Dynamic 3D Scene Analysis through Synthesis Feedback Control

Reinhard Koch

Abstract **–The analysis of 3D scenes consisting of nonrigid moving objects from 2D image sequences is discussed. A parametric description of dynamic objects is extracted and the time variant scene parameters are estimated throughout the sequence by employing an analysis by synthesis approach. From the parametric scene description images are synthesized and compared with the original input images of the camera. Frame differences between both images are evaluated to estimate a scene parameter update. The analysis system is applied to video phone scenes as a data compression algorithm where the scene parameters are transmitted and the output sequence is synthesized at the receiver.**

Index Terms **– 3D scene analysis, dynamic scene analysis, image sequence analysis, 3D motion estimation, 3D object tracking, 3D shape reconstruction, 3D shape adaptation, analysis by synthesis, object–oriented video phone coding.**

I. Introduction

Modelling of 3D scenes from 2D image sequences has been a research topic for a long time as Aggarwal and Nandhakumar [1] showed in their overview of this field. The goal of such modelling is to extract a compact description of the scene for purposes of reconstruction [2], recognition [3], or data compression [4]–[6]. When analyzing complex scenes with multiple moving flexible objects a complete description of all properties of the scene is necessary. In previous works the different properties 3D object shape, 3D object motion, and object surface were treated separately. Great effort went into developing algorithms that estimate 3D object shape from various sources, termed shape from motion, stereo, texture, and others. [7]–[10]. On the other hand research was conducted to find solutions to the problem of rigid object motion [11], [12]. Only recently the problem of dynamic nonrigid bodies and nonrigid motion was addressed [13], [14].

When imposing specific restrictions onto the scene the complexity of the problem can be reduced. One such restriction is the analysis of a moving person in front of static background viewed by a monocular camera. This scenario is relevant for the application of video telephone systems and researchers have focussed on this type of scene for some time. Special attention was given to the treatment of human faces and several models were developed to describe facial mimic. Rydfalk [15] described the facial model CANDIDE, a parametric 3D facial mask with the capability to exhibit facial mimic actions based on the FACS model of facial actions introduced by Ekman and Friesen [16]. Several researchers adopted this model to perform 3D motion compensation of human faces [17]–[19] and to estimate the facial mimic of a speaking person [5]. A similar approach was carried out by Aizawa et al. [20] for the model–based analysis of a human face. These works, however, focus on special aspects of the problem, like 3D motion compensation and image synthesis. They neglect, for example, the issues of initial scene segmentation, automatic positioning of their facial mask, proper treatment of the shoulder part etc. No complete approach to a fully automated image analysis system for video phone scenes has been discussed so far.

The image analysis and image coding group at the University of Hannover has been working on all aspects of automatic modelling of human head – shoulder parts in video phone scenes for several years [4], [21]. From the analysis of a monocular image sequence a compact parametric description of moving three dimensional bodies is extracted [22], [23]. The scene parameters are transmitted and a scene model is constructed at the receiver, where a synthesized image sequence is computed and displayed to an observer [24].

In this contribution an approach to automatically extract 3D object shape and 3D object motion from a video sequence is presented.The analysis system adapts a model scene towards the real scene by comparing the visual reconstruction of the model scene with the real input sequence. The model adaptation is guided by measurable differences between the synthesized model reconstruction and the real camera input sequence. An analysis system following such a principle is called **analysis by synthesis control**. Based on physical models of the scene, scene parameters are adapted until the image difference between real input sequence and synthesized model sequence is minimized. In this approach key ideas from the distinct fields of computer aided design, computer graphics, computer vision, and control theory are merged to yield the system presented. As one application the modelling of video phone scenes is chosen but the approach is capable to work for a wide variety of different tasks as discussed in various contributions [25]–[29].

The paper is organized as follows. In Section II the general principles of analysis by synthesis are reviewed. Section III discusses the parameters of the scene model and the measurement model whereas in Section IV different procedures to compute a first 3D model estimate are developed. The parameter estimation for 3D motion and 3D shape based on the analysis by synthesis scheme is derived in detail in Section V. A strategy for the modelling of video phone scenes is discussed in Section VI.

II. Principles of Analysis through Image Synthesis

A dynamic 3D scene consists of time variant 3D objects with each object described by object shape, object position, and object surface reflectance. A light source illuminates the scene viewed by a camera. The projection of the scene onto the camera target at constant time intervals forms a 2D image sequence of the scene.

The goal of image analysis is to determine all parameters of the 3D scene from the sequence of images recorded by the camera. The measurements available are the recorded 2D image intensity distributions of the image sequence. This is an ambiguous problem because an infinite number of possible 3D scenes exist that generate the same 2D image intensity distribution. Direct measurements of the geometrical scene parameters from the image intensity distribution are possible only in special cases where the photometric surface description is known, as assumed in shape from shading or shape from texture measurements. For a general scene methods must be developed to indirectly extract the 3D scene parameters from the 2D image sequence.

A. Scene analysis as parameter estimation problem

The problem of scene analysis can be interpreted as the problem of finding the parameters that describe the 3D scene. Measurements available are the images generated by the 3D to 2D image formation process. It is not possible to directly calculate the 3D scene through inversion of the 3D to 2D projection due to the ambiguity in the image formation process, but the 3D to 2D projection itself can be computed using perspective geometry once the parameters of the scene are known.

This leads to the idea of a model based analysis through synthesis system. The scene is interpreted as a time variant system that can be described by a parametric model whose parameters are unknown and can not be measured directly. Some indirect measurements, however, can be performed transforming the unknown system parameters into accessible measurements through a known measurement transform. The system parameters are then estimated indirectly by modelling the dynamic system and the measurement process. The model scene is projected onto a model camera and compared with the real camera input. The measurable differences between real and modelled images are then minimized through controlled changes of the model parameters. The system needs initialization with a starting parameter set and error minimization is carried out in a feedback loop that calculates the parameter changes directly from the image differences. The combination of model parameters that minimizes the image differences forms the best approximation of the scene.

A system conforming to this approach is shown in Fig. 1. The analysis is separated into four main modules: 3D model scene description, initial model estimation, image synthesis, and feedback parameter estimation.

