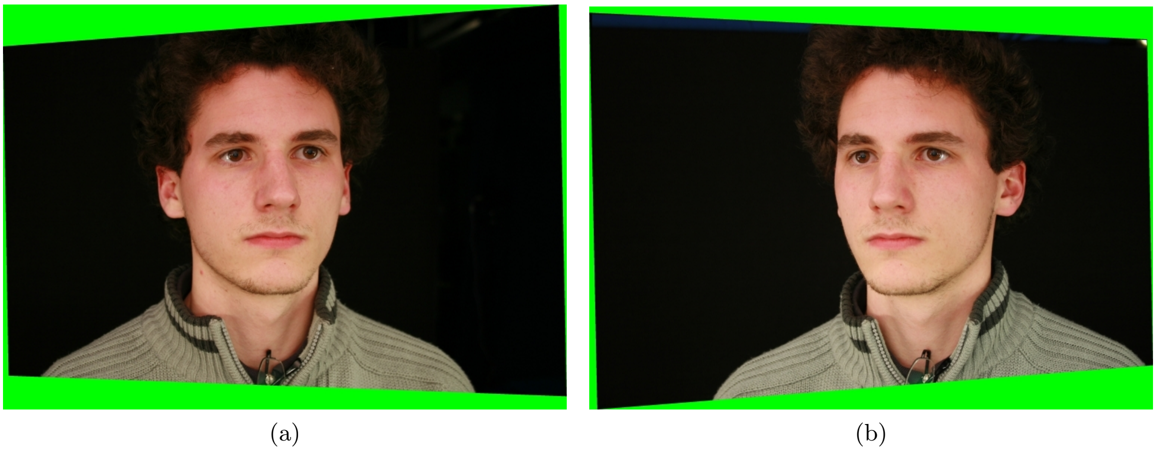


Figure 3.2: The example of rectified stereo image pair - left view (a), right view (b). The background is marked by green colour.

3.4.3 Face extraction

Although both cameras are pointed and zoomed on the head of person captured, a face covers smaller portion of whole area of an image. Especially after rectification when the dimensions of image are increased. It is profitable for following processing to extract a face area in both rectified images. The correspondence search can be then restricted only these regions.

The extraction is done by segmenting image according to a sample of skin. The representatives of skin colour are collected from small window in the face area of image segmented. Every component of colour is normalised by an intensity to gain a robustness to illumination changes (e.g. red component $r = R/I$). General skin colour is modelled separately in each normalised colour component. The skin distribution in every component is described by Gaussian function. The mean μ and variance σ^2 of function are computed from sample colours. During segmentation every pixel is classified as skin or non-skin according to the relationship of its own colour to these functions. A similarity of pixel colour to general skin colour for red channel is expressed in Equation 3.9 where *tol* defines some portion of variance.

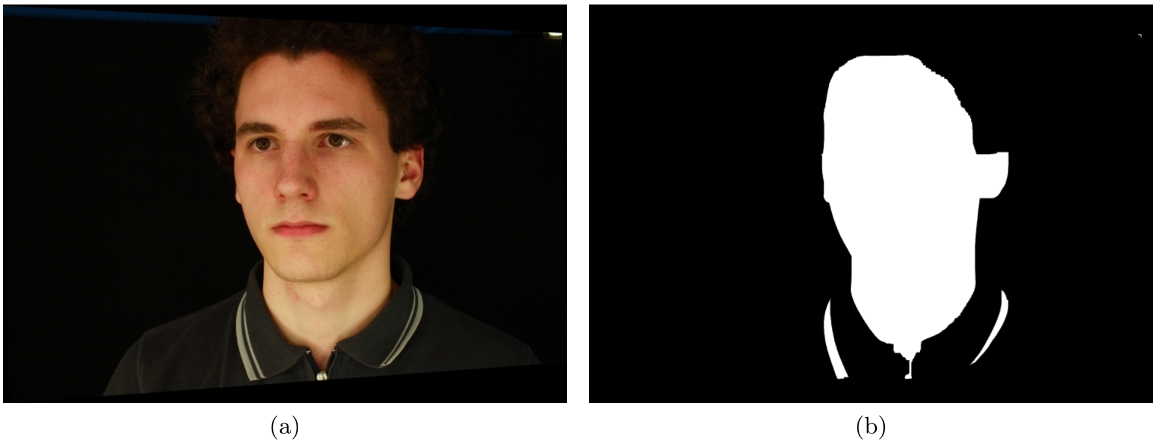
$$s_r = \frac{(r - \mu_r)^2}{\sigma_r^2 tol} \tag{3.9}$$

When the condition in Equation 3.10 is correct, the pixel is marked as a skin. The result of segmentation is binary mask with same dimensions like original image.

$$\frac{1}{3}s_r + \frac{1}{3}s_b + \frac{1}{3}s_g < 1 \tag{3.10}$$

The classifier is simple, its sensitivity can be adjusted only by coefficient *tol*. It also assumes that normalised colour components are independent which is not completely correct. However, it achieves reasonable results.

Small areas of skin which are not correctly classified during the segmentation always exist. Furthermore, non-skin face parts (e.g. eyes, eyebrows, lips) are also excluded from binary mask. But they all belong to a face (a hair are not considered as a part of face). To obtain compact face region, binary mask produced by the segmentation is modified by morphological operators [21]. At first, binary dilation operator is applied to the mask to fill the holes in the face region. Afterwards, binary erosion operator with same size is applied to return back the outer boundary of face region. The example of face extraction is shown in Figure 3.3.



(a)

(b)

Figure 3.3: The example of face extraction - original image (a), extracted face region (b).

3.5 Construction of disparity space

After the image pre-processing the actual stereo correspondence search can be started. The first phase of the process is the construction of disparity space. The concept of disparity space and related terms are explained in Section 2.4.2. The foundation for the disparity space is the pair of rectified face images. The image captured from the left viewpoint is chosen as the reference image and the image captured from the right viewpoint is the matching image. The information about all possible variants of matching between the images are gathered inside disparity space. The amount of variants is reduced by the epipolar constraint.

The precision of disparity space is limited by minimal unit on each axis - 1 pixel. It is necessary for exact mapping - pixel to pixel between the reference and matching image by Equation 3.8. The resolution of reference image sets an extent of disparity space along axes U and V. The maximal range of disparity derived from widths of images is needlessly large because the corresponding depth range contains large volume of empty space in front of the head and behind it. The minimal disparity range needed for correct reconstruction of whole face is limited by the depth of the closest point on the face to the camera pair and the farthest point from it. The nearest point specifies the highest possible disparity and vice versa. The small margin is added to minimal range to ensure an accuracy of matching on the borders. Despite the reduction on axis D the disparity space can be huge in the case of high-resolution images. Therefore, it is necessary to reduce the number of points in the disparity space which are actually examined during the correspondence search. A large portion of disparity space is excluded by face regions extracted from both rectified images. Only the pixels inside face region in the reference image are relevant for stereo matching. The general disparity range is shrunk individually for each pixel according to the face region in the matching image because every potentially corresponding pixel have to be inside the region. The last reduction is a sampling of reference image by uniform *scanning step* s_s along rows and columns. The correspondence is searched only for this grid of pixels.

3.5.1 Computation of disparity space image

The disparity space image DSI is defined only for the parts of disparity space which were not excluded in the described reduction. The slice through the DSI along row V'_L is depicted in Figure 3.4. It can be seen that DSI is limited by the face region in the reference image (white colour). The value in the point $DSI(U'_L, V'_L, d)$ describes a similarity between the window centered around pixel $[U'_L, V'_L]$ in the reference image and the window centered around pixel $[U'_R, V'_R]$ in the matching image. The relationship between these pixels through disparity d is expressed by Equation 3.8. The matching windows have square shape and same fixed size $2w + 1$. The comparison of windows with same shape implicitly assumes the ordering and window similarity constraint (see Section 2.4.1).

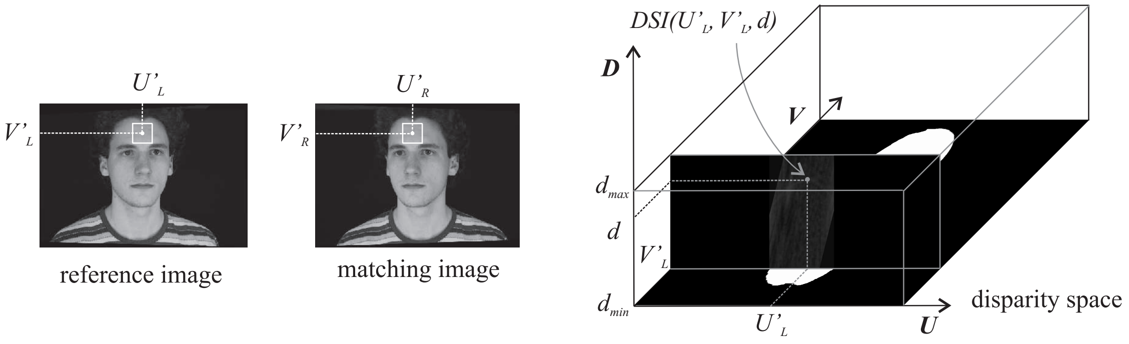


Figure 3.4: The relationship between compared windows and the disparity space image.

The normalised cross-correlation is employed as a measure of window similarity (see Section 2.4.4). The intensity only is used for the comparison. Although, the emphasis is put on similar appearance of face in both views during the capturing, small differences between same surface areas are present in the images. They are caused by slightly different camera gain or weak reflections on a skin. The n_{CC} brings bigger robustness against this view-dependent appearance in comparison to other measures such as SSD or SAD. On the other hand, the computational time is longer.

The computation of DSI for a compact part of the disparity space can be significantly accelerated [28]. The windows around close pixels are overlapping, therefore many calculations for n_{CC} are repeated. At first, the formula for $n_C C$ has to be modified. Equations 2.14, 2.15 and 2.16 for a covariance and variances can rewritten to Equations 3.11, 3.12 and 3.13.

$$cov = \frac{1}{N}\left[\sum_{i=-w}^{w}\sum_{j=-w}^{w}\left(I'_L[U'_L+i, V'_L+j]I'_R[U'_R+i, V'_R+j]\right)\right] - \bar{I}'_L[U'_L, V'_L]\bar{I}'_R[U'_R, V'_R] \quad (3.11)$$

$$var(I'_L) = \frac{1}{N}\left[\sum_{i=-w}^{w}\sum_{j=-w}^{w} I'_L[U'_L+i, V'_L+j]^2\right] - \bar{I}'_L[U'_L, V'_L]^2 \tag{3.12}$$

$$var(I'_R) = \frac{1}{N}\left[\sum_{i=-w}^{w}\sum_{j=-w}^{w} I'_R[U'_R+i, V'_R+j]^2\right] - \bar{I}'_R[U'_R, V'_R]^2 \tag{3.13}$$

Now each formula contains several different aggregations over window which can be computed independently (average $\bar{I}'_L[U'_L, V'_L]$ is sum of intensities over the window as well).

The sum of some generic pixel value over the window for every pixel in the rectangular image can be done effectively by *box-filtering*. Simple demonstration is in Figure 3.5(left) where 3×3 window is used. The 1D row buffer is initialised by sum of 3 values in vertical direction for every horizontal position (grey frames). The column buffer is one value which is initialised by sum of first 3 entries in the row buffer. After initialisation the column buffer contains the sum of pixel values in first window on the row (grey frame). The value from column buffer is stored into another image with same dimensions. The aggregation over next window (black frame) is computed by subtracting starting value and adding new value from row buffer to the column buffer. This procedure is repeated along whole row. The row buffer is similarly updated by one substraction and addition for every entry when the processing continues on the next row (black frames). The total number of basic operations such as plus and minus is significantly smaller in the case of box-filtering in comparison to separate summing over the window for every pixel in the area.

Assume that n_{CC} has to be computed for every pixel inside rectangular area in the reference image for particular disparity level. The area with same size is situated in the matching image according to the disparity value. The sum of intensities and squared intensities is computed for every pixel in both areas by box-filtering. The $var(I'_L)$ and $var(I'_R)$ can be easily calculated for every pixel in both areas then. The *cov* has another term which sums the multiples between areas which can be pre-computed by box-filtering as well. Further speedup is gained when the computation is extended to continuous disparity range. The pre-computation for the area in the reference image is necessary to perform only once. But the area in the matching image is shifting along rows with changing disparity value (see Figure 3.5(right)). Thus, the array of windows for summing is different for every disparity level. However, these arrays are almost overlapping in case of 1-pixel disparity change as it is depicted in Figure 3.5(right). It is effective to pre-compute the sums of intensities and squared intensities by the box-filtering for every pixel in extended area in the matching image. This extended area contains every instance of original area shifted according to the disparity range. When the step on next disparity level is made, the data for shifted area are only retrieved and $var(I'_R)$ is directly calculated. The term with multiplication of areas in the *cov* cannot be pre-computed across disparity range, the box-filtering is exploited only within each disparity level. The calculation of final n_{CC} value for arbitrary pixel pair is a matter of few basic operations over pre-computed values. The computational time for whole DSI is almost invariant with respect to the matching window size.

The Figure 3.5(right) illustrates a situation when the potentially corresponding pixels are mapped by d_{max} outside the matching image or the part of windows around them is mapped by $d_{max}-1$ outside. In both cases the n_{CC} is not computed for these pairs of pixels and the DSI is not defined in this point. The reduction of disparity space by scanning step has an influence on the box-filtering where the update of buffer aggregates s_s elementary

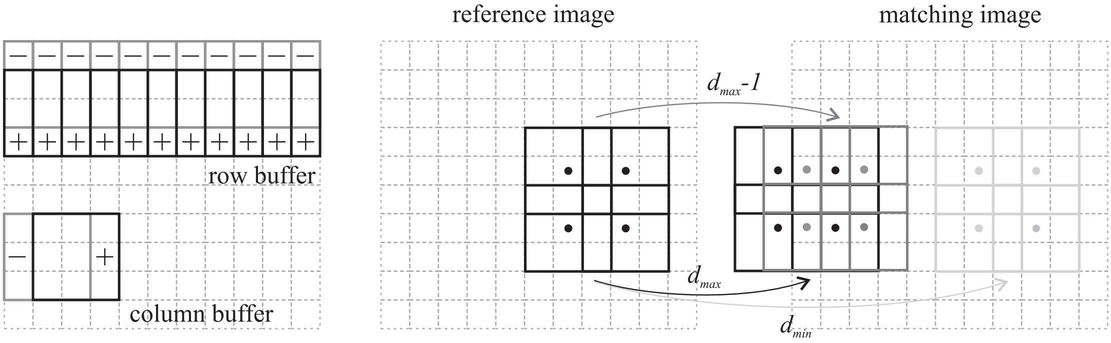


Figure 3.5: The box-filtering for computing DSI. Basic idea of box-filtering with two buffers (left). Mapping small pixel grid in the reference image to the matching image through disparity range where $s_s = 2$ (right).

steps. The face regions are not involved in the box-filtering because it would complicate and slow down the pre-computation. The check of examined pixel pair against the regions is done before final computation of n_{CC}.

3.5.2 Quality of disparity space image

A face captured in the stereoscopic image pair has a response inside the DSI. It has a form of surface marked by high correlation values. Because a disparity can be seen as an inverse depth, the surface in the disparity space is skewed version of the real face surface. The illustration of DSI is shown in Figure 3.6. It is a slice through disparity space along row across the area under eyes. The row coordinates are increasing from left to right. The disparity is increasing from top to bottom. The matching score is plotted in grey-scale values (brighter colour means higher n_{CC}). Black areas is a part of disparity space excluded by face region in the reference image where DSI is not defined. The DSI is not defined in red areas as well because the disparity range is limited by the face region in the matching image. The searched surface is represented by the brightest, thin ridge going from the left to right. Its position is ambiguous in some areas because the course of DSI is very rugged. General articulation of DSI is caused by the fact that a human skin doe not provide strong texture. For instance the object with noise texture has strong, visible ridge in comparison to the rest of DSI. The DSI is more rugged for the face areas which are have slightly view-dependent appearance (e.g. eyebrows). Finally, the parts of face which are more oriented to one of the views or completely occluded have an ambiguous response in the DSI as well. The reason is a violation of window similarity constraint. It is visible in Figure 3.6 where the ridge is not recognisable in the central part (a nose) and near the borders (sides of face). The nose and sides of face are slanted with respect to the camera pair, therefore the window comparison gives not completely correct results. The error is increasing with the window size and the slope of surface.

The parameter which influences the course of DSI is the size of matching window. The examined neighbourhood of pixel has to have certain size to provide enough information for unambiguous matching. This is dependent on the amount of texture detail available around pixel but bigger window brings more information. On the other hand, the window size should be small as possible to reduce a incorrectness for slanted surfaces. The DSI becomes smoother with increasing size of window. The response of face is stronger and

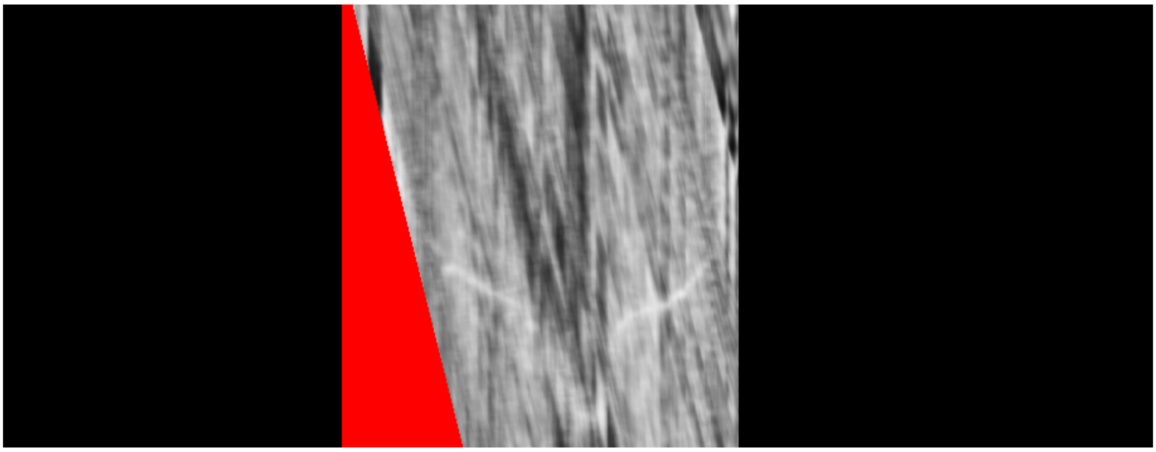


Figure 3.6: The slice through the disparity space with visualised disparity space image.

more recognisable but it looses details. However, ambiguous parts are not much improved in comparison to better ones. The experiments showed that the windows with size under 21 pixels produce weak response of face. The most rapid improvement is in the range from 21 to 31 pixels. The windows bigger than 31 pixels bring only small changes.

3.6 Construction of disparity map

The most complex task in a chain of 3D face capture is an extraction of disparity map from the DSI. The goal is to find the DM which precisely describes the surface present in the DSI. It is done by assigning one disparity value to each desired pixel in the reference image (see Equation 3.14). To reflect the reduction of a disparity space, the DM is determined only for the pixels sampled by scanning step within face region in the reference image (DSI is defined only for these pixels as well). It is important to mention that the adjacency between pixels in context of DM is understood with respect to the scanning step in the following text. The neighbouring pixels in the DM are distant s_s pixels in full resolution of the reference image. The size of any region in the DM or operator applied on the DM is measured in the sampled pixels so it covers larger area in the full resolution.

$$DM(U'_L, V'_L) = d \tag{3.14}$$

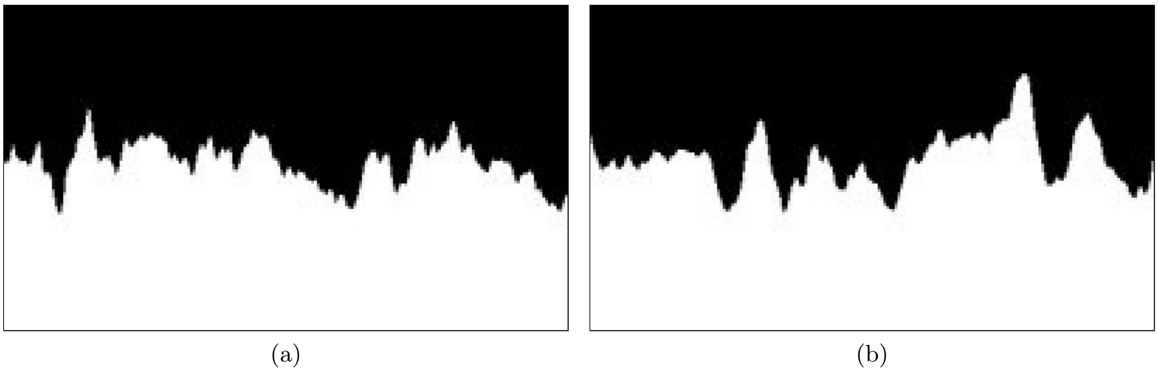
A priori knowledge of reconstructed object gives beneficial constraints on the searched DM. A human face has a compact, continuous, and smooth surface. It conforms the ordering constraint. The majority of surface is fronto-parallel to the baseline because the face is captured from front. Thus, it is suitable for the window similarity constraint. Moreover, the neighbourhood consistency constraint can be exploited for the extraction of DM. It assumes that neighbouring pixels in a picture are projections of adjacent spatial points so their disparity values should be similar.

Three different techniques are presented for the construction of disparity map. First technique is a representative of local methods for stereo correspondence search. Second technique is a representative of global methods based on the graph cuts. Third technique is a hybrid between first two techniques which combines their advantages. All three methods work with same DSI constructed by the presented procedure. The disparity map can be post-processed by various methods after the construction.

3.6.1 Local approach

Traditional local methods examine individually the correlation function in the DSI for each desired pixel in the reference image. The choice of final disparity for the pixel is straightforward, the disparity value of global maximum in the correlation function is recorded into the DM (recall that the n_{CC} is a matching score). However, this method leads to the noisy result for a human face. The reason is the articulation of DSI. The correlation functions of pixels are generally very rugged with multiple strong local maxima as it is illustrated in Figure 3.7(a). The global maximum does not often belong to correct disparity value. The concept from [34] is adopted instead but it is changed to be more reliable for the less-quality DSI. At first, a subset of pixels from the reference image which have strong correspondence in the matching image is found. This pixel group helps to resolve the rest of pixels which have more ambiguous correlation functions.

The pixel with strong correspondence can be recognised by high global maximum in its correlation function and small ratio between the second highest and highest maximum. Experimental example of such correlation function is in Figure 3.7(b). It is highly probable that the disparity with the highest matching score leads to the correct matching pixel in the matching image. Therefore, this disparity value is assigned to the DM for such pixel. The statistics over the reference image are gathered to gain a knowledge about the highest matching score and the ratio between the second best and best score in the correlation functions of pixels. These statistics are exploited for setting a *matching score threshold* t_s and a *ratio threshold* t_r. The purpose of the thresholds is to filter the set of pixels with strong correspondence. This set should cover the largest portion of DM as possible and it should be spread over whole face region. Because the statistical results are used, a possibility that pixels with incorrect disparity (outliers) are added to the DM exists. It is important to eliminate their amount as much as possible because they can negatively influence consequent resolving other pixels.


(a)

(b)

Figure 3.7: The correlation function of the pixel with weak correspondence (a) and strong correspondence (b). Horizontal axis is a disparity and vertical axis is n_{CC} (from -1 to 1).

Several variants of filtering were investigated. Firstly, the pixels with strong correspondence were extracted only with respect to the matching score threshold. But many points which have the highest score on the surface in the DSI were skipped because their global maximum is small with respect to the statistics. In the other hand, many outliers were added to the DM. Secondly, the filtering was made by the ratio threshold. It performed significantly better. The ratio between the outliers and correct pixels was smaller than in the previous case. However, the best results are obtained by the combination of both thresholds.

The pixel is accepted to the initial set when it passes through both thresholds. The size of set can be adjusted by the level of each threshold. The basal value of each threshold is the mean of equivalent characteristics. Moreover, the positive or negative multiple of its standard deviation can be added. In the practice the thresholds are left on the mean values because the resulting DM is a reasonable compromise between the number of correct pixels and outliers. Common values of thresholds are approximately 0.6 for the t_s (range from -1 to 1) and 0.8 for the t_r (range from 0 to 1). The example of DM filled by the pixels with strong correspondence is shown in Figure 3.8. It is restricted by the face region. The grey colour expresses the value of disparity on particular pixel position. Brighter colour means bigger disparity, thus smaller depth. Black colour marks undefined disparity values.

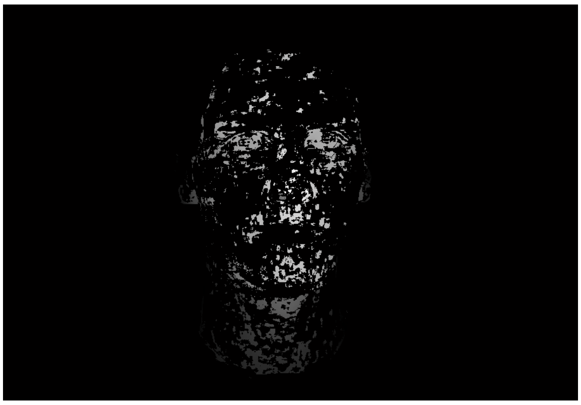


Figure 3.8: The partial disparity map filled for the pixels with strong correspondence.

The initial set of pixels with strong correspondence gives the foundations for incremental filling the rest of DM. Main role is played the neighbourhood consistency constraint. According to that the correct disparity for a pixel should be similar to the disparity values of pixels in its neighbourhood. The pixel with weak correspondence has ambiguous correlation function where it is hard to choose between several comparable local maxima. Experimental example of such correlation function is in Figure 3.7(a). The choice would be easier if the final disparity is searched only in some part of correlation function. The estimate of disparity can be provided by the closest pixels with the strong correspondence. Final value is searched nearby this estimate. There is much higher probability that correct disparity is chosen in comparison to picking weak global maximum. This brings an opportunity that the surface will be followed even through ambiguous parts of DSI.

The DM is completed in an iterative process. A starting point is the set of pixels with strong correspondence which already have entry in the DM. The current set of unresolved pixels is searched for pixels which are on the edge of resolved area in every iteration. The disparity estimate is computed for such pixel from its 8-pixel neighbourhood. The estimate is an average of known values in the DM in mentioned neighbourhood. Afterwards, the closest local maximum to average disparity value is found in the correlation function and its disparity is recorded into DM. During this iterative resolving DM the initial groups of pixels with strong correspondence are spreading in every direction until whole face region is filled and DM is completed.

The outliers from initial set of pixels are spread as well and create incorrect regions in the

DM. The sharp steps exist between the outlying regions and the rest of relatively correct DM. The surface reconstructed from such DM contains extruded peaks in the places of outlying regions which degrade quality of 3D model. To cope with this problem, *a disparity difference threshold* t_d is established. It exploits the neighbourhood consistency constraint on higher level that the disparity difference between adjacent pixels should not exceed certain value. The iterative resolving DM is modified according to this fact. After finding the local maximum closest to average disparity its disparity value is compared to all known disparties in the neighbourhood (those used for average computation). The pixel is resolved in the given iteration only if all differences are below $t_d s_s$. The multiplication with s_s makes t_d independent from the density of DM and more connected to the character of captured surface itself. The threshold t_d is actually maximal disparity difference between adjacent pixels in the full resolution of the reference image. The effect of t_d on resulting DM is that it contains 'worm-like' holes in the places where the rapid disparity changes were previously. They clearly separate outlying regions what results to the patches disconnected from main body of face model. The DM becomes sparser when the threshold t_d is decreasing. The value of threshold $t_d = 3$ pixels separates almost all outlying regions but it does not eliminate pixels near the disparity steps in well-resolved areas of DM.

The surface in the DSI becomes stronger with enlarging matching window, especially for the areas of face which are fronto-parallel to the camera pair. The correlation functions are less articulated what leads to the larger amount of pixels with strong correspondence. Therefore, the set filtered by thresholds t_s and t_r contains less outliers which would contaminate the DM. The fronto-parallel areas of face mainly gain better matching in the DM. Because of that the local technique favours larger matching windows. However, the larger size of window causes that the pixels in the filtered set are more concentrated around strong face features (e.g. eyes, nose) and more uniform parts are less covered (e.g. forehead). It is not suitable for consequent iterative resolving. Moreover, the surface in the DSI looses details with increasing window size, thus the DM describing it as well. The window with size 31 × 31 pixels gives best results for this method.

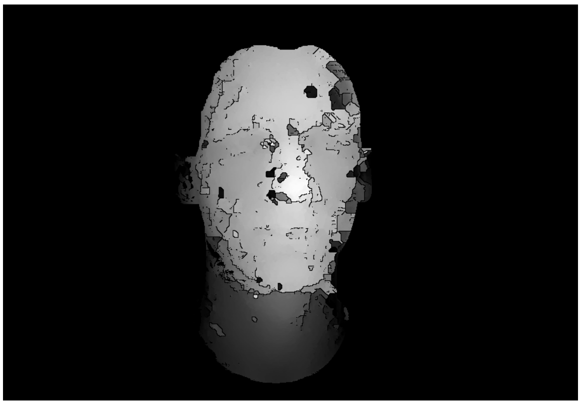


Figure 3.9: The disparity map by local approach.

The DM constructed by presented local approach suffers from imperfections (see Figure 3.9). It contains a number of strongly outlying regions concentrated in slanted areas of face (e.g. sides of nose, sides of face). This is caused by the ambiguity of DSI for these areas. Also more correctly resolved slanted areas contain steps. In some cases the changes of disparity are big enough to be separated by 'worm-like' holes. The reason is that the processed pixels have a tendency to copy the disparity from the neighbourhood instead of its modification along the surface which should be followed. The DM is in these areas piece-wise smooth. Main advantage of local approach is a speed.

3.6.2 Global approach

The global approach to the construction of DM is based on the energy minimisation. The energy function is defined for the DM. It evaluates the shape of DM with respect to the DSI and smoothness requirements. The optimal solution for the DM is obtained by the minimisation of this energy function. The graph cut is employed as a tool for a minimisation. A background from the graph theory is available in Section 2.4.5. The graph with certain form is constructed to reflect the defined energy function. The energy minimisation procedure is then transformed to the minimum cut problem. The final DM is extracted from minimum cut in the graph as the solution with minimal energy. This technique was originally proposed for multiple camera matching in [23]. However, it can be easily simplified for pure stereo matching. Another difference in comparison to this work is the used matching cost. The variance of the intensities in potentially matching pixels is the matching cost in [23]. Whereas, the n_{CC} between matching windows is exploited in this work what is not usual in the context of global methods generally. The majority of them skips the aggregation of matching costs in the DSI.

Various classes of energy functions exist which differ in a definition of smoothness of the DM. The DM modeled as piece-wise smooth is most convenient for capturing general scenes. Because it allows sharp depth discontinuities in the captured scene. The presented minimisation technique is limited to one type of function which models the DM as everywhere smooth. This model constrain the shape of objects which can be accurately described. However, it is suitable for a human face.

The input of used energy function is the DM as it can be seen in Equation 3.15 [22].

$$E(DM) = \sum_{p \in \mathcal{P}} DSI'(p, DM(p)) + \sum_{(p,q) \in \mathcal{N}} \lambda |DM(p) - DM(q)| \tag{3.15}$$

The function consists of two main terms. First term is called a data term. The $\mathcal{P}$ is a set of processed pixels in the reference image. The $DM(p)$ returns the disparity d for pixel p. The $DSI'(p, d)$ accesses n_{CC} for the disparity d in the pixel p and transforms it into the matching cost c. The reason of transformation is that the number summed in the energy function should decrease with better matching because lower energy means better DM. But n_{CC} is higher when there is higher probability that the pair of pixels is corresponding to each other. Therefore, it is normalised to the range 0..1 and inverted to become the matching cost in Equation 3.16.

$$c = 1 - \frac{n_{CC} + 1}{2} \tag{3.16}$$

The sum of c over the set $\mathcal{P}$ evaluates the quality of DM from the viewpoint where the correspondence of each pixel is seen separately. During minimisation the data term is responsible for choosing the best correspondence for every pixel against the required smoothness of DM.