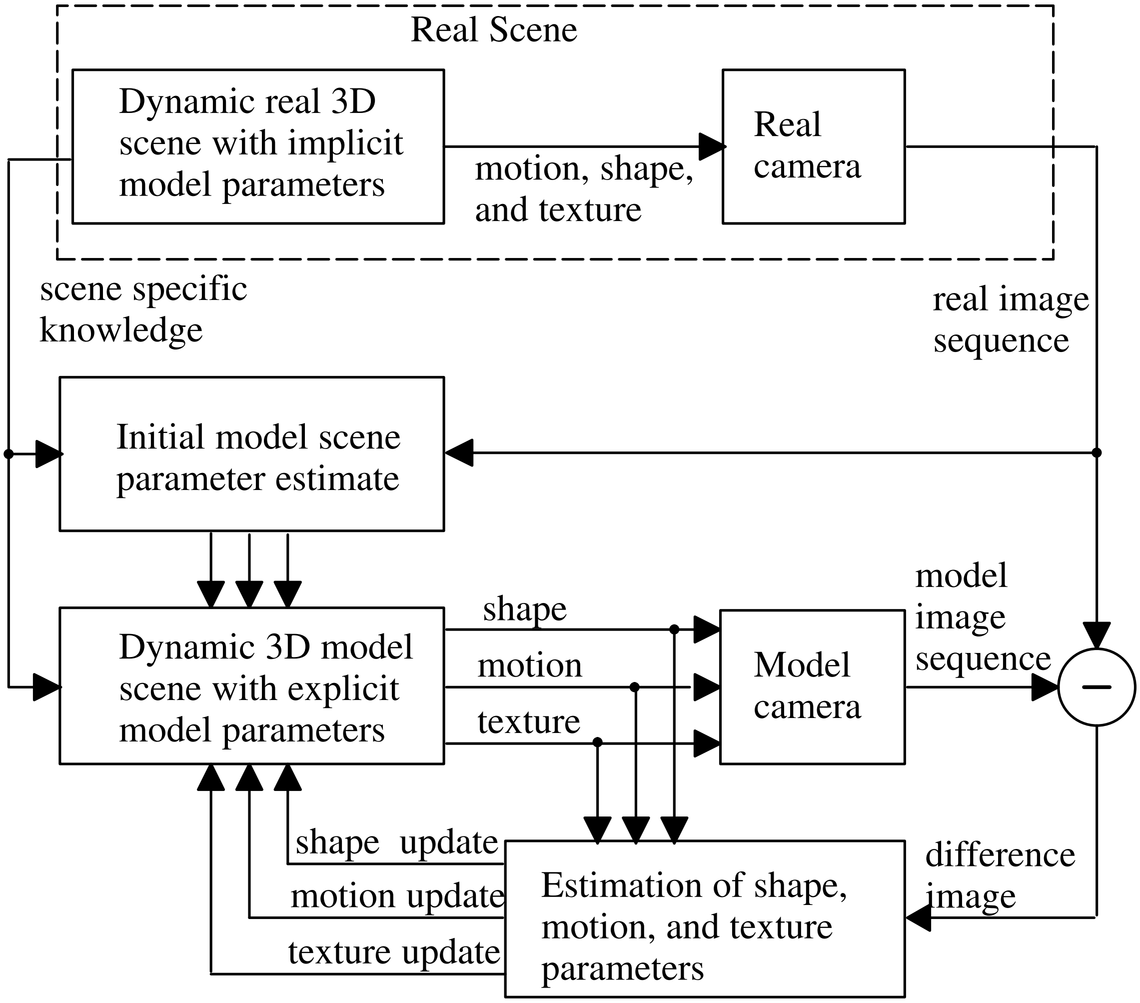


Fig. 1: General structure of model based analysis by synthesis.

– The **model scene** is described parametrically. It approximates the real scene through time variant textured 3D surface models. For this parametric model description surface boundary representations developed for 3D computer aided design techniques are chosen.

- The **initial model estimation** describes a first approximation of the 3D scene. It can be estimated through computer vision algorithms like stereoscopy, shape from various sources, and scene specific knowledge.
- The **image synthesis** is performed through perspective projection of the model scene onto a synthetic camera target. The natural surface colors of the real scene are mapped onto the model surface through surface texture mapping. These textured surfaces are then projected to yield a naturally looking reconstruction of the real image.
- The **feedback parameter control** is used to adapt the model scene parameters. The synthesized image is compared to the real image in a feedback loop which minimizes image differences through parameter estimation from image intensity gradients or through parameter search.

III. Model Scene Description

According to Fig. 1 an approximation of the 3D scene is needed with an explicit description of shape, motion, and surface texture parameters. Geometric properties of an object like surface shape, position in space, and dynamic behavior are modelled through deformable triangular surface meshes while photometric properties like surface reflectance and surface color are stored in texture maps attached to the object. The term 'texture' in this context will be used in the computer graphics sense to describe the photometric color intensity pattern on the object surface.

B. Geometric properties

The following section will give some definitions of scene geometry used and will illustrate these definitions by an example.

- Scene geometry is measured relative to a 3D scene coordinate system. The scene consists of a set of dynamic 3D objects. Each object can be deformable and moves freely throughout the scene.

- An object that does not change its position is called still, otherwise it is a moving object. Parameter of object motion is the object ego motion.
- An object that does not change its shape is called rigid, otherwise it is flexible. Parameter of object shape is the 3D surface representation.
- Each object is formed by a simply closed 3D surface. Flexible objects may be divided into a hierarchically structured set of rigid components while each component describes a coherent rigid part of the object surface. Components are interconnected by deformable surfaces. Each component has an individual set of motion parameters that contributes to the total object motion in a hierarchical fashion.

The concept of hierarchical object composition is best described by an example. Assume a video phone scene with static background and the head–shoulder part of a moving person's body as a time variant object. The body is a flexible object that consists of the components head and shoulder. The head itself can be described as skull and face, whereas the face has as independent components eyes and a mouth. The connection between the rigid components head and body is the flexible neck. The motion of the head may be different from the body motion but both are dependent on each other and on the global motion of the person. The dependencies of the components to each other are described through a dynamic component hierarchy. Fig. 2 shows a wireframe representation of the object from front and left side and the component hierarchy of the scene.