The second term in the energy function is called a smoothness term. It compares the disparity in the adjacent pixels p and q. The adjacency is defined by the relation $\mathcal{N}$ which describes 4-pixel neighbourhood. The absolute difference between examined disparities is multiplied by a *smoothness coefficient* λ. The $\lambda|DM(p) - DM(q)|$ operates as a penalty for the change of disparity in the DM. It exploits a principle of neighbourhood consistency constraint. The linearity of the penalty is the reason why the model of DM is everywhere smooth. The sum of these penalties over the relation $\mathcal{N}$ evaluates the quality of DM in terms of smoothness. During minimisation the smoothness term enforces smooth, continuous shape of DM against choosing best correspondence for each pixel.

The minimisation of this energy function can be directly mapped on the computation of one minimum cut in a graph. The graph with 3D grid topology is embedded into the disparity space. Edge capacities in the graph are set in a way that the cost of cut in the graph exactly corresponds to the energy function in Equation 3.15. The minimum cut is equivalent to optimal DM with respect to the defined energy function. It is guaranteed that the resulting DM corresponds to a global minimum of energy function. This is allowed by the fact that the used energy function belongs to the class of energy functions for which it is computationally possible to find global minimum [31]. For instance, this is not possible for the class with piece-wise smooth model.

The construction of oriented graph in the disparity space is similar to one proposed in [23]. The nodes of graph are created for every point $[U'_L, V'_L, d]$ of disparity space which belongs to the pixel included into the DM. The chain of nodes for each desired pixel spans across full disparity range. There is a 6-connectivity between adjacent nodes, so the graph has 3D grid topology. Figure 3.10 shows simple example of graph in the disparity space where the reference image has resolution 4 × 3 pixels and the disparity range contains 4 values (no face region is considered). The chains of nodes along the axis D are linked by data edges. All data edges are oriented in the direction of increasing disparity value (from source to sink). The data edge between the nodes $[U'_L, V'_L, d]$ and $[U'_L, V'_L, d+1]$ has a capacity equal to the matching cost $c = DSI'(U'_L, V'_L, d+1)$ (illustrated in the detail of interconnection scheme in Figure 3.10). If the DSI' is not defined for some position, the capacity is set to an infinity. This situation can happen when the position leads to the pixel outside of face region in the matching image. One auxiliary layer of nodes is added under the layer with minimal disparity because of the 'shift' of matching costs from nodes to edges. The nodes in each disparity layer are interconnected by the smoothness edges in regular 2D grid. The node $[U'_L, V'_L, d]$ is linked to its closest neighbours $[U'_L - 1, V'_L, d]$, $[U'_L + 1, V'_L, d]$, $[U'_L, V'_L - 1, d]$ and $[U'_L, V'_L + 1, d]$. Two opposite smoothness edges with the capacity equal to smoothness coefficient λ connect each pair of nodes. Special edges with infinite capacity connect the source node with all nodes in the auxiliary layer and all nodes in the layer with maximal disparity to the sink node. Moreover, every data edge has a dual infinite edge oriented in an opposite direction.

A graph cut can be depicted as a surface which divides nodes into a group with source node and a group with sink node (see Figure 3.10). The set of data edges severed by the cut corresponds to the data term in Equation 3.15. The sum of their capacities equals to the sum of matching costs over the set $\mathcal{P}$ in the energy function. Similarly, the set of smoothness edges severed by the cut corresponds to the smoothness term in Equation 3.15. The smoothness edges link the node only to the adjacent nodes in the 4-pixel neighbourhood that is modelled by the relation $\mathcal{N}$. Assume that the cut goes through the data edges on different disparity levels for two neighbouring pixels. To overcome the disparity difference, it has to severe some smoothness edges. The sum of capacities of these edges is expressed by the penalty $\lambda|DM(p) - DM(q)|$. The capacity of only one edge from the pair between two nodes is added into the sum. It is given by the definition of graph cut where the edge is considered as severed only if its orientation leads from source group to sink group. It can be seen that the linearity of smoothness term is implicitly given by a topology of the

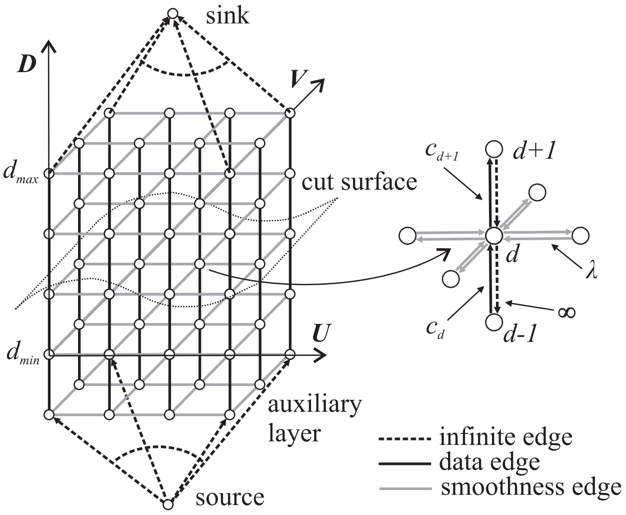


Figure 3.10: The graph constructed in the disparity space by global approach. Detail of interconnection scheme is included.

graph. Other form of smoothness term cannot be modelled by this graph. The DM is simply determined by the disparity values corresponding to the data edges severed by the cut. Therefore, direct connection exists between the surface of the cut and the surface in the DSI which should be described by the DM. The minimum cut is equivalent to optimal DM regarding defined energy function and it is the best description of the surface in the DSI.

The cut has to split the graph in certain way to insure a correctness and completeness of DM. It has to provide exactly one disparity for every pixel included in the DM. The edges with infinite capacity are used to prevent incorrect cuts. To maintain low cost, the minimum cut is forced to avoid edges with infinite capacity. Therefore, the edges to source and sink nodes guarantee that the source and sink are in different groups of nodes after splitting a graph by the cut. Furthermore, the cut severs every chain of nodes in direction of disparity axis so the disparity is known for every pixel in the DM. A purpose of dual infinite edges to data edges is to avoid folds in the surface of cut. The fold in the cut means two and more severed data edges for one pixel. Thus, it leads to an inconsistency of DM (more disparity values for one pixel). Finally, the data edges with infinite capacity push the cut out of the parts of DSI where the matching cost is not defined.

In practice the minimum cut is found in the graph by maximum flow algorithm. It is possible according to the minimum cut - maximum flow theorem (see Section 2.4.5). The maximum flow saturates the set of edges severed by the minimum cut. Thus, the DM is simply extracted from the saturated data edges. The chosen maximum flow algorithm is from family of algorithms based on augmenting paths. But it is optimised for the graphs with grid topology which are often used in the computer vision [10].

The shape of DM can be adjusted by the size of matching window and λ. The DSI becomes more blurred with increasing window size. It is reflected on the DM where the fronto-parallel areas of face are oversmoothed so the course of disparity is almost flat. Moreover, the slanted areas contain flat regions connected by big steps. Decreasing size of window improves the DM, slanted and also fronto-parallel areas become more correct and smooth. The face details such as eyes are better described as well. However, a noise is becoming apparent together with appearing face details in the case of small windows. The reasonable compromise is the matching window with the size 11 × 11 pixels. The λ influences the smoothness of the DM on a local scale. The increment of λ smooths the DM, however small face details are destroyed. The decreasing λ a bit reduce a flatness in the case of larger matching windows. For smaller windows it brings the details to the DM but it less eliminate the noise. The λ set to approximately one eighth of the average matching cost leads to the best results ($\lambda = 0.025$ for 11×11 window). Similar experimental observation was made in [18].

The DM constructed by the presented global approach describes reasonably the response of face in the DSI. Figure 3.11 shows that the DM is smooth in the majority of face region. The method shows an ability to follow the surface even through ambiguous parts of DSI. The DM contains small stepiness in the strongly slanted areas. The outlying regions are present only for occluded areas such as ears. The transition between the chin and the neck is problematic because of sharp depth step. The used energy function does not support the rapid changes in the disparity, thus the transition is oversmoothed (several small steps instead of one big). However, the resulting DM is a significant improvement with respect to the local approach. On the other hand, the constructed graph is huge for the used resolution of face images (tens of millions nodes). This implies high memory demands. Moreover, the construction of large graph and maximum flow computation in it take long computational time.

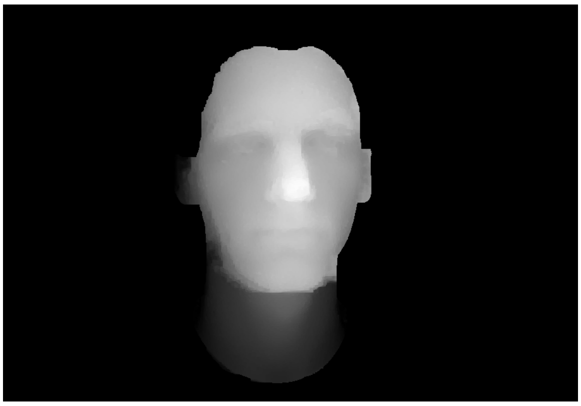


Figure 3.11: The disparity map by global approach.

3.6.3 Global approach with the estimate by local technique

The high-quality results of global approach are overshadowed by its computational and memory demands. To address this problem, a technique combining the local and global approach was developed. Basic concept is to exploit the DM produced by the local technique as an initial estimate for the global approach. The local solution is used for a reduction of search space. The following global optimisation of DM is performed on the reduced graph. This fact leads to the memory savings and shorter computational time. The survey in [32] presented different methods of reducing search space for the stereo correspondence using graph cuts. The alpha-expansion algorithm based on multiple graph cuts was used for energy minimisation. It is a difference from this work, however the methods for search space reduction are applicable to the 3D grid graph as well. The three techniques were compared - an estimate by local technique, hierarchical computation and best-candidate approach. The initialisation by fast local technique brought the biggest memory and time savings.

The estimate of DM is computed by the proposed local technique. It is a basis for modelling *a volume of interest* in a disparity space for the global optimisation. The first step in modelling is a creation of thin layer around initial DM. A thickness of layer is determined by an *offset* o_l. The boundary surfaces of layer are duplicates of initial DM shifted by o_l along D axis in both directions as it is shown in Figure 3.12. The disparity range set by the layer for each pixel is then expanded according to the pixel's neighbourhood. The minimal boundary of range is updated if lower minimal boundary exists within *an expansion region* (equivalently for the maximal boundary). The expansion region has square shape with size $(2w_{er} + 1) \times (2w_{er} + 1)$ pixels and it is centered around the examined pixel. The reason for layer expansion is an elimination of influence of outliers in the estimate of DM. Figure 3.12 depicts a situation where the strong outlier is situated in the centre of the estimate. The thin layer created by o_l does not include the correct DM. But the expansion of layer creates the volume of interest big enough to contain whole correct DM. Hence, there is still an opportunity to extract correct DM from the volume of interest. The full elimination of outlying region is possible when the expansion region has similar size. Another reason for expansion is to provide the chance to find smooth DM when the estimate contains step changes of disparity. If the estimate of DM does not provide a disparity for a pixel, the volume of interest is enlarged to full disparity range for this pixel. Finally, the volume of interest is trimmed by the volume where the DSI is defined.

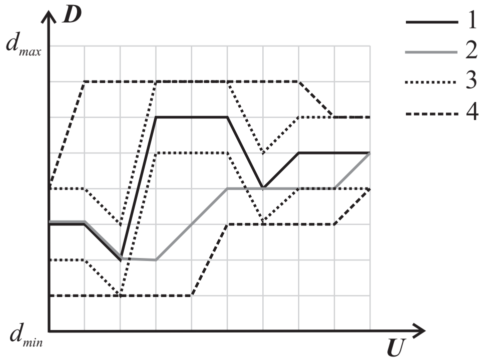


Figure 3.12: Modelling the volume of interest in a disparity space. A slice along row in the reference image is illustrated (10 pixels and 9 disparity layers). (1) the estimate of DM, (2) correct DM, (3) the layer created by $o_l = 1$, (4) the layer expanded by $w_{er} = 2$.

The construction of reduced graph starts by creating 3D grid of primary nodes for the volume of interest inside the disparity space. Figure 3.13 illustrates the reduction of graph where the grey thin grid in the background represents an extent of full graph (all disparity levels for each pixel). The interconnection scheme is same as in purely global approach except for the edges to the sink and the source. Because the first and last layer of nodes do not exist. Figure 3.13 does not depict both smoothness edges and the infinite dual edges to the data edges for the simplicity. The orientation is marked only on the edges to terminals for the same reason. The graph constrained by the volume of interest can be very rugged (especially in the case of face). Thus, the chains of nodes for adjacent pixels have usually different lengths. It brings a possibility to create a step in the DM which is not penalised by the smoothness term because there are no smoothness edges. The solution is to wrap upper and lower boundary of graph by auxiliary nodes (see Figure 3.13). These nodes are the ending points for missing smoothness edges. The arrow in Figure 3.13 marks a situation when valid cut inside the volume of interest is correctly penalised by additional smoothness edge. Data edges between auxiliary nodes have infinite capacity because those disparity values are outside the volume of interest. After the addition of auxiliary nodes each chain of nodes corresponding to one pixel can be linked to the source and the sink. The node on minimal disparity level in the chain is connected to the source and the node on maximal disparity level to the sink. It does not matter if it is primary or auxiliary node.

The difficulty with the graph in this form is that the time of graph cut computation is not dependent only on the number of nodes and edges but also on the shape of graph. When the volume of interest has a shape close to a cuboid (searched DM is 'flat'), the graph has a shape of regular 3D grid. The speed of computation significantly grows with the decreasing number of nodes and edges in the graph. When the volume of interest has a rugged boundaries, the reduced graph has complex shape distant from regular 3D grid (unfortunately the case of DM for human face). The computational time decreases slower or can even grow with the reduction of nodes and edges. The actual trend strongly depends on the complexity of graph. The reason of this behaviour is the optimisation of the used maximum flow algorithm to the regular grid topology of graph. To cope with this problem, extra infinite edges to the source and the sink are added (gray dashed edges in Figure 3.13). They connect the terminals to the rest of nodes which are on the boundary of graph but not in direction of D axis. In terms of maximum flow computation a 'liquid' is now quicker brought into main body of graph. The result of this modification is that the shape of graph has smaller influence on the speed of maximum flow algorithm. Therefore, the reduced graph brings also shorter computation of DM except the memory savings.

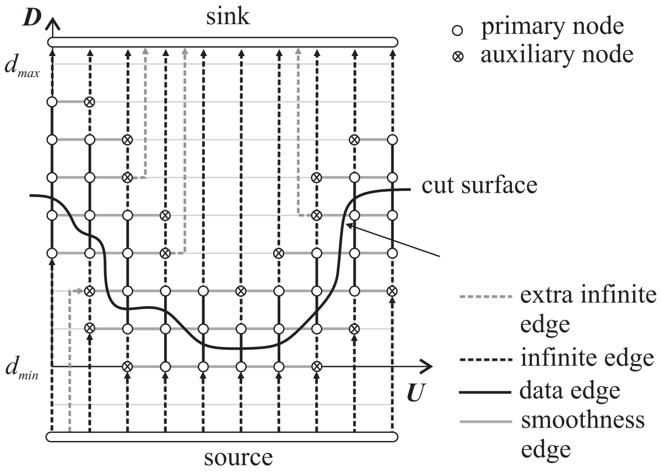


Figure 3.13: The slice of reduced graph along a row in the reference image (10 pixels and 9 disparity layers). The depiction of edges is simplified except for the edges to terminal nodes.

The minimum cut in the reduced graph minimises an energy inside the volume of interest. The obtained DM can differ from the DM computed from the non-reduced graph. A solution identical to the one obtained through the purely global technique can be found if optimal DM is entirely inside the volume of interest. The reduction of the graph does not influence the shape of minimum cut in that case. The reason is that the excluded parts of full graph consist of the edges with limited capacity which were not saturated with maximum flow. These parts are replaced by the edges with infinite capacity after the graph reduction. Therefore, the situation is same in terms of flow computation and the maximum flow saturates same set of edges defining optimal DM. If the volume of interest does not fully contain the optimal DM, the minimum cut is forced by the boundaries of graph to different solution. This solution is optimal within the volume of interest but sub-optimal with respect to whole DSI.

The effect of matching window size and λ on the DM is same as in purely global approach because the principle of energy minimisation did not change. The size of reduced graph is directly adjusted by the offset o_l and the size of expansion region w_{er}. The memory consumption grows together with increasing o_l and w_{er}. The time of DM computation decreases with shrinking graph. The quality of initial estimate by local technique significantly influences the resulting DM and computational demands. The regions with incorrect matching decrease an extent of graph reduction because o_l and w_{er} have to be enlarged to eliminate their influence. Large outliers in the initial DM can even prevent the graph cut from finding globally optimal solution because the volume of interest does not contain optimal DM. Moreover, the outliers cause more complex shape of graph what leads to longer maximum flow computation. Therefore, the parameters of local technique have to be adjusted to achieve the best estimate of DM. In practice the extra DSI has to be computed because the local technique achieves better results with larger matching windows. The global technique uses another DSI created by smaller windows. The computational overhead connected with the estimation of DM is primarily generated by the mentioned construction of DSI. The time of DM construction by local technique is negligible in comparison to it. Moreover, the initial DM is post-processed to fill the 'worm-like' holes (see Section 3.6.4). The reason is to avoid the expansion of the volume of interest to full disparity range for these areas. It leads to more complex shape of graph.

A tradeoff between the efficiency and the accuracy of DM is present in the proposed method. A price for the solution similar to the purely global approach is small reduction of memory consumption and computational time. However, the experiments showed that significant speedup and memory savings with respect to the global technique can be gained with the reasonable loss of precision. Figure 3.14 shows an example of the constructed DM. It is similar to the DM by purely global approach. The side of face only is worse solved because the estimate contained many strong outliers in that part.

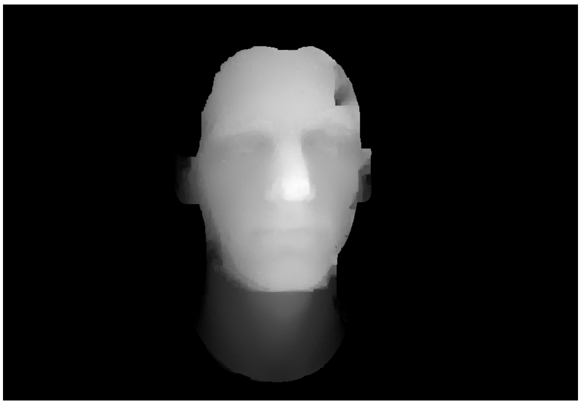


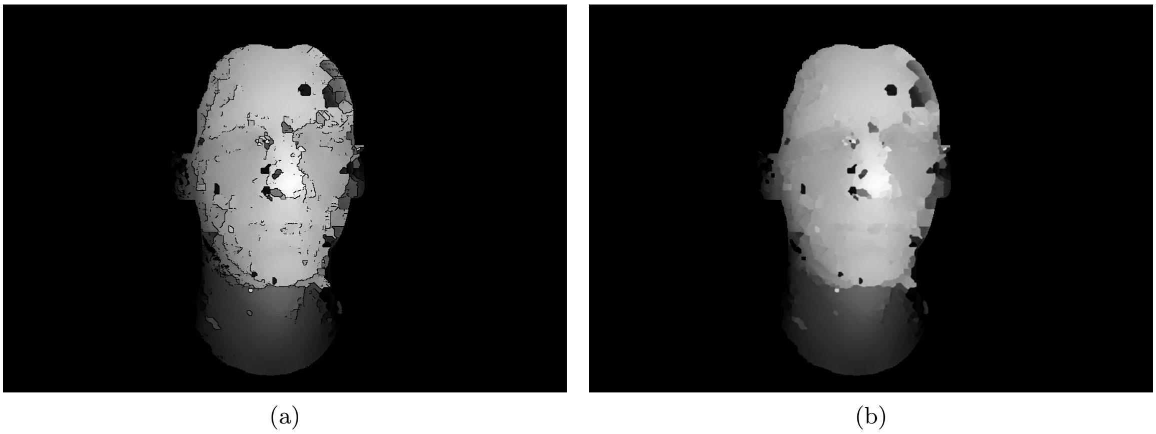
Figure 3.14: The disparity map by global approach with the estimate by local technique.

3.6.4 Post-processing disparity map

The DM constructed by any of the presented methods can be post-processed to eliminate unwanted artefacts. Several procedures which target different problems are explained in this section.

The first procedure is a *hole filling*. The holes with undefined disparity are present in the DM created by the local approach (see Figure 3.15(a)). They are a result of disparity difference threshold. Their presence is welcomed when the 3D model is reconstructed from the DM. The outlying regions in the DM are transformed into the standalone patches which are completely disconnected from the face surface. Otherwise, they would create the extruded peaks from the main surface what more degrades the visual quality of the model. However, the hole filling is valuable when the DM is used as an estimate for the global technique. It provides at least some estimate of disparity for these pixels and prevent the growth of graph (see Figure 3.15(b)).

At first, the holes in the DM have to be detected. The binary mask defining the pixels with known disparity is modified by the morphological closing operator. The operator is a square with size $(2w_h + 1) \times (2w_h + 1)$ pixels. The w_h needs to be only few pixels to fill the 'worm-like' holes. The substraction of old mask from the closed mask gives the set of pixels missing disparity. The consequent process is similar to the resolving the pixels with weak correspondence in the local technique. The holes are filled iteratively, the pixels on the edges first. The average disparity is calculated from the known values in the 8-pixel neighbourhood of the examined pixel and assigned to its position in the DM.

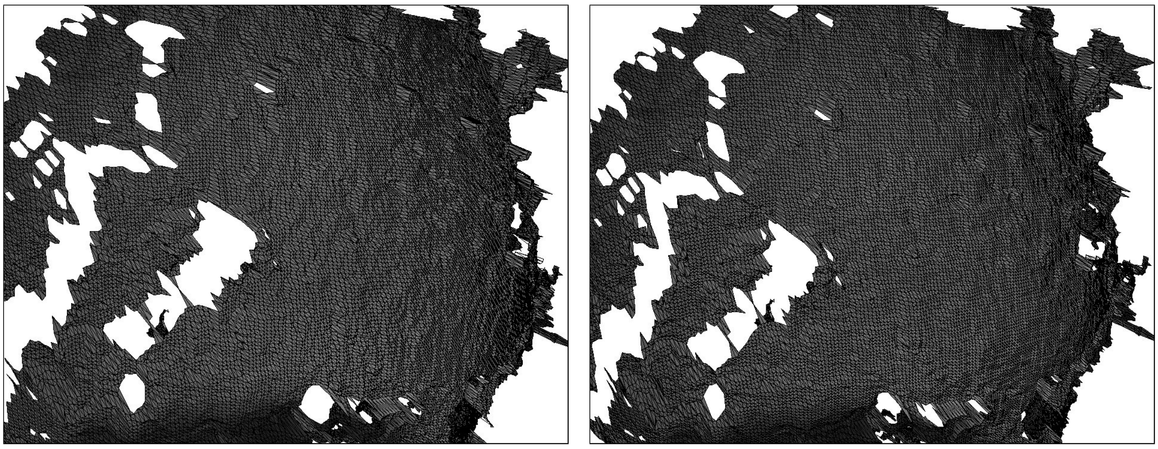

(a)

(b)

Figure 3.15: The disparity map with holes by the local technique (a) and after the hole filling (b).

In all previous stages the disparity is expressed as integer number. The whole pixel units are necessary for the exact alignment of matching window with the pixel grid in the matching image. But the surface reconstructed from integer disparity map contains steps between depth layers defined by the individual values of disparity. The the smoothness of surface can be enhanced by the *sub-pixel disparity refinement*. The technique used for the surface reconstruction does not require integer disparity values. The refinement is done by fitting parabola to the matching score values in the discrete correlation function for every pixel in the DM [20]. The small part of the correlation function in the DSI is examined around the disparity value recorded in the DM. It is sufficient to consider only 2 disparity levels on each side of this value. The continuous parabola approximates the sequence of n_{CC} values in these 5 positions. It is fitted in the least-square sense. The disparity in the DM defines the position of some maximum in the correlation function, therefore the fitted parabola is concave and forms continuous peak. The position of parabola peak on the disparity axis defines the new floating point value of disparity for the processed pixel.

The actual effect of this refinement depends on the articulation of the DSI. The reconstructed surface is slightly smoothed in the case of DSI by larger matching windows. Thus, it is suitable for the local method of constructing the DM. Figure 3.16 shows that the transitions between depth layers are not visible after the refinement. But the improvement is on the level of the depth resolution and it does not effect the bigger steps on the surface. In the case of the DSI by smaller matching windows the refinement does not smooth the surface but brings small noise. Therefore, it is not useful for the post-processing the DM created by the global approach.



(a) (b)

Figure 3.16: The surface reconstructed from original disparity map (a) and disparity map after sub-pixel disparity refinement (b).

The surface of model reconstructed from the integer DM by any of presented techniques suffers from local bumpiness. Its extent exceeds the transitions between depth layers. It is caused by a noise in the DSI. The sub-pixel disparity refinement does not cope with this problem because it operates on smaller scale than the size of present bumps. Therefore, the *2D Gaussian operator* is used to explicitly smooth the DM. It covers the area $(2w_g + 1) \times (2w_g + 1)$ pixels. The strength of smoothing is adjusted by the standard deviation σ_g. Because the 2D Gaussian operator is separable, the DM is filtered consecutively by 1D filter in the horizontal and vertical direction. The result of filtering are new floating point disparity values for all previously defined positions in the DM. The independence of Gaussian operator from the DSI allows its application on the result of global technique. It is an advantage over the sub-pixel disparity refinement. The adjustability of smoothing is beneficial for the elimination of the mentioned bumpiness. However, the small facial details (e.g. a shape of eye) are disappearing together with the bumpiness. It is necessary to find balancing value of σ_g. Figure 3.17 illustrates that the Gaussian smoothing DM enhances visually the quality of surface.

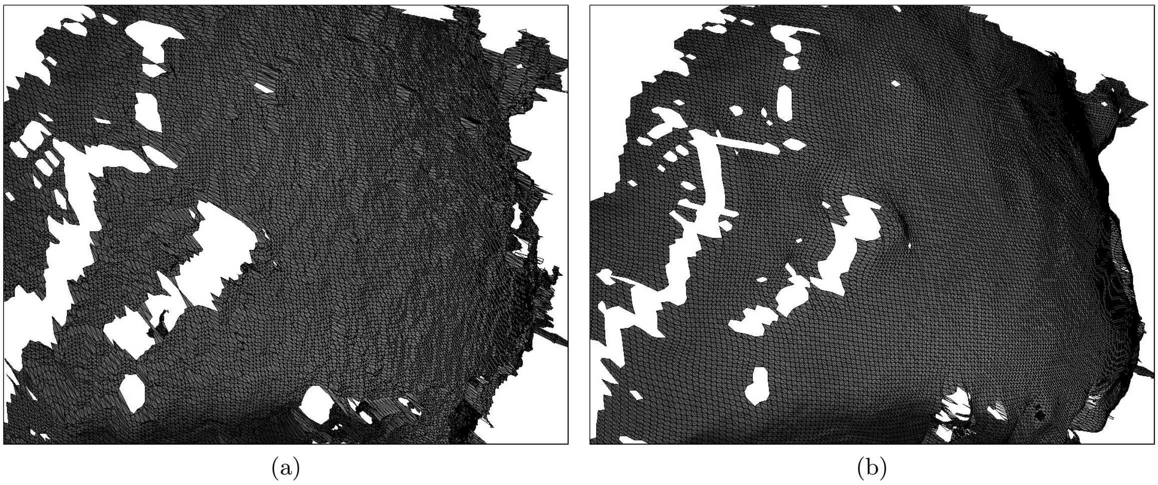


Figure 3.17: The surface reconstructed from original disparity map (a) and disparity map after smoothing (b).

3.7 Surface reconstruction

The last stage of 3D capture is a transformation of DM to the 3D model of object. Several ways of 3D point reconstruction from the pair of matching pixels are presented in Section 2.3.3. The computation of spatial points from DM results in the non-structured cloud of 3D points in the world coordinate system. The interconnections between the points are defined by the adjacency of their projections in the DM. The final model of face is represented by the textured triangle mesh.

3.7.1 3D point computation

The algebraic solution presented in [34] is used for the computation of spatial point. The position of point can be calculated from information which are already available from previous stages. No iterative optimisation is necessary like in the geometrical methods (Equation 2.7). It is simpler than the calculation based on baseline length (see Equation 2.12). Because it does not need the length of baseline and the disparity value is not transformed in the normalised image coordinates. Finally, it computes the overdetermined system of linear equations only with two variables in comparison to the algebraic method represented by Equation 2.9.

The image coordinates of two corresponding projections of 3D point are derived from the DM. The projection p'_L in the left rectified image (the reference image) is simply defined by the position in the DM. It has to be multiplied by the scanning step s_s first to obtain the actual pixel coordinates $[U'_L, V'_L]$ in the left rectified image. The pixel coordinates $[U'_R, V'_R]$ of the projection p'_R in the right rectified image are calculated from $[U'_L, V'_L]$ and the disparity value on the processed position of DM (see Equation 3.8). The projection matrices H'_L and H'_R defining the projection of spatial point $\tilde{P}_W$ on the rectified images are known from the image rectification. The relationship between the image pixels and 3D point is expressed by Equation 3.17.

$$\tilde{p}'_L = H'_L \tilde{P}_W \quad \tilde{p}'_R = H'_R \tilde{P}_W \tag{3.17}$$

The extended variant necessary for the next step is shown in Equation 3.18.

$$\begin{bmatrix} w'_L U'_L \\ w'_L V'_L \\ w'_L \\ 1 \end{bmatrix} = \begin{bmatrix} H'_L \\ \mathbf{0}|1 \end{bmatrix} \cdot \begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} \quad \begin{bmatrix} w'_R U'_R \\ w'_R V'_R \\ w'_R \\ 1 \end{bmatrix} = \begin{bmatrix} H'_R \\ \mathbf{0}|1 \end{bmatrix} \cdot \begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} \tag{3.18}$$

These equations can be joined using $\tilde{P}_W$ as it is expressed in Equation 3.19.

$$\begin{bmatrix} H'_L \\ \mathbf{0}|1 \end{bmatrix}^{-1} \cdot \begin{bmatrix} w'_L U'_L \\ w'_L V'_L \\ w'_L \\ 1 \end{bmatrix} = \begin{bmatrix} H'_R \\ \mathbf{0}|1 \end{bmatrix}^{-1} \cdot \begin{bmatrix} w'_R U'_R \\ w'_R V'_R \\ w'_R \\ 1 \end{bmatrix} \tag{3.19}$$

The unknown coefficients w'_L, w'_R in Equation 3.19 are computed by QR decomposition. Each of them can be substituted into Equation 3.17 and express the position of 3D point. Potentially, two slightly different spatial points can be obtained due to inaccuracies in the calibration and the image rectification. The spatial average of these two 3D points would lead to the occlusions of points in the left view in some cases what does not fit a reality. Therefore, the coefficient w'_L only is used to acquire the 3D point which is consistent with other points from the reference image.

3.7.2 Construction of surface

The combination of a *DM* with the computed 3D points forms a range image of face from the left view. Each 3D point is directly linked to the position in the *DM* which was computed from. A triangulation between spatial points is determined from the neighbourhood of their positions in the *DM*. The *DM* is traversed along rows and the square groups of four positions are examined. According to the existence of spatial points for these positions 0, 1 or 2 triangles are created. The possible constellations of 3D points with their triangulation are illustrated in Figure 3.18. The result is compact triangle mesh representing a face. This technique is a simple adaptation of complex solution introduced in [15]. They presented a step discontinuity constraint which terminates a triangulation when the processed 3D points are too distant from each other in terms of a depth. It is not exploited because the surface of human face does not contain such large depth steps. However, the similar task is performed by the disparity difference threshold on a level of the *DM* in the local approach. The reason is that the local technique is inclined to produce the *DM* with depth steps.