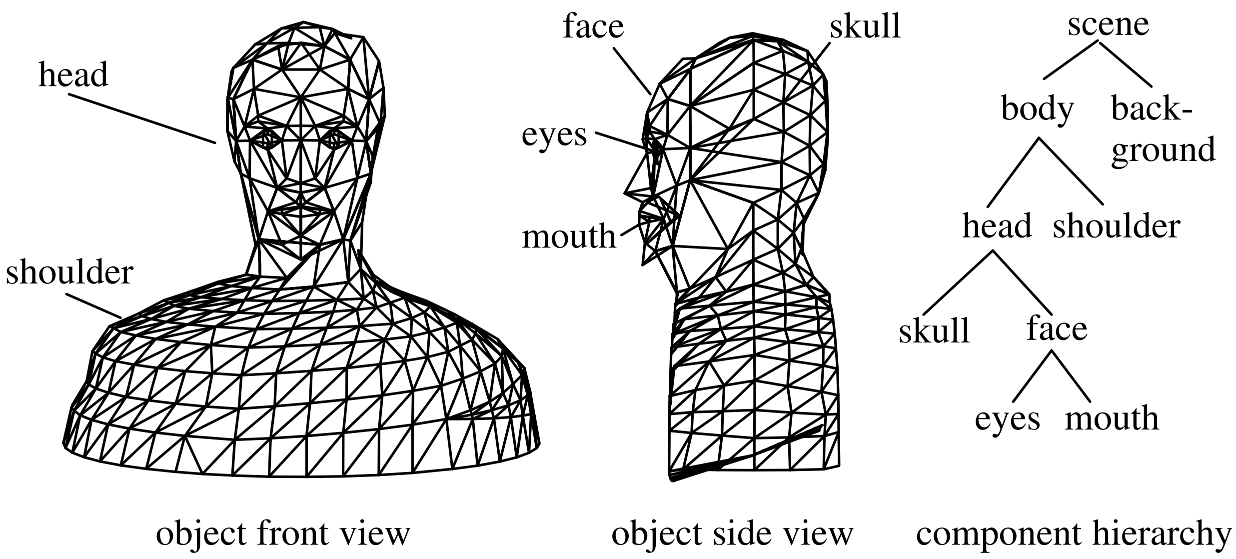


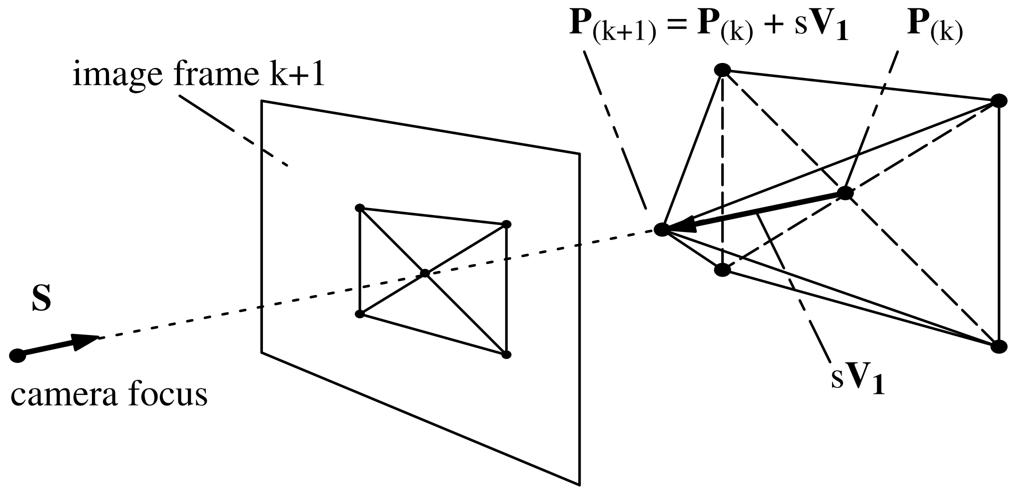
Fig. 2: Component hierarchy of object video phone scene.

1) Object shape

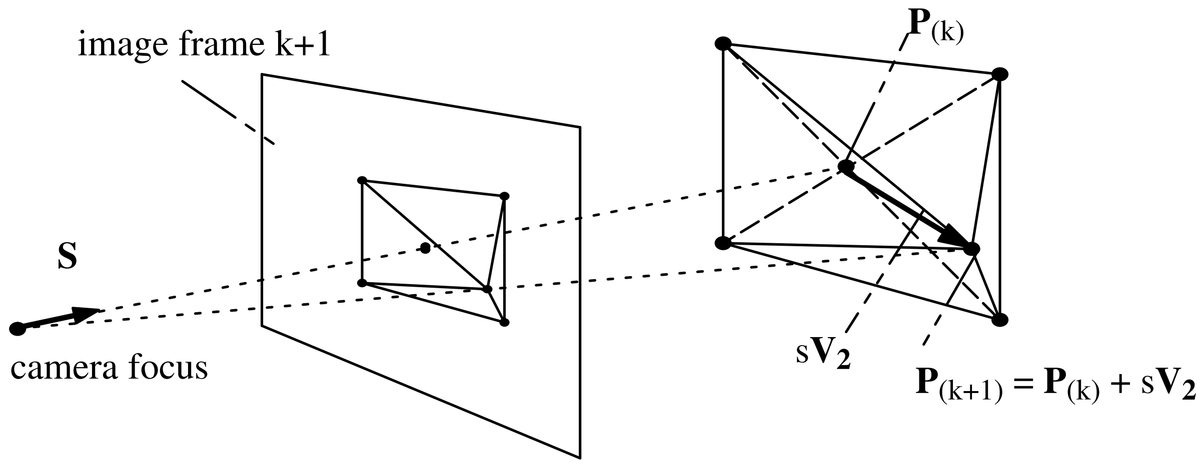
Each object is composed of a mesh of plane triangular surface patches that approximate the assumed surface shape. The position of each surface triangle is determined by three control points that span the triangle in space. Neighboring triangles share control points to form a closed surface. To change the surface shape of a component and to model flexible deformations of component surfaces from frame k to frame k+1, each control point $\mathbf{P}_{(k)}$ of the surface mesh can be shifted in space individually. Local deformations of the component surface are carried out as follows. Fig. 3 displays part of a textured 3D surface mesh with control point $\mathbf{P}_{(k)}$ and its adjoining triangular surface patches (dashed lines) whose surface normals point roughly towards the camera at frame k. The control point $\mathbf{P}_{(k)}$ is projected onto the camera target along the line of sight unit vector $\mathbf{S}$. Between frame k and k+1 the local shape is deformed by shifting the control point $\mathbf{P}_{(k)}$ to $\mathbf{P}_{(k+1)}$ along shape vector $\mathbf{V}$ (solid lines). $\mathbf{V}$ is defined as unit vector and the parameter s expresses the local shape deformation of the mesh at $\mathbf{P}_{(k)}$.