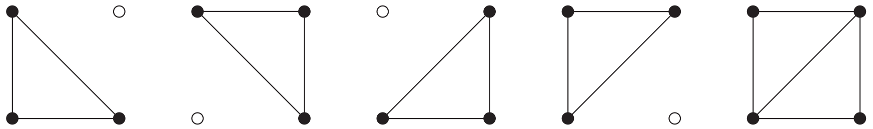
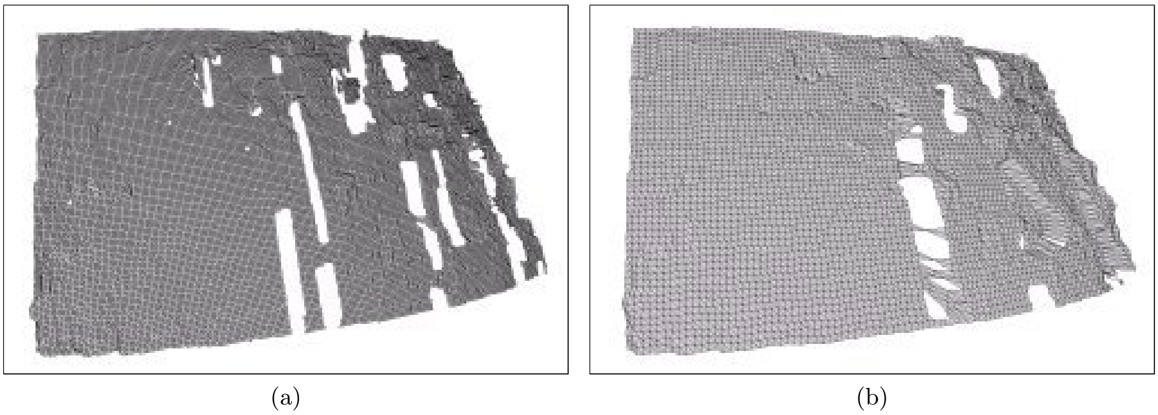


Figure 3.18: The possible constellations of the present 3D points and the resulting triangulation.

A density of triangle mesh can be adjusted by the scanning step (see Figure 3.19). The experiments showed that s_s equal to 4 pixels is a reasonable compromise between the level of detail of reconstructed model and computational demands for 8.2 megapixel images. The reference image is stored as an texture for the mesh. The texture coordinates are assigned to each 3D point. They are actually the pixel coordinates of the projection of 3D point in the reference image. The textured triangle mesh is stored in VRML format [7].



(a)

(b)

Figure 3.19: The surface patch from a forehead reconstructed using $s_s = 2$ (a) and $s_s = 4$ (b).

Chapter 4

System implementation

The implementation of 3D face capture system is explained in this chapter. The structure of capture rig is described including the camera and light setup. The image acquisition and calibration procedure are outlined. The description of the implemented software is included.

4.1 Capture rig

The capture rig is based on two high-resolution digital still cameras connected to computer and supplemented by the system of lights. The individual components of system and their main properties are described in this section.

4.1.1 Camera setup

The foundation is the digital still camera Canon EOS 30D. It is single-lens reflex, auto-focus, auto-exposure camera with built-in flash. It is equipped by EF-S 18-55 mm lenses and 8.2 megapixel high-sensitivity CMOS sensor. Natively supported image file formats are JPEG and RAW (Canon proprietary format CR2). A communication interface is USB 2.0 Hi-Speed. More technical details can be found in [1].

The scheme of rig with approximate dimensions is shown in Figure 4.1. The practical experiences showed that the baseline long 20 cm and the distance between the person and cameras equal to 70 cm are suitable. The cameras are approximately in the height 135 cm. This configuration guarantees proper acquisition of the front part of a face. The capture volume created by camera fields of view contains whole head and the neck of scanned person. The focused volume restricted by the depth of fields of cameras is smaller but sufficient. Its depth range is approximately 20 cm in the direction of camera view. A panel with dark colour is situated behind a person to have neutral and uniform background in the images.

The existing USB interface is used for controlling a camera and transferring images to a computer. Each camera is connected by passive USB extension cable to USB hub D-Link DUB-H4 [2]. The hub is connected by one USB cable to the computer and is supplied by external power adaptor. This solution saves native computer USB ports and gives a possibility to work with both cameras simultaneously.

The transfer of images to a computer can be controlled from the side of computer. This fact is important because it is not possible to manipulate with cameras after the calibration. A communication with camera is driven by Picture Transfer Protocol which is commonly used by digital camera manufacturers. The software EOS Utility for the operating system

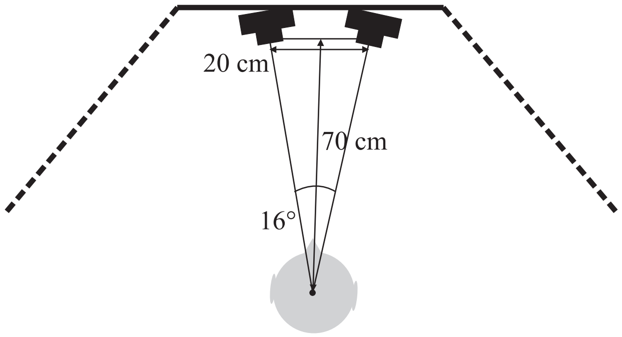


Figure 4.1: The camera setup in a metal construction.

Windows is supplied with a camera [1]. It allows to transfer images and control remotely a reduced set of camera parameters. The image transfer can be realised by the application Gtkam which is based on the library libgphoto under the operating system Linux [4].

It is necessary to trigger the cameras simultaneously because the face has to be captured in exact same spatial position in both images. This is not possible to realise it by the EOS Utility. Because it has support for remote triggering but it can communicate with only one camera at the moment. A hardware solution is employed instead. The camera has special port for hardware remote triggering by the remote switch RS-80N3 from the manufacturer [1]. Special remote triggering system for multiple cameras was manufactured from original remote switch and extension cables (many thanks to Mr. Birt).

4.1.2 Light setup

The purpose of light system is to provide uniform light coverage of face without shadows or reflections. It is achieved by strong omnidirectional light scattered in the capture space. Directional lights are not present in the capture environment to avoid the reflections. The light setup was changed several times because of changing availability of lighting components and a relocation of rig to another laboratory. Two most used variants are described.

Figure 4.2(left) illustrates the first variant. White light tubes are attached on the metal construction which partially surrounds the scanned person. The horizontal set of lights in the height of cameras illuminates a central part of face from the front and the sides. Vertical tubes provide proper illumination of chin and neck. The wall behind rig is illuminated by strong directional light to reflect scattered light on a top of head. A face is not underlit by the light tubes on the construction then. However, the upper part of face is slightly darker in comparison to lower part.

The second variant is illustrated in Figure 4.2(right). The difference from first variant is a lack of illumination from the sides. But the one camera pair covers primarily the frontal part of face which is properly illuminated. Another difference is vertically and horizontally symmetric light configuration in the shape of letter H. It provides same amount of light to the upper and lower part of face which eliminates the unbalance of brightness in the fist variant of setup.

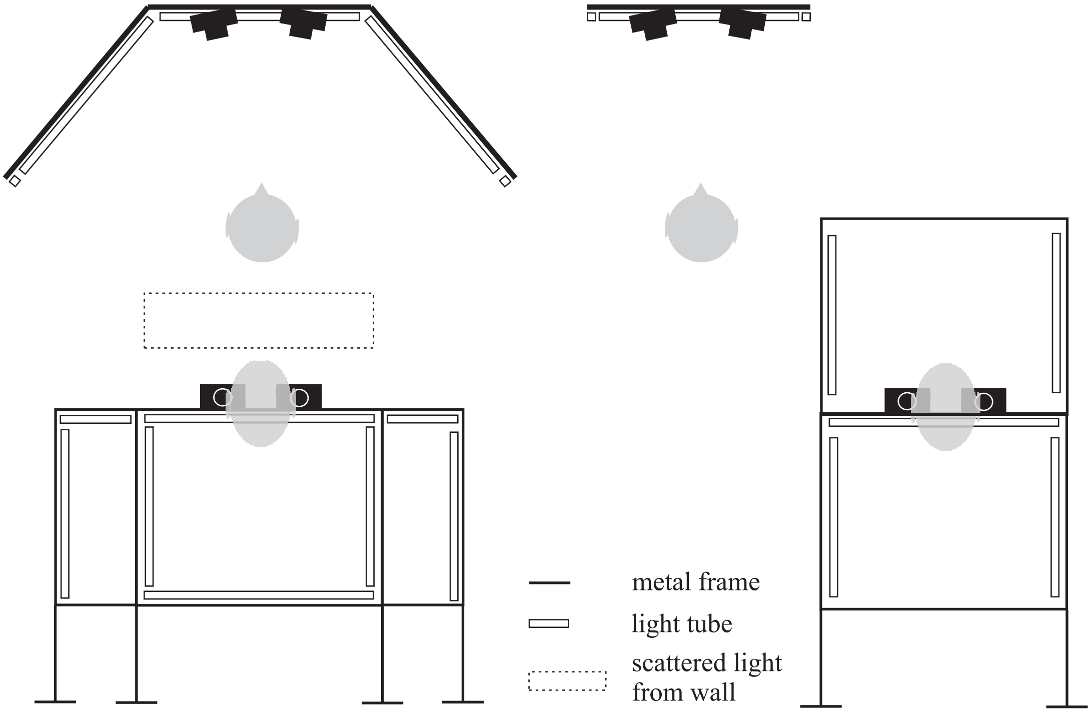


Figure 4.2: The top and front view on the light setup - first variant (left), second variant (right).

4.2 Image acquisition

The face images are captured with red-eye correction but without a flash to avoid reflections. The cameras are manually focused on the capture volume because auto-focusing causes a change of focal length. Any change of optical system is not permitted after the camera calibration. Both cameras work in manual programme where a exposure time and an aperture can be set by user. It gives a opportunity to adjust the depth of field. It is desired to enlarge the focused volume as much as possible, however there is a trade-off with a quality of images. The depth of field can be widened by decreasing aperture size. But less amount of light passing through optical system has to be compensated by a shutter speed and ISO speed. The maximum time of exposition is experimentally one tenth of second. The reason is that the images would be blurred by a movement of head during longer time. The remaining possibility is to increase a sensitivity of sensor described by ISO speed. But higher sensitivity deforms correct colour acquisition and brings a noise. Final compromise leads to approximately 20 cm wide depth of field which is reasonably large to include whole face.

Full manual control of settings prevents several problems with the synchronisation of cameras. Same settings on both cameras which are not automatically adjusted lead to same internal processing a sensor data. It means for instance that the cameras have similar gain and the colours on the images do not differ much. Despite the fact that the cameras receive the external impulse for capturing at same time, the internal logic postpones it a bit individually. If both cameras are identically configured, the actual times of capture by cameras are not very distant as well. However, some amount of desynchronisation between cameras still exists in terms of a time or properties of images because of their individual manufacturing.

The images are acquired in the highest resolution (3504x2336 pixels) and the colours are represented in sRGB colour space. They are stored in the RAW format because it has several advantages. It stores sensor data which are not compressed and processed by a camera. The image post-processing is completely in user hands and possible loss of information can be controlled. It is important because every captured information is beneficial for the stereo matching. The pictures are post-processed in the software Digital Photo Professional [1]. It is proprietary software for the operating system Windows distributed together with a camera by manufacturer. Different colour appearance of face between images can be reduced by adjusting brightness and white balancing. However, full manual configuration of cameras and their close distance ensure similar face appearance. Thus, a software improvement is necessary occasionally. Finally, the images are losslessly saved in the TIFF format which is more suitable for following processing. The alternative for post-processing RAW format and a conversion to standard formats under the operating system Linux is open source software dcraw [3] and products derived from it.

4.3 Calibration

The cameras are calibrated by the widely used Camera Calibration Toolbox for Matlab [9]. It provides complete framework for an extraction of reference points from images, a calibration and an analysis of solution. The camera model implemented in the toolbox is briefly explained in Section 2.2.2. But it is not exploited completely in practice. Experiments showed that skew coefficient α and third coefficient of radial distortion k_5 are negligible or cause numerical instability of solution for the used type of cameras. Because of these facts they are excluded from the calibration procedure and assumed as zero. The camera model with this modification is similar to the model in [35].

The calibration object is shown in Figure 4.3. The reference points used for the calibration are the corners of black squares. The dimensions of each square are 15×15 mm and they are arranged in 7×7 matrix. The calibration object provides together 196 co-planar reference points. The coordinate system related to the calibration object has an origin in one of the corners of pattern. Its x and y axes are aligned with the sides of pattern and z axis is perpendicular to the board surface. The calibration board is captured in 8-10 different positions with respect to the camera pair. One position is chosen as principal and it defines the world coordinate system. The calibration board in principal position should be oriented in a way that whole square pattern is properly observed by both cameras. Because the extrinsic parameters of cameras are computed with respect to the world coordinate system. The intrinsic parameters of each camera are computed from the full set of images of calibration board in different positions. It is not allowed to move or change the parameters of cameras after capturing the calibration images.

The processing calibration images by the toolbox consists of several steps. At the beginning a user have to specify the properties of calibration object. Then he has to mark four extreme corners of pattern on all calibration images. They should be marked in certain order to give the information how the coordinate system is related to the calibration object. These information are used for the automatic extraction of the rest of square corners and determining their spatial positions. The specified parameters of camera model are then analytically estimated and iteratively optimised (see Section 3.3). The result of calibration procedure for each camera consists of the focal lengths f_x, f_y, the coordinates of principal

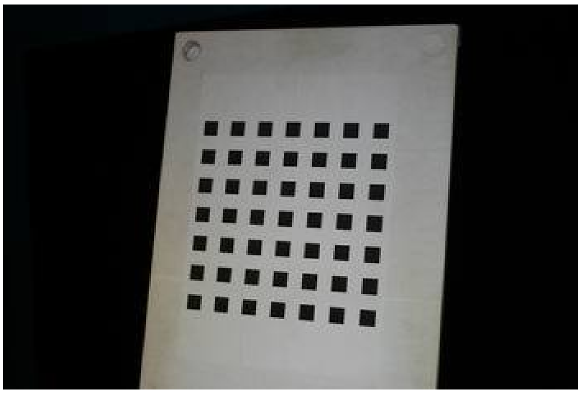


Figure 4.3: The calibration object - white board with periodic square pattern.

point u_0, v_0, the radial and tangential distortion coefficients k_1, k_2, k_3, k_4, the rotation angle vector and the translation vector with respect to the world coordinate system.

The quality of calibration can be described in terms of the re-projection error. The error for one reference point is a difference between its extracted position and the position after re-projecting through computed parameters (see Equation 3.1). The statistics of re-projection error through all reference points in all images gives the insight to an accuracy of solution. The standard deviation of error in horizontal and vertical direction is used for final assessment. The solution can be refined by adjusting extracted corners and repeating optimisation. The results of calibration are crucial for following stages, especially for surface reconstruction. The reason is that majority of calculations using view geometry needs these information. Because of this fact the highest possible accuracy should be achieved. Otherwise, the potential error will accumulate in next stages and could lower precision of whole system.

To simplify and automatize the calibration procedure, the own interface to the toolbox was implemented. The majority of settings are preset because the properties of capture rig and calibration object are known. The user needs to run one Matlab function for each camera. The inputs are the set of calibration images for the camera, the index of image where the calibration board is in the principal position and the size of search window for automatic detection of square corners. Only user interaction is marking the extreme corners of pattern. Otherwise, the user does not need to proceed between individual steps through GUI such as in the original toolbox interface. The output is the text file with the calibration results which is used by other programs in the face reconstruction.

4.4 Software

The implemented software provides a solution for all image processing stages in the 3D face capture system. It comprises all procedures in the image pre-processing, the construction of disparity space, the construction of disparity map and the surface reconstruction. The methods and algorithms described in Chapter 3 are implemented in a form of classes. The classes are integrated into one library which provides all designed functionality. An access to this functionality is realised by several command-line programs. The inputs to this chain of programs are the calibration data of each camera and the pair of face images. The output is the resulting 3D model of face.

The mentioned library and programs are implemented in the C++ programming language. They were developed using the compiler GCC and GNU Make system. The target platform is the operating system Linux. The implementation is extensively based on the Recognition and Vision Library (RAVL) [6] developed in the Centre for Vision, Speech and Signal Processing at the University of Surrey in United Kingdom. Another used external library is MAXFLOW [5].

The RAVL provides a base C++ class library together with a range of computer vision, pattern recognition, audio and supporting tools. It has own easy-to-use make system built above the standard GNU Make which simplifies building large projects. It is distributed under the Lesser GNU Public License. The implemented software works with the types and data structures such as dynamic arrays or lists provided by the RAVL (alternative to STL functionality). Another area of used functionality is connected with advanced mathematical calculations - operations with vectors and matrices, solving systems of linear or non-linear equations, computation of statistical quantities, etc. The powerful loading and saving of images or triangle meshes is exploited as well. The basic image processing routines such as morphological operators or bilinear interpolation are occasionally called as well.

The library MAXFLOW provides the computation of maximum flow/minimum cut in a graph. The algorithm works with the augmenting paths and is optimised for the graphs with grid topology [10]. A user can construct own oriented graph with the source and the sink node. After the computation the value of maximum flow and the division of nodes by the minimum cut is known. It is distributed under own research license. The library is used in the global optimisation of DM by the graph cut.

4.4.1 Library `StereoFaceCapture`

The library `StereoFaceCapture` contains a set of C++ classes. The classes represent the important data structures or functionality blocks identified during the design of system. The classes are independent on each other to a great extent. Each of them encapsulates the closed cell of related operations and is useful as standalone unit. A speed of computation and memory consumption were taken into account during a development.

The class `DistortionRemoverC` is dedicated to the image undistortion (see Section 3.4.1). It basically encapsulates a camera optical system described by the camera model in Section 2.2.2. It needs only intrinsic camera parameters. The difference is that only two radial distortion coefficients k_1, k_2 are used. The coefficient k_5 is already excluded during the calibration of camera. Main function is the transformation of distorted image to the image without lens distortion according to the set intrinsic camera parameters. The foundation are the methods for the bidirectional transformation between the image coordinates and normalised image coordinates through the camera intrinsic matrix A or the combination of A and lens distortion. They allow the mapping between distorted and undistorted pixel according to Equation 3.2. To make the undistortion procedure effective, the mapping table is pre-computed for whole area of output image for the certain size of input and output image. The table can be reused for the transformation of images if the sizes of images are not changed.

The class `RectificatorC` is dedicated to the rectification of image pair (see Sections 2.3.2, 3.4.2). The first function is to compute the the new projection matrices H'_L, H'_R from original matrices H_L, H_R for both cameras. The image transformation matrices Q_L, Q_R are computed as well. Second function is a transformation of undistorted images to their rectified equivalents through the matrices Q_L, Q_R. The additional product of the transformation are the binary masks which mark the relevant area in each rectified image (the rest of pixels is filled by background colour). The masks are stored as black and white images.

The class `FaceExtractorC` is dedicated to the extraction of face region from the image (see Section 3.4.3). The model of skin colour is obtained from the square window with size 1/100 of image height. It is assumed the face covers the majority of image so the window is placed in the centre of image. The extraction is controlled by the sensitivity parameter tol and the size of morphological closing operator. The obtained region can be optionally shrunk by erosion operator. The reason is that the matching window will contain only a skin even when it is placed on the edge of face region. The result is a binary mask in a form of black and white image. The skin modelling and initial detection of skin are implemented by the RAVL class.

The class `PointMatchC` represents the correlation function in the DSI (see Section 2.4.2). It contains the coordinates of the pixel which it belongs to. It is capable to store the set of the pairs - disparity and matching score. The disparity values can be arbitrarily situated on the disparity axis. It supports only integer disparity values. The methods for the detection of local maxima and their sorting according a height are provided.

The class `DisparitySpaceC` is a representation of the disparity space (see Sections 2.4.2, 3.5). The important properties are the rectified pair of images (the reference and matching image) and the equivalent pair of binary masks with face regions. It also contains the DSI which is constructed from the mentioned images. The DSI is built as an array of the instances of the class `PointMatchC`. The stored set of pairs (disparity - matching score) in every instance is continual sequence within the set disparity range. The extent of DSI can be limited by the given rectangle within the reference image. Using the disparity range and the region in the reference image the arbitrary part of disparity space with the cuboidal volume can be represented by the object of this class. It gives an opportunity to cover only interesting parts of full disparity space by the group of these objects. It saves a memory and overall time of DSI construction. The stored DSI is defined only for the pixels sampled by the scanning step s_s in the reference image. The sampling is done with respect to the upper left corner of the reference image. Even when the object is limited to the arbitrary part of full disparity space, the sampling is done with respect to the global coordinate system of disparity space. It preserves the consistency of the sampling in more objects in different parts of full disparity space. The embedded DSI can be computed in two ways. The classical way computes individually the n_{CC} for each pair of matching windows. The optimised way uses the box-filtering technique.

The class `DisparityMapC` is a representation of the DM (see Sections 2.4.2, 3.6). It supports the floating-point disparity values. The first reason is the post-processing of DM which produces real values. The second reason is easy extension to the methods of the construction of DM which calculate the disparity on the sub-pixel level. Similarly to the class `DisparitySpaceC` the object can store only a part of DM corresponding to the rectangular region in the reference image. Again the combination of objects can define the DM only for interesting parts of the reference image. The main property is the scanning step s_s. The sampling is solved by the approach used the class `DisparitySpaceC`. It is global with respect to the reference image in full resolution despite a restriction by rectangular region. Similar design with the class `DisparitySpaceC` simplifies the usage of these classes by the stereo correspondence methods. The object of `DisparitySpaceC` can be seen as an input and the output of these methods is the object of `DisparityMapC`. Except being advanced data structure the class `DisparityMapC` implements the post-processing methods - hole filling, sub-pixel disparity refinement and Gaussian smoothing (see Section 3.6.4).

The class `StereoCorrespondenceFinderC` is dedicated to the stereo correspondence search (see Sections 3.5, 3.6). It integrates several phases - the construction of disparity space, the construction of DM and post-processing DM. The inputs are the rectified images and the binary masks with face regions and the result is final DM from which the surface can be reconstructed. The main properties are the object of class `DisparitySpaceC` and the object of class `DisparityMapC`. Therefore, the part of tasks included in the stereo correspondence search is internally performed by these subobjects. The functionality implemented directly in this class is connected with the construction of DM. All three approaches proposed in the design chapter are included in this class. The call of method runs the chosen approach using the set parameters. At first, the DSI is computed by the object of class `DisparitySpaceC`. The local technique examines the correlation functions in the DSI using the functionality of class `PointMatchC`. The disparity of global maximum is recorded for the pixels with strong correspondence into the object of class `DisparityMapC`. The remaining instances of `PointMatchC` are copied into a dynamic array which is repeatedly traversed until the whole DM is resolved. The global technique passes through the DSI and constructs the graph using the library MAXFLOW. The library provides the methods for the maximum flow computation and the detection whether a node belong to the source set or sink set. The different classification of adjacent graph nodes in the direction on the disparity axis indicates the resulting disparity value for the pixel. It is recorded on the equivalent position in the object of class `DisparityMapC`. The global approach with estimate uses the methods for the local technique to obtain the estimate of DM. The additional objects of class `DisparitySpaceC` and `DisparityMapC` are temporarily created. The volume of interest is described by two 2D arrays with disparity values which have same size as the desired DM. Another two arrays define how the graph has to be wrapped by auxiliary nodes. The graph is constructed according to these 4 arrays. The computation of DM is same as in the global technique. The object of class `DisparityMapC` performs the Gaussian smoothing at the end of all methods. The computation of DM can be reduced to the rectangle in the reference image as well.

The class `SurfaceReconstructorC` is dedicated to the surface reconstruction (see Section 3.7). It works with the object of class `DisparityMapC`. Other important properties are the reference image and the rectified projection matrices. The 3D model is reconstructed from the DM as a triangle mesh. The object of `DisparityMapC` also provides the texture coordinates for the mesh pointing to the reference image. The 3D model can be saved into VRML format what is alowed by the input-output system in the RAVL.

4.4.2 Programs

The captured images are processed by the group of command-line programs and one simple GUI application. They are based on the library `StereoFaceCapture`. The tasks from the image pre-processing are realised by standalone programs because they are quite independent and can be used separately. The stereo matching and surface reconstruction are integrated to one program. The input for this set of applications is the pair of images obtained by the capture rig and the calibration data of both cameras computed by the Matlab routine mentioned in the previous section. The data flow between the programs consists of the images, text files and binary files. The images transferred between the programs can be in the arbitrary format supported by the RAVL (e.g. PPM, PGM, JPEG, TIFF, PNG).

The program Undistortion performs the removal of lens distortion from the image. The inputs are the image and the text file with the calibration data of camera which has taken that image. The output is the undistorted image in the same format as the input image.

The program Rectification rectifies the image pair. The precondition is that the images does not contain the lens distortion. The second input are the calibration data of both cameras. The radial and tangential distortion coefficients are ignored. The main output is the rectified pair of images in the same format as input images. The supplementary binary masks defining the relevant area in each image are another result. The masks are stored as black and white images in the PGM format. The white colour represents the relevant area. The third outcome is binary file for each image which contains the data describing the performed transformation. For instance, the file for left rectified image contains the matrices H'_L, Q_L and the principal point coordinates u'_{0L}, v'_0 (see Section 3.4.2).

The program FaceExtraction performs the extraction of face region in the image. The input is the image with the face. The parameter *tol*, the size of closing operator and the size of erosion operator are set through the program arguments (see the class FaceExtractorC). The result is the binary mask defining the face region in the input image. The mask is saved as black and white image in the PGM format. If the rectified image is processed, the optional input is the binary mask from the rectification. The face region is then searched only in the relevant area of the image.

The program DisparityRange is the simple GUI application. It is implemented using the extension for GUI in the RAVL. The purpose is to roughly find the nearest and the farthest point of the face in the images. The rectified images are an input. The user chooses the approximate position of the nearest and the farthest point in both images. The horizontal coordinates of these points in both images are saved into the text file. The disparity range for the stereo correspondence search is calculated from these values.

The program Stereo is the main application which creates the 3D model from the images. The first input is the rectified image pair. The second input is the binary masks with the face region for each rectified image. The third input is the binary files with the data about the rectification for each image. The fourth input is the configuration file with all parameters necessary for the processing. The set of used parameters differs according to the chosen stereo correspondence method. The optional fifth input is the text file from the application DisparityRange. The user has an option to set the disparity range directly through the configuration file. The output is the textured 3D model saved in the VRML format. It is one text file with the description in the VRML language and one image for a texture. The program can optionally visualise the intermediate results from the processing as images and print the status information.

Chapter 5

Results

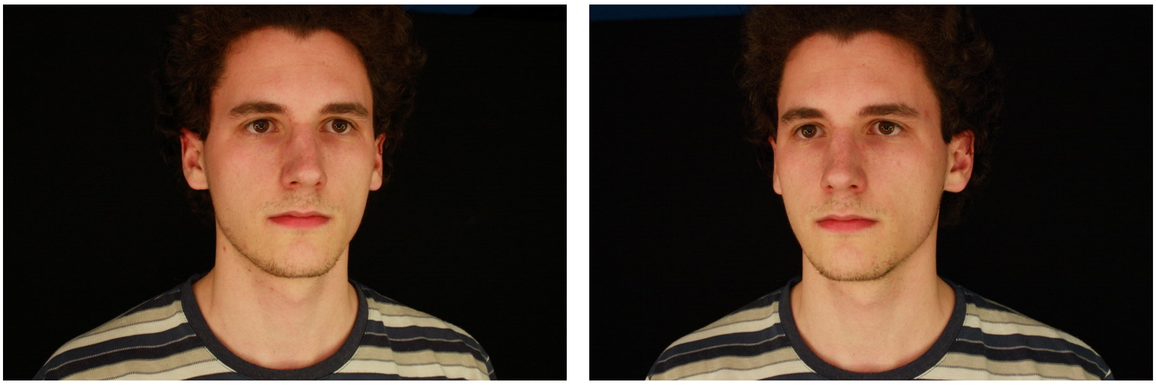
The results by the developed 3D face capture system are presented in this chapter. Except the final 3D models of face the intermediate outputs produced by individual processing stages are shown. The evaluation is focused on the stereo correspondence search in the image pair. The three implemented approaches to this problem are compared in terms of a computational time, memory consumption and accuracy of resulting 3D model.

5.1 Test conditions

The results were obtained by processing particular stereo pair of face images. The images were captured by the capture rig described in Section 4.1. All distances depicted in the scheme were adhered. The first variant of light setup illustrated in Figure 4.2(left) was used. The digital still cameras were calibrated by the procedure explained in Section 4.3. The acquisition of images was performed according to the recommendations in Section 4.2. They were processed by the software described in Section 4.4. The following hardware configuration was used for the processing images: 8 × Dual Core 3GHz P4 Xeon EM64T, 32GB RAM. Only one processor was computing the tasks because all programs are single-threaded. The computational times presented later in this chapter might be influenced by the other processes on the server. It was not possible to reserve the server exclusively for these tests. However, the tests were performed during the minimal load of server so the times should be close to precise measurements.

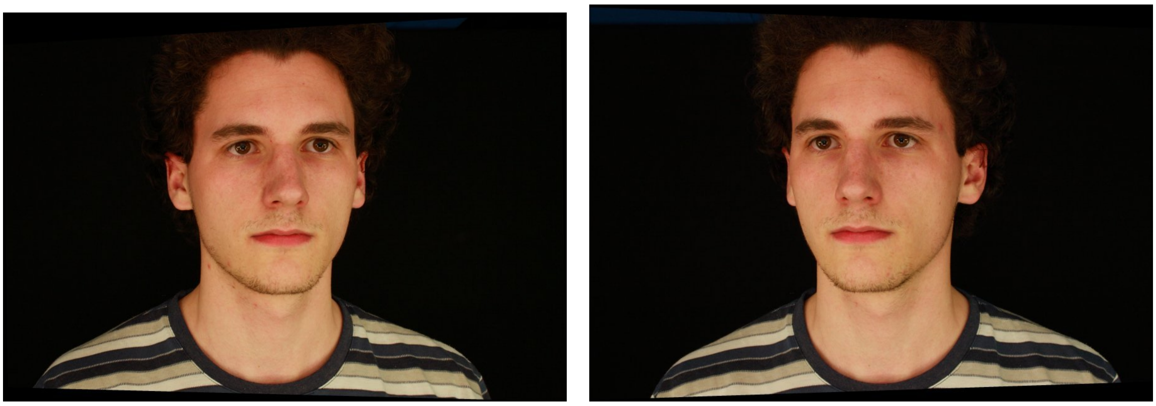
5.2 Image pre-processing

The input images of the face are shown in Figure 5.1. Both of them have a resolution 3504×2336 pixels. The images are pre-processed at first. They do not change noticeably during the removal of lens distortion because the used digital cameras do not suffer from severe radial and tangential lens distortion. Therefore, their versions after image undistortion are not showed. The undistortion takes $4.31s$ for each picture. The image pair after the rectification is displayed in Figure 5.2. Both original images are transformed to the areas with a shape of irregular quadrangle which are embedded inside the rectified images (background outside the areas has black colour). The transformation is not strong because the camera coordinate systems are in almost same height and they are rotated with respect to each other by small angle. New resolution of the left rectified image is 3653×2520 pixels and the resolution of the right rectified image is 3578×2520 pixels. The important fact is that the height of pictures is same and the face features have same vertical position in both of them. The time of rectification is $26.49s$. Finally, the face regions are extracted from the rectified pictures. The model of skin colour is estimated from the 25×25 window in the centre of each image. The classification of pixels as a skin is done according to the parameter $tol = 70.0$. The mask is finalised by the morphological closing operator with the size 70×70 pixels. Large operator is necessary for including large non-skin regions such as eyes into the face mask. The computational time is $2m55.53s$ for each picture. The original masks contained also large pieces of T-shirt because simple segmentation technique does not distinguish between a skin and bright stripes on the T-shirt. These regions were manually erased to provide precise computational demands of following processing related to the face. The used binary masks defining the regions are depicted in Figure 5.3.



(a) (b)

Figure 5.1: The input image from the left (a), the input image from the right (b).



(a) (b)

Figure 5.2: The left rectified image - the reference image (a), the right rectified image - the matching image (b).

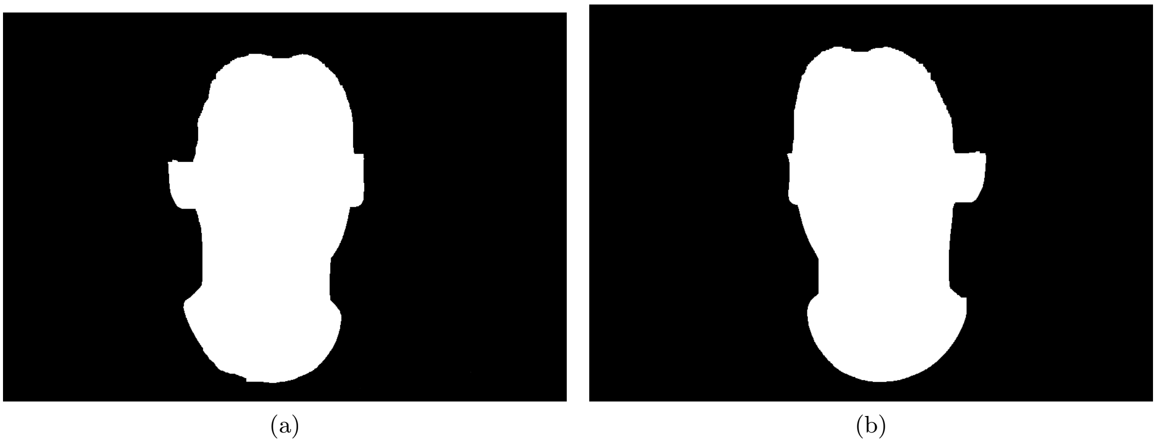


Figure 5.3: The binary mask of face region in the reference image (a) and matching image (b).