$$\mathbf{P}_{(k+1)} = \mathbf{P}_{(k)} + s \cdot \mathbf{V} \tag{1}$$

Depending on the direction of **V** different types of local distortions can be observed. In general geometric object shape as well as projected surface texture in the image plane are distorted. If **V** points along the line of sight **S**, however, the object shape can be changed without visible distortions of the projected image. This feature is useful when adjusting the object shape independently of the image projection as viewed in Fig. 3a. A shift of $\mathbf{P}_{(k)}$ along **S** does not change the visible projection of the mesh from frame k to frame k+1. This type of shape modelling is appropriate when a rigid object with unknown distance from the camera needs to be adapted. If, on the other hand, **V** is constructed to lie in a plane perpendicular to **S** then it is possible to distort the projected object surface like a rubber skin and to model flexible distortions in the image plane like a person's facial mimic. This feature is displayed in Fig. 3b. With this approach visible distortions of flexible surfaces can be modelled. Both modelling techniques will be applied during shape analysis.



3a) surface deformation $\mathbf{V_1}$ along line of sight **S**



3b) surface deformation $\mathbf{V_2}$ perpendicular to line of sight **S**

Fig. 3: Local surface modelling through control point shifting.

2) Object motion

A set of six motion parameters is associated with each component. Component motion is defined as rotation of the component control points around the component center followed by a translation of the component center, measured between two frames k and k+1. The component center **G** is the mean position vector of all N component control points. Each component control point $\mathbf{P}_{\mathbf{i}(k)}$ at frame k is transformed to its new position $\mathbf{P}_{\mathbf{i}(k+1)}$ in frame k+1 according to the general motion Eq. (2) between frame k and k+1.

$$\mathbf{P}_{\mathbf{i}(k+1)} = [\mathbf{R_G}] \cdot (\mathbf{P}_{\mathbf{i}(k)} - \mathbf{G}) + \mathbf{G} + \mathbf{T}, \qquad i = \{1, 2, ...N\} \tag{2}$$

$$\text{with } \mathbf{T} = \begin{pmatrix} T_x \\ T_y \\ T_z \end{pmatrix} = \begin{matrix}\text{translation}\\ \text{vector}\end{matrix}, \qquad \mathbf{G} = \begin{pmatrix} G_x \\ G_y \\ G_z \end{pmatrix} = \sum_{i=1}^{i=N} \frac{\mathbf{P_i}}{\mathbf{N}} = \begin{matrix}\text{component}\\ \text{center}\end{matrix}, \qquad \text{and}$$

$$[\mathbf{R_G}] = \begin{bmatrix} \cos R_y \cos R_z, & \sin R_x \sin R_y \cos R_z - \cos R_x \sin R_z, & \cos R_x \sin R_y \cos R_z + \sin R_x \sin R_z \\ \cos R_y \sin R_z, & \sin R_x \sin R_y \sin R_z + \cos R_x \cos R_z, & \cos R_x \sin R_y \sin R_z - \sin R_x \cos R_z \\ -\sin R_y, & \sin R_x \cos R_y, & \cos R_x \cos R_y \end{bmatrix} \tag{3}$$

Component rotation can be expressed by a rotation vector $\mathbf{R} = (R_x, R_y, R_z)^T$ that describes the successive rotation of the object around the three axes $(x, y, z)^T$ parallel to the scene coordinate system centered at **G**. From this vector the rotation matrix [$\mathbf{R_G}$] in Eq. (3) is derived when the identity matrix [**I**] is rotated around the coordinate axes with R_x first, R_y second and R_z last. Because [$\mathbf{R_G}$] is derived from the rotation vector **R**, the six parameters of **T** and **R** suffice to describe the 3D component motion.

B. Photometric properties

Each object in the real scene has a surface intensity distribution that can be measured by the camera. Several surface properties such as surface roughness, absorption and reflectance in combination with scene illumination form a color intensity distribution that varies with time. To simplify the modelling, a Lambertian reflectance model is assumed. This assumption leads to a surface intensity distribution, defined as surface texture, which is invariant of the viewing position. Restricting the scene illumination further to being diffuse leaves changes of surface properties as the only source of changing scene texture. When such a textured object is viewed by a camera the projection of surface texture onto the camera target forms the image. To model surface texture the inverse projection using texture mapping can be applied, provided the surface geometry was determined in advance. The real image is projected onto the model surface and the projected texture is stored separately for each triangular surface patch as a texture map.

C. Camera model for image synthesis

With the model scene as described above, complex time variant 3D scenes can be modelled. To synthesize images from the model scene a camera model is introduced [30]. The camera is controlled by its focal length and the position of the camera coordinate system relative to the scene coordinate system. It may move freely in space, recording scene projections at fixed time intervals as an image sequence.

The camera model is defined by a set of four vectors and one scalar as shown in Fig. 4. The camera coordinate system is spanned by the vector triple **{A, L, R}** that describes camera orientation while the focal point position is determined by **C**. The focal length f of the camera controls the distance of the target image plane from the camera focus along **A**. **L** and **R** span the line and row coordinates of the target image plane. 3D scene coordinates **P** are denoted by capital letters while projections of scene coordinates onto the image plane at **p** are written with small letters. Perspective transformation of a point **P** onto **p** is performed through Eq. (4)

$$\mathbf{p} = \begin{pmatrix} p_x \\ p_y \end{pmatrix} \quad \text{with } p_x = f\frac{(\mathbf{P}-\mathbf{C})\cdot\mathbf{L}}{(\mathbf{P}-\mathbf{C})\cdot\mathbf{A}} \quad \text{and } p_y = f\frac{(\mathbf{P}-\mathbf{C})\cdot\mathbf{R}}{(\mathbf{P}-\mathbf{C})\cdot\mathbf{A}} \tag{4}$$