5.3 Parameters and intermediate results of techniques

The left rectified image is the reference image for the stereo correspondence search. The matching is searched for the sampled grid of pixels. The sampling is determined by the scanning step $s_s = 4$ pixels. The disparity range is fitted to the volume of face. The maximal disparity value 1785 represents the depth layer approximately $0.5cm$ in front of the tip of nose. The minimal disparity value 1400 represents the depth layer approximately $0.5cm$ behind the edge of ears. The average distance between the depth layers defined by 1-pixel steps of disparity is round $0.5mm$. Together 386 depth layers slice the volume of face and provide a place for situating 3D points of a model.

The local approach to the construction of DM creates the DSI using the matching window with size 31×31 pixels. The slice through the DSI along the row in the area of forehead is showed in Figure 5.4(a). Only the part of slice within the face region in the reference image is depicted. The axis U points in the horizontal direction from left to right and the axis D in the vertical direction from top to bottom. The minimal disparity value is associated with the top row of image. The values of DSI image are rendered in grey scale whereas minimal n_{CC} is mapped on black colour and maximal on white colour. Red colour marks the areas with undefined DSI. The response of face within the DSI is visible as a bright ridge. The pixels with the strong correspondence are chosen by the matching score and ratio threshold set to the means of corresponding characteristics. Specifically, $t_s = 0.608745$ (range -1..1) and $t_r = 0.832531$ (range 0..1) for this pair of images. The pixels with strong correspondence are 15.57% of all processed pixels. The maximal disparity threshold t_d is 3 pixels. The resulting DM is showed in Figure 5.5(a). The maximal disparity value is mapped on white colour and minimal value on black colour. Unprocessed background defined by the mask in Figure 5.3(a). The slice through DM is illustrated in the context of DSI by yellow curve in Figure 5.4(a).

The global approach creates the DSI using the matching window with size 11×11 pixels. The slice through the DSI is showed in Figure 5.4(b). It can be seen that the DSI is more rugged with decreasing size of the matching window. The responce of the face inside is less recognisable as well. The smoothness coefficient λ is set on the value 0.025. The resulting DM is showed in Figure 5.5(b). The slice through DM is illustrated in Figure 5.4(b).

The global approach with the estimate by local technique has same parameters as standalone local and global method. The estimate of DM is modified by the hole filling with 5×5 operator. It is depicted in Figure 5.5(c). The volume of interest is modelled by the offset $o_l = 10$ and the expansion region with the size 15×15. Figure 5.4(c) illustrates the slice through DSI with some additional information specific for this technique. The green curve represents initial estimate of DM by the local method. The course of final DM is present as well. The minimal disparity boundary of the volume of interest is marked by cyan colour and maximal boundary by magenta colour. The reduced graph is constructed only for the part of DSI between these two curves. It can be seen that the graph has complex shape for human face. The resulting DM is showed in Figure 5.5(d).

The resulting DM by each technique is post-processed. This procedure involves only smoothing by Gaussian operator. The size of operator is 13×13 and standard deviation of Gauss function is $\sigma = 3.0$. Finally, the 3D model is reconstructed from each DM.

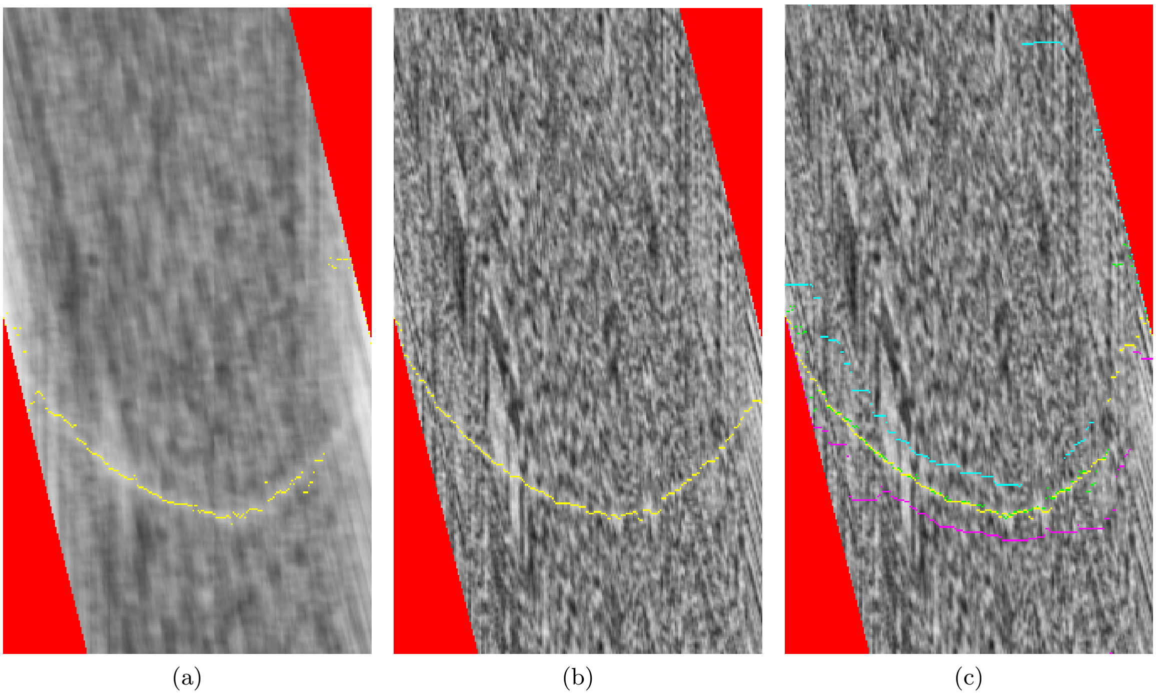
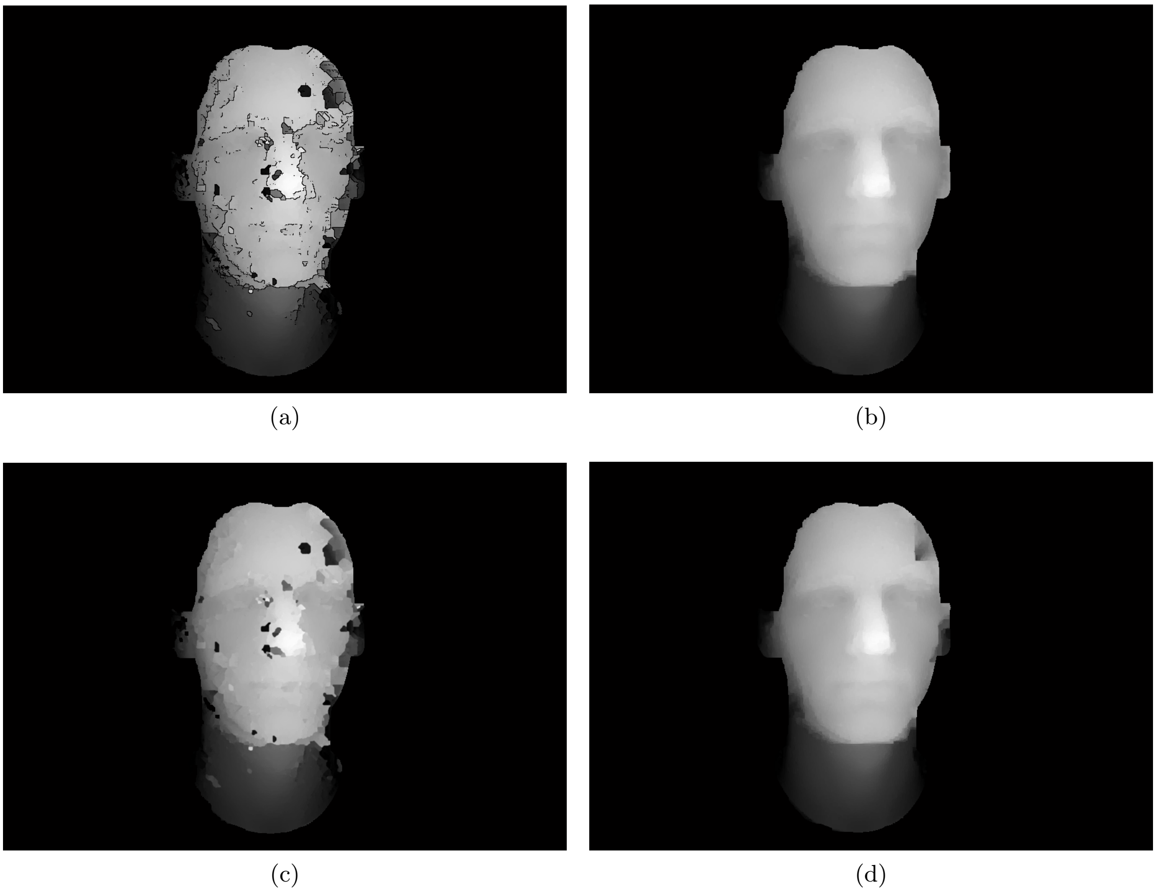

(a) (b) (c)

Figure 5.4: The slice through DSI for the local approach (a), global approach (b) and global approach with an estimate by local technique (c). The DM together with the resulting DM are visualised.

5.4 Comparison of techniques

The 3D models and measured computational demands are obtained using the parameters of individual techniques mentioned in previous section. Table 5.1 sumarises the times of important stages of processing for each approach. The image pre-processing is same for all techniques. Total time comprising the image undistortion, rectification and face extraction is shown. The construction of DSI takes very similar time for every method because



(a) (b) (c) (d)

Figure 5.5: The *DM* by the local approach (a), global approach (b) and global approach with an estimate by local technique (d). The initial *DM* for the global approach with an estimate by local technique (c).

the computation is almost independent on the size of matching window. The 1 second difference between the local approach and the others is caused by 31 × 31 window against 11 × 11 windows. The most interesting time is the construction of *DM*. It illustrates the computational complexity of each approach to the creation of *DM*. The local technique is clear winner. The purely global approach takes the longest time as expected. The global approach with the estimate brings significant improvement over the global method. It is important to mention that the time for global technique with estimate includes the computation of initial *DM* by local method. The time of this computation is similar as the sum of *DSI* and *DM* construction in the original local technique. The time of smoothing *DM* is negligible in comparison to the rest of stages, therefore it is not included in Table 5.1. The reconstruction of 3D surface from the *DM* does not differ much among the methods in terms of a time. The numbers of the vertices and triangles in the 3D model created by each approach are written in the last two rows of Table 5.1. The numbers are smaller for the local technique because the *DM* contains the 'worm-like holes' (see Figure 5.5(a)). The surface is not reconstructed from these areas so the time is lower than in other methods. The total time comprises all stages including the image pre-processing. Finally, Table 5.1 contains also the maximal value of memory consumption during the whole processing. The significant consumers of memory are the images and the data structure for DSI (same for all techniques). In the methods using graph cut the data structure for the graph is responsible for the majority of memory requirements. Moreover, the global method with estimate has additional DSI.

	local	global	global with estimate
pre-processing	$6m26.17s$		
construction of DSI	$1m39s$	$1m38s$	$1m38s$
construction of DM	$0m5s$	$68m38s$	$17m4s$
surface reconstruction	$9m7s$	$11m13s$	$11m21s$
total	$17m17.17s$	$87m55.17s$	$36m29.17s$
memory	$0GB992MB$	$16GB448MB$	$4GB0MB$
vertices	113015	121579	121579
triangles	214266	241720	241720

Table 5.1: The comparison of techniques - the computational times of individual stages, memory consumption and the properties of 3D model.

It is interesting to compare the purely global approach and the global approach with estimate in detail. As it was mentioned before, the tradeoff between the accuracy of DM and the computational demands exists for the global technique with estimate. The size of the volume of interest (therefore the size of graph) is set by the parameters $o_l = 10$ and $w_{er} = 7$. These values were chosen to demonstrate how the computational demands can be reduced with the acceptable loss of accuracy in comparison to purely global approach. The Table 5.2 lists the number of nodes and edges in the graph constructed by each technique. The graph is significantly reduced by using the estimate what leads to 75.68% reduction of overall memory consumption (see Table 5.1). Also the computation of DM is 4.02 times shorter. The absolute difference between DM produced by each technique (see Figure 5.5(b),(d)) is visualised by in Figure 5.6. The maximal difference of disparity is marked by black colour and zero difference by white colour (a background is black). The DMs are compared before the post-processing. Maximal difference is 199 disparity levels but 95.74% of DM is the same. The energy of the solution according to Equation 3.15 is increased by 5.88% (see Table 5.2). The solutions differ only for the right side of a face where the estimate by local technique (see Figure 5.5(c)) contains strong outlying regions. It may be concluded that the computational demands are significantly lowered and only small loss of precision is present.

	global	global with estimate
nodes	47051073	10277033
edges	280879848	60358142
energy	34893.9	36946.0

Table 5.2: The comparison of purely global approach and the global approach with the estimate - the number of nodes and edges in the graph and final energy value.

The visual quality of 3D face models reconstructed by individual techniques is illustrated in Figures 5.7, 5.8, 5.8. Each model is visualised in the shaded variant without a texture and

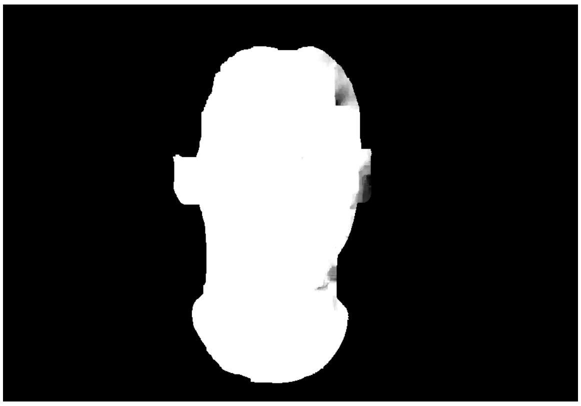


Figure 5.6: The absolute difference between the DM constructed by the purely global approach and global approach with estimate.

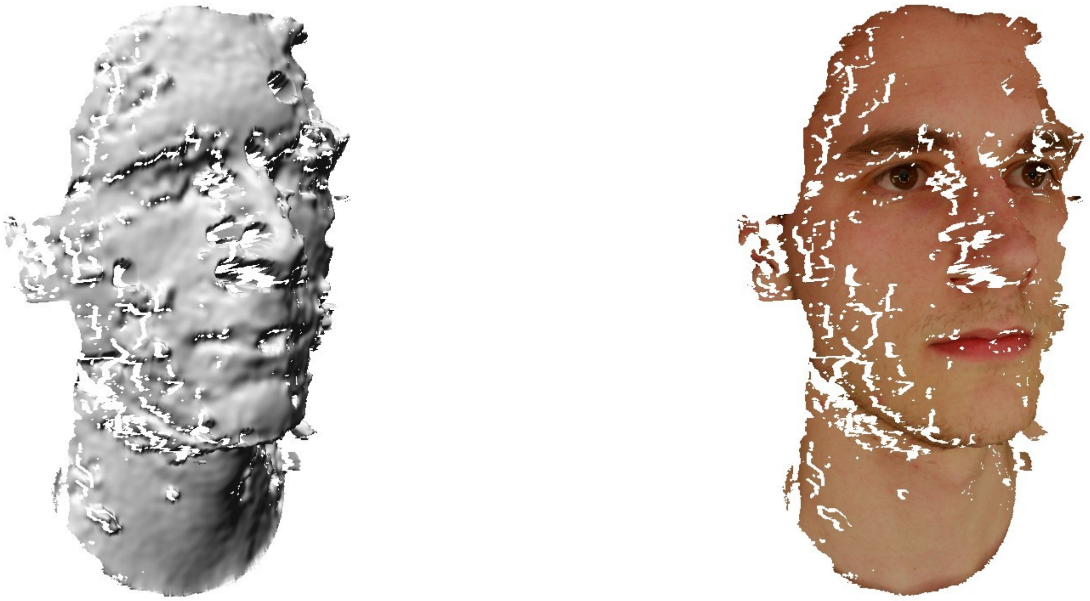
in the textured variant. The shaded model gives some insight of the actual shape of surface and the textured variant demonstrates final result of 3D face capture. The accuracy of the model is assessed by the subjective visual comparison with real face because no ground truth data were available.