To simplify measurements the scene oriented projection equation (4) can be transformed into a camera oriented projection equation using as camera coordinate system **C** = $(0,0,0)^T$, **L** = $(1,0,0)^T$, **R** = $(0,1,0)^T$, and **A** = $(0,0,1)^T$. This leads to the simplified projection equation (5).

$$p_x = f\frac{P_x}{P_z}, \qquad \text{and} \qquad p_y = f\frac{P_y}{P_z} \tag{5}$$

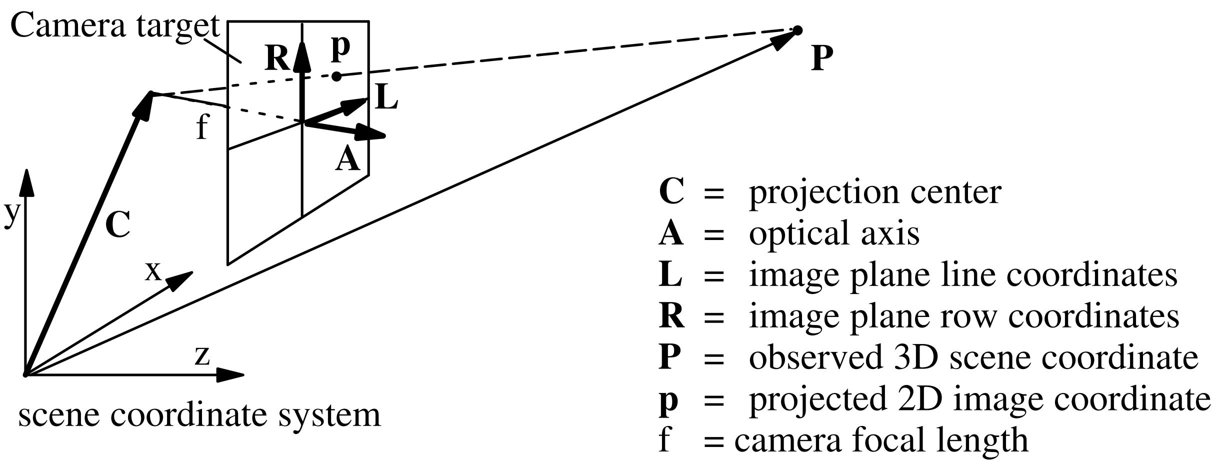


Fig. 4: Camera model.

IV. Initial Model Scene Generation

An important step in the analysis–synthesis system is the initial model estimation. A first guess of the scene geometry is needed for the system to work. The better the first approximation the less effort it is to correct the model. Depending on the application, the measurements available, and on the knowledge about the scene contents, different procedures can be utilized to estimate an initial 3D model scene. For this system the conditions applying to video phone scenes are evaluated. Monocular or at most stereoscopic image sequences are available and fully automatic modelling is a necessity. These requirements lead to a two step method. In the first step the objects are located in the observed sequence by estimating 2D object silhouettes. From these 2D object silhouettes the 3D objects are modelled in a second step using available measurements as well as scene specific knowledge.

A. Extraction of 2D object silhouettes

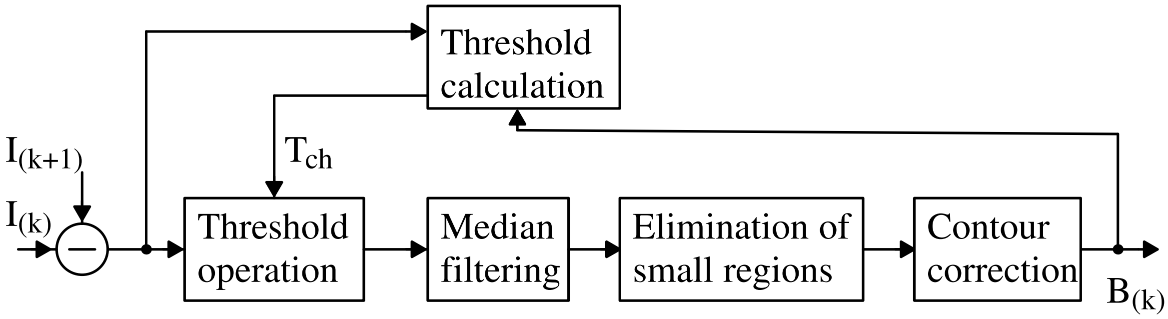
An object silhouette is the projection of the outer object boundary onto a camera target from a particular viewpoint. All image areas that conform to a specified definition of object are selected to form a mask whose outer boundary is the object silhouette. Depending on the assumptions chosen, different object properties like object texture, motion, and shape can be utilized. Examples of image segmentation using assumptions about these properties, depending on the knowledge about the imaging situation, are described in the following.

1) Texture segmentation through known background texture

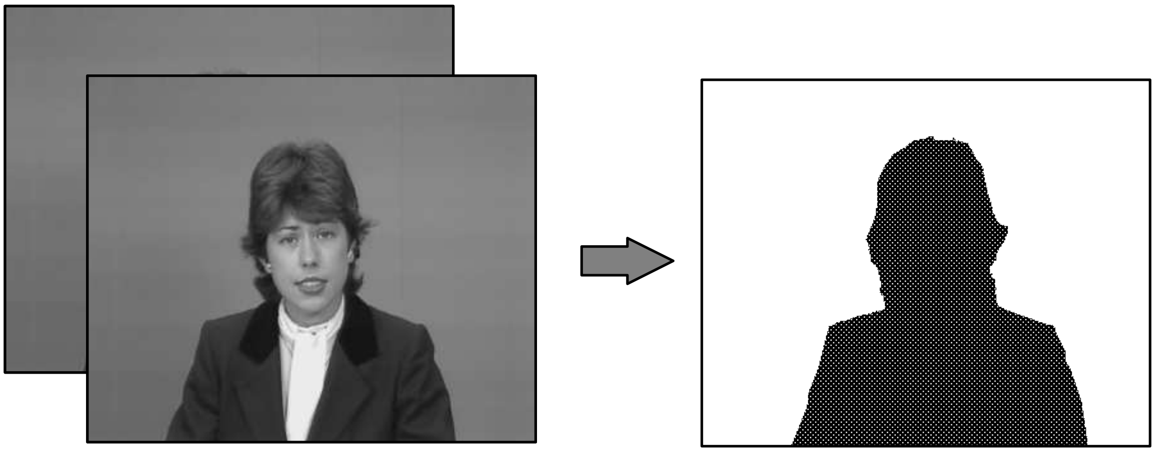
All image areas that contain a prior defined texture are selected as background; the residual areas contain the objects. This principle is easy to apply and often used in laboratory environment where the imaging situation can be controlled [25]. Objects are placed in an environment with known background texture. The background texture is extracted and the boundary between object and background forms the object silhouette. The approach is also known as 'blue screen' in television where the background has a known blue color. One image from one camera is sufficient for silhouette extraction.