The local technique has the least computational complexity but the model of face has the least quality (see Figure 5.7). The rather 'flat' fronto-parallel areas such as a forehead or the parts under the eyes are the best reconstructed. The estimate of accuracy is $\pm 2mm$ from real face. The slanted parts such as the sides of nose or cheeks contain holes. The outlying regions from the DM are transformed to the separate patches which are disconnected from the main surface. The 'skirts' on their edges and the edges of holes are caused by the Gaussian smoothing DM. The facial details such as eyes or lips are coarse with small holes in the places of rapid change of surface. The imperfections of model considerably deteriorate its overall look.

The model by the global technique has far better quality in comparison to the local technique (see Figure 5.8). The surface of the face is smooth and continuous without holes or steps. The main improvement is visible in the slanted areas. The global optimisation of DM is able to cope with the ambiguity of DSI for these areas. The details of eyes or lips are finer than in the local technique. The cavity in the eye is a consequence of the transparency of a cornea. The central part of face differs from the real face round $\pm 2mm$. The eyebrows and the nose are on the margin for this precision. The sides of nose are too flat. The appearance of eyebrows is a bit noisy. The matching is complicated for the eyebrows because they look different in each image. Only areas around the ears are completely incorrect. It is understandable because they are almost occluded in one of the images. The flat transition between the chin and neck is present because this part of face is not captured at all. The overall look of the textured model is realistic.

The model by the global technique with estimate is very similar to the model by the purely global approach (see Figure 5.9). The difference is noticeable on the right side of the face. The ear is differently reconstructed, however it is not correct even in the global technique. The actual loss of accuracy is a cavity on the right side of forehead.

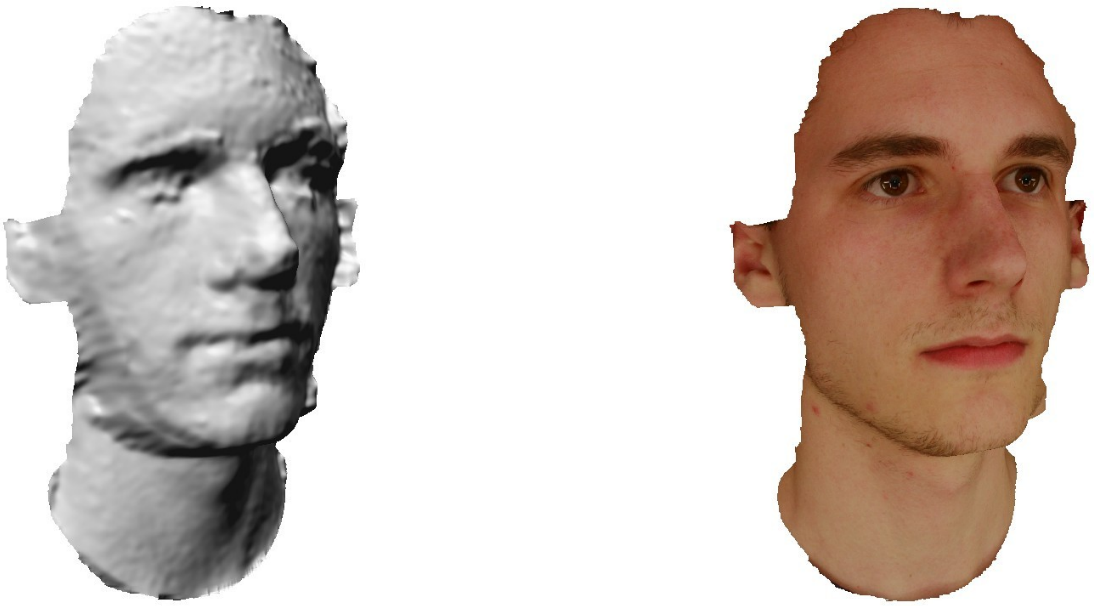
All models suffer from small bumps even in the well reconstructed areas. It is a residual from the local bumpiness in the raw DMs which is partially eliminated by Gaussian smoothing. The large imprecisions on the sides of the face model by the global techniques are not serious for practical use. These part of 3D model can be replaced by the surface reconstructed from another stereo image pair covering the side of face.



(a)

(b)

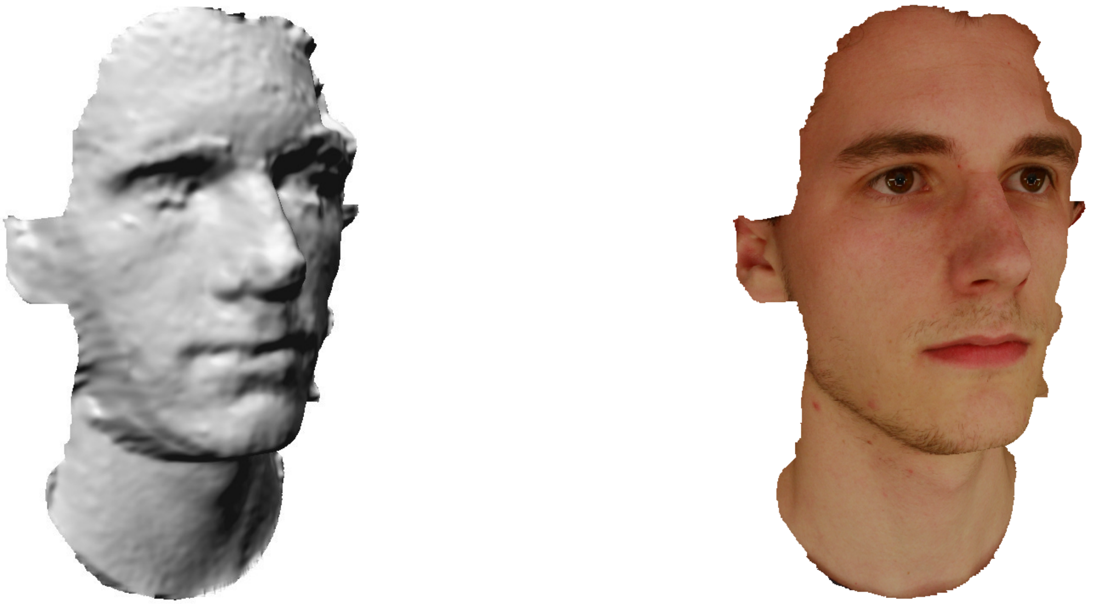
Figure 5.7: The 3D face model reconstructed by the local approach - shaded variant (a), textured variant (b).



(a)

(b)

Figure 5.8: The 3D face model reconstructed by the global approach - shaded variant (a), textured variant (b).



(a)

(b)

Figure 5.9: The 3D face model reconstructed by the global approach with the estimate by local technique - shaded variant (a), textured variant (b).

Chapter 6

Conclusion

The objectives of the thesis were fulfilled. The theoretical background about the stereo matching images and generally 3D capture systems was gained (see Chapter 2). Own 3D face capture system was designed and appropriate algorithms were chosen for each stage of processing (see Chapter 3). Several different approaches to stereo matching face images were proposed. Complete chain from the image acquisition to the 3D model reconstruction was implemented regarding hardware and software side (see Chapter 4). The results produced by the system were evaluated and discussed (see Chapter 5).

The developed 3D face capture system is based on the passive stereo photogrammetry. Own capture rig allows simultaneous acquisition of face shape and appearance. The images are taken under uniform illumination of whole face by one stereo pair of digital still cameras with high resolution (8.2 megapixel). The capturing is quick and comfortable for a person. The cameras are calibrated to acquire the transformation between a 3D point and its projection in an image. The calibration data are necessary for the image pre-processing and surface reconstruction. The image pre-processing consists of the image undistortion, rectification of image pair and face region extraction. It simplifies the consequent stereo correspondence search. The stereo matching is performed by three different methods. The face surface is reconstructed according to the found correspondence between images as textured triangle mesh. The final 3D model covers only the face with part of neck (hair are not included).

Main contribution of this thesis is an exploration of three different approaches to the stereo correspondence search in the case of face reconstruction. Local approach is an extension of traditional local techniques. It is more robust in the case of rugged correlation functions in a disparity space. Global approach is a known global technique which optimises the disparity map with respect to the defined energy function. The energy is minimised by a graph cut. The third approach is proposed as the combination of previous two methods. The estimate by local technique is exploited for the reduction of search space in following global optimisation by the graph cut. All mentioned methods can be used for the reconstruction of arbitrary object, however they exploit some constraints which are given by the shape of human face.

The local approach is very fast in comparison to other two methods but the model of face has the least quality. The slanted parts of face with respect to the camera pair contain many holes and outlying patches of surface (e.g. sides of face or nose). The facial details such as eyes or lips are coarse. The resulting 3D model is not suitable for practical use. However, the disparity map constructed by this technique provides reasonable estimate for the hybrid method.

The global approach produces the 3D model with good quality. The surface of face is smooth and continuous without holes or steps. The facial details (e.g. eyes or lips) are finer than by the local technique. The central part of face differs from the real face round $\pm 2mm$ according to subjective assessment. The sides of face are not correctly reconstructed. The 3D model looks realistic. Main disadvantage is the long computational time and huge memory consumption especially when the high-resolution images are processed.

The global approach with the estimate by local technique creates similar 3D model like purely global technique. A noticeable degradation of the model is caused by the imperfections in the estimate of disparity map. The differences appear on the sides of face. The tradeoff between the precision of 3D model and the computational demands exists in this method. However, the computational time and memory consumption can be significantly reduced with the modest loss of accuracy. This method is the most promising for practical use.

The results proved that the human skin provides enough details in high-resolution images for precise correspondence search. The sides of the face are not correctly reconstructed because one stereoscopic camera pair with short baseline cannot properly cover whole face. In practice it can be easily solved by an addition of another camera pairs pointing on these areas.

The system can be improved or extended in several ways. The main limitation in terms of model quality is an ambiguity of disparity space image. Correlation functions in slanted areas could be improved by the adaptation of matching windows with respect to local orientation of the surface. It is also desirable to cope with a noise present in the functions of matching score which cause local bumpiness of the surface. Explicit smoothing a disparity map could then be omitted. But post-processing the disparity space image has to preserve the present information, especially about small facial details. Another possibility is a change of the computation of matching score which would not generate the rugged disparity space image. Addressing the computational demands, it is worth to try a hierarchical approach to the construction of disparity map using multiple graph cuts. The comparison with proposed technique exploiting the estimate by fast local method would be interesting. A drawback of all described stereo matching techniques is their dependency on many parameters which have to be adjusted manually for different capture configurations in order to gain the best results. An automatic setup of these parameters from general informations about a capture configuration would ease a practical usage. Finally, several camera pairs could be added to the capture rig to cover properly whole face. The surfaces created from each image pair can be merged into single 3D model. As a result, larger portion of object would be correctly modelled.

Bibliography

[1] Canon eos 30d. [online]. [cit. 2008-05-13]. URL: http:///web.canon.jp/Imaging/eos30d/index.html.

[2] D-link dub-h4. [online]. [cit. 2008-05-13]. URL: http://www.dlink.com/products/?pid=152.

[3] Decoding raw digital photos in linux. [online]. [cit. 2008-05-13]. URL: http://cybercom.net/~dcoffin/dcraw/.

[4] gphoto digital camera software. [online]. [cit. 2008-05-13]. URL: http://www.gphoto.org.

[5] Maxflow library. [online]. [cit. 2008-05-15]. URL: http://www.adastral.ucl.ac.uk/~vladkolm/software.html.

[6] Recognition and vision library. [online]. [cit. 2008-05-15]. URL: http://www.ee.surrey.ac.uk/Research/VSSP/Ravl/.

[7] Virtual reality modeling language. [online]. [cit. 2008-04-29]. URL: http://www.web3d.org/x3d/specifications/vrml/.

[8] J.-Y. Bouguet. *Visual methods for three-dimensional modeling*. PhD thesis, California Institute of Technology, Pasadena, California, USA, 1999.

[9] J.-Y. Bouguet. Camera calibration toolbox for matlab: www.vision.caltech.edu/bouguetj/calib-doc. Technical report, MRL-INTEL, 2003.

[10] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. In *IEEE Transactions on PAMI*, volume 29, pages 1124–1137, September 2004.

[11] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice-Hall, New Jersey, USA, 2003.

[12] A. Fusiello, E. Trucco, and A. Verri. A compact algorithm for rectification of stereo pairs. In *Machine Vision and Applications*, pages 16–22, March 2000.

[13] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, United Kingdom, 2000.

[14] J. Heikkila and O. Silven. A four-step camera calibration procedure with implicit image correction. In *IEEE Proceedings of the Conference on Computer Vision and PatternRecognition (CVPR '97)*, page 1106, 1997.

[15] A. Hilton, A.J. Stoddart, J. Illingworth, and T. Windeatt. Implicit surface-based geometric fusion. *CVIU*, 69(3):273–291, March 1998.

[16] H. Ishikawa and D. Geiger. Occlusions, discontinuities, and epipolar lines in stereo. In *The Fifth European Conference on Computer Vision*, June 1998.

[17] Martin Klaudíny. Optické skenovanie 3d objektov, kalibrácia systému. Bachelor's thesis, Vysoké učení technické v Brně, 2006.

[18] V. Kolmogorov. *Graph based algorithms for scene reconstruction from two and more views.* PhD thesis, Cornell University, January 2004.

[19] V. Kolmogorov and C. Rother. Comparison of energy minimization algorithms for highly connected graphs. In *ECCV (2)*, pages 1–15, 2006.

[20] J. Migdal. Depth perception using a trinocluar camera setup and sub-pixel image correlation algorithm. Technical report, Mitsubishi Electric Research Laboratories, Cambridge Research Center, May 2000.

[21] R. Owens. Mathematical morphology. [online], 1997. [cit. 2008-04-29]. URL: http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/OWENS/LECT3/node3.html.

[22] N. Paragios, Y. Chen, and O. Faugeras, editors. *Handbook of Mathematical Models in Computer Vision.* Springer, 2006.

[23] S. Roy. Stereo without epipolar lines: A maximum-flow formulation. *International Journal of Computer Vision*, 34(2/3):147–161, 1999.

[24] S. Roy and I. J. Cox. A maximum-flow formulation of the n-camera stereo correspondence problem. In *ICCV*, pages 492–502, January 1998.

[25] Sébastien Roy and Marc-Antoine Drouin. Non-uniform hierarchical pyramid stereo for large images. In *VMV*, pages 403–410, 2002.

[26] S. Rusinkiewicz, O. Hall-Holt, and M. Levoy. Real-time 3d model acquisition. In *ACM Transactions on Graphics (SIGGRAPH)*, pages 438–446, July 2002.

[27] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Technical Report MSR-TR-2001-81, Microsoft Research, November 2001.

[28] C. Sun. A fast stereo matching method. In *Digital Image Computing: Techniques and Applications*, pages 95–100, Auckland, New Zealand, December 1997. Massey University.

[29] Ch. Sun. Fast stereo matching using rectangular subregioning and 3d maximum-surface techniques. *International Journal of Computer Vision*, 47(1/2/3):99–117, May 2002.

[30] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. F. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields. In *ECCV (2)*, pages 16–29, 2006.

[31] O. Veksler. *Efficient graph-based energy minimization methods in computer vision*. PhD thesis, Cornell University, August 1999.

[32] O. Veksler. Reducing search space for stereo correspondence with graph cuts. In *BMVC06*, page II:709, 2006.

[33] E. W. Weisstein. "euler angles" from mathworld – a wolfram web resource. [online]. [cit. 2008-04-29]. URL: http://mathworld.wolfram.com/EulerAngles.html.

[34] I. A. Ypsilos. *Capture and Modelling of 3D Face Dynamics*. PhD thesis, CVSSP, University of Surrey, United Kingdom, September 2004.

[35] Z. Zhang. A flexible new technique for camera calibration. Technical Report MSR-TR-98-71, Microsoft Research, December 1998.