2) Motion segmentation through temporal change detection

All moving parts of the image are defined as object, the static image areas contain background. The moving object area can be detected from a temporal image sequence through temporal change detection [31]. This procedure is shown in Fig. 5a.



5a) Structure of the change detection algorithm



temporal image sequence $I_{(k)}$, $I_{(k+1)}$ moving object silhouette $B_{(k)}$

b) Change detection in a video phone sequence (test sequence CLAIRE)

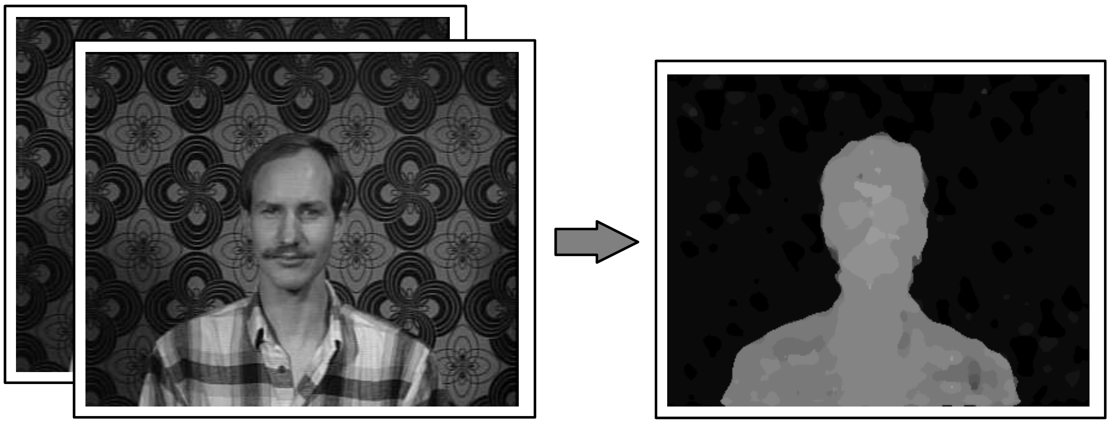
Fig. 5: Motion segmentation through temporal change detection.

Two subsequent frames $I_{(k)}$ and $I_{(k+1)}$ from a monocular image sequence are subtracted and a threshold that is adaptive to background image noise is applied. Areas with grey level difference above the threshold are assumed to contain changed areas. In a post processing step small regions and holes are eliminated and the borders are adjusted to grey level contours in the original images. The detected area contains a superimposition of the object silhouettes from frame $I_{(k)}$ and $I_{(k+1)}$

because the object motion from frame k to k+1 uncovers parts of the background in $I_{(k+1)}$. This area of uncovered background can be detected once the object motion parameters as described in Section V are applied. A detailed discussion on temporal change detection can be found in [33]. The corrected mask is labeled $B_{(k)}$, its border being the object silhouette. This procedure requires the object to have a non uniform texture and to move between both images. It may require more than two images to reliably segment the object. A typical application for video telephony is shown in Fig. 5b where a person moves in front of a static background. On the left side two subsequent frames from the video phone test sequence CLAIRE are shown. The moving part is segmented and the area of the detected object silhouette mask is marked black.

3) Shape segmentation using depth maps

An object in front of other objects or background can be separated using depth information. When a stereoscopic image sequence is available the scene depth can be measured through triangulation of corresponding image points [27]. Changes in scene depth are used to segment the object from the background because discontinuities in the depth map are interpreted as object silhouettes. Two simultaneous images $C_{1(k)}$ and $C_{2(k)}$ from different and known viewpoints are required for this segmentation. Fig. 6 shows on the left an image pair from the stereoscopic video phone sequence MICHAEL. Two cameras displaced with a focal distance of 60 mm and a verging angle of approximately 3 degrees between the optical axes were used to simultaneously record a stereoscopic sequence of a person moving in front of a static background. To evaluate scene depth, the corresponding points between the images C_1 and C_2 are searched. The search space can be reduced to a one dimensional search along the epipolar lines of the camera system because the relative orientation of the cameras to each other is known. The algorithm used to search for corresponding image points is a block oriented epipolar search based on maximum cross correlation evaluation of a small region around each image pixel [27]. The distance between two corresponding points along the epipolar line contains the depth information that is recorded in a depth map as shown in Fig. 6. Areas near to the camera are shown light, whereas the parts being further away like background are marked dark.



stereoscopic image pair $C_{1(k)}$, $C_{2(k)}$

depth map $D_{(k)}$ with object area (light color) and background (dark color)

Fig. 6: Shape segmentation using depth maps.

B. *Inferring 3D shapes from 2D silhouettes*

The object silhouette is a powerful clue to the 3D shape of an object. Certain types of symmetric objects are completely determined by their silhouettes. Generalized cylinders are an example of such objects. Often the 3D shape of an object can even be inferred from a single 2D silhouette using scene specific knowledge. This condition holds for video phone scenes. It is assumed that the object is a moving person and that the object silhouette encloses the front view of a real person. This knowledge can be exploited to create a 3D person model from a single object silhouette. In this case all information is contained in the silhouette and object depth will be a function of its silhouette only. For that purpose Ostermann [31] developed an algorithm that derives contour lines of constant object depth through erosion of the object silhouette. The depth value for each contour line is chosen as a function of its distance from the outer silhouette which determines the global shape. Fig. 7a displays an object silhouette from which the contour lines in Fig. 7b are derived by shrinking the outer silhouette. As global shape an ellipsoidal function in the xz–plane is assumed. For each contour line in the xy image plane, a z–coordinate value is calculated according to its distance from the outer silhouette and a 3D surface between the contour lines is approximated by polygons. The surface is constructed symmetrically to the image plane which results in a closed surface with ellipsoidal horizontal cross sections of the body. The resulting 3D surface is displayed in Fig. 7c where the body was rotated 30 degrees around its vertical axes and shading was applied to visualize the object shape. This yields a rather good approximation of a human body as needed in video telephony. It is, however, not possible to determine absolute 3D geometry and the model shape is scaled with the parameter of unknown scene depth as long as only one monocular view of the scene is available.

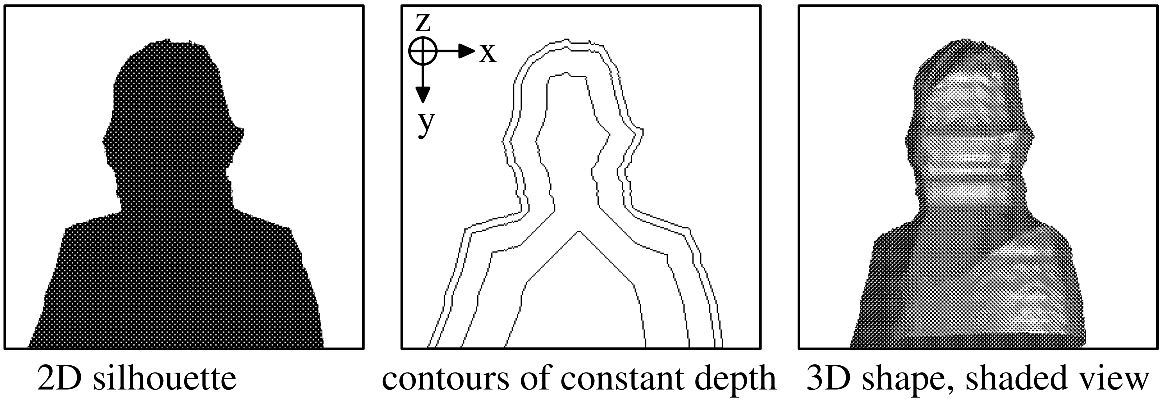


Fig. 7: 3D shape from 2D silhouette.

C. *Exploiting scene specific knowledge*

In a video phone scene it is possible to further exploit the scene specific knowledge for building an even more appropriate 3D model of a person. If it is assumed that the object to be modelled is a person with eyes and mouth, the model can be extended to contain a more realistic facial region. A strategy to reliably detect eyes and mouth area in video phone sequences and to automatically improve the 3D model of a person was presented in [29]. Areas of rapidly changing image content are interpreted as eyes and mouth of a speaking person and characteristic geometric and texture distribution features of eye and mouth prototypes serve to verify the facial position. The coarse 3D model generated from the silhouette information can then be improved with a 3D facial mask containing eyes and mouth as displayed in Fig. 2. The facial mask is a modified version of the facial mask CANDIDE, which is automatically inserted into the coarse object model generated by the change detection silhouette.

D. Surface modelling from depth maps

The methods described above permit surface inference based on 2D information and scene specific knowledge without the possibility to measure absolute 3D geometry. From a monocular view only relative geometry scaled to scene depth can be obtained. With a stereoscopic camera system as described in Section IV, absolute scene geometry can be estimated and represented by a depth map. As long as the scene contains enough textured surface areas, a dense and continuous 3D surface description is created. Only those parts of the scene that are visible to both cameras can be modelled, therefore no closed surface is obtained from a single image pair. Fig. 8 shows results of stereoscopic surface modelling from a single image pair of the sequence MICHAEL.

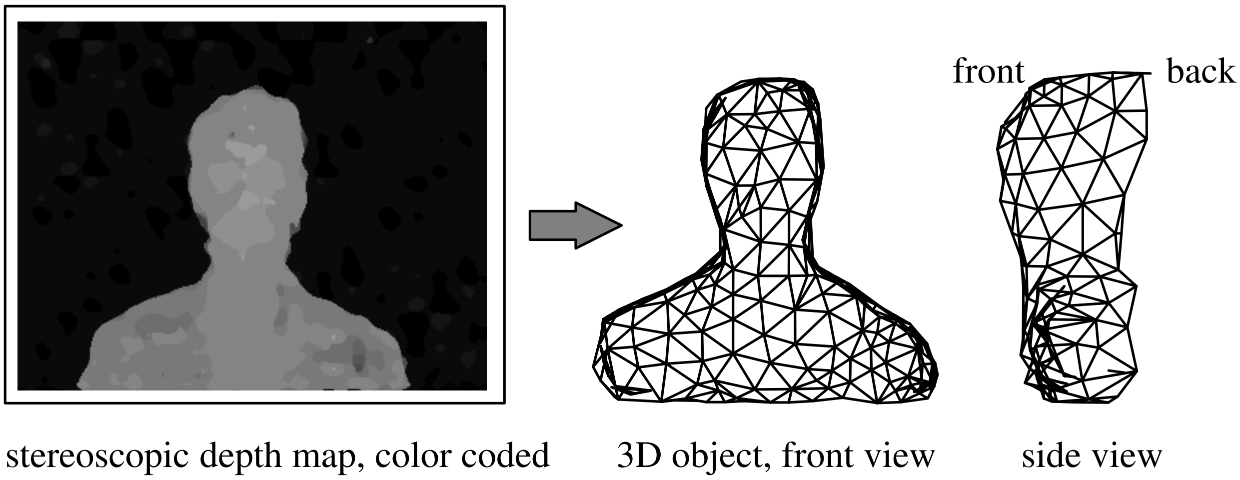


Fig. 8: Surface modelling from a depth map.

The depth map on the left is shown color coded with light areas being near to the camera and dark areas being further away. The depth map is then segmented according to scene depth and the depth map is sampled to construct a 3D surface model. A wireframe representation of the resulting object surface is shown on the right. The side view of the object shows that at least the coarse object shape of a human person is modelled without the need of shape assumptions and with absolute scene geometry.

E. Surface texture mapping

Different procedures to generate a first geometric representation of the model scene were presented above. It is now necessary to determine the photometric properties of the observed objects. All photometric properties are summarized in a surface intensity map that is projected onto the camera target. For each triangular surface patch of the model object the appropriate image patch is mapped onto the object surface in a texture map. When more than one view of the object is available, the view with the highest possible image resolution for each patch is selected as texture map. Fig. 9 explains the process of texture mapping and image synthesis. The surface is approximated through plane triangular surface patches with control points $\mathbf{P_1}$, $\mathbf{P_2}$, and $\mathbf{P_3}$. The control points are projected into image frame 1 and the circumscribing rectangle is stored as texture map for each triangle separately, which can be replaced throughout the sequence if needed. Through object or camera motion the projection of the control points change from frame 1 to frame k. From these changes an affine transform is calculated that maps the image texture from the texture map onto image frame k. A point $\mathbf{p_i}$ in frame 1 is mapped to frame k at position $\mathbf{p'_i}$ through the mapping equation.

$$\mathbf{p}'_i = \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix} \cdot \mathbf{p}_i + \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \tag{6}$$

The six parameters of the mapping equation Eq. (6) can be computed when inserting the perspective projection of control point $\mathbf{P_i}$ at frame 1 as $\mathbf{p_i}$ and its projection at frame k as $\mathbf{p'_i}$ in Eq. (6) for each of the three control points $\mathbf{P_1}$, $\mathbf{P_2}$, and $\mathbf{P_3}$. The resulting six linear equations are then solved for the parameters a_1, a_2, b_1, b_2, c_1, and c_2. Linear intensity interpolation is provided to map those intensity values where the transform projects a pixel between the regular pixel grid of the texture map. Mapping Eq. (6) provides a transform that does not account for perspective foreshortening of the triangle which will cause a geometric distortion. As long as a single triangle is small compared to its distance to the camera this projection error can be neglected.

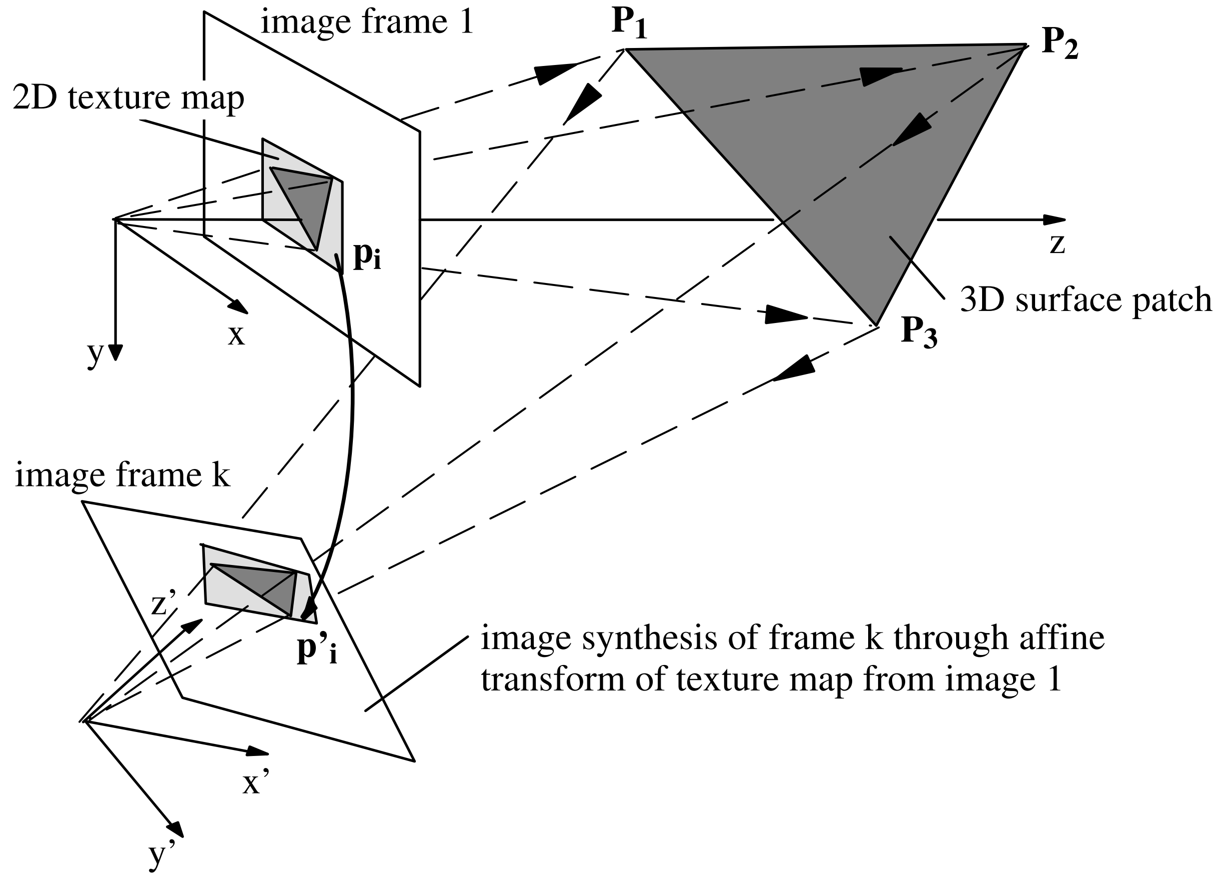


Fig. 9: Surface texture mapping and image synthesis.

V. Scene Parameter Estimation

The initial scene estimation computes a first approximation of the real object shapes. It provides a static and maybe incomplete description of the real scene with the necessity to adapt shape, motion, and texture parameters throughout the analysis process when more measurements become available. The parameter estimation is performed using an analysis through synthesis feedback control system which will be discussed in this Section. Modelling of object motion and object shape are treated separately in this approach. This separation may lead to problems when object shape and object motion have to be estimated simultaneously. For 3D motion estimation the object shape is assumed to be known, while shape estimation requires the prior determination of object motion. This dependency affects the analysis and will sometimes lead to modelling errors. As long as the initial shape approximation is reliable this dependency can be neglected.

A. Requirements for analysis through synthesis control

The only information available to the analysis system is the surface texture projected onto the camera target throughout the image sequence. From this sequence the shape and motion parameters have to be derived. Assume a scene with an arbitrarily shaped, moving textured object observed by a camera **C** during frames k and k+1 as shown in Fig. 10. The object moves between frame k (dashed object silhouette) and k+1 (solid object silhouette) according to the general motion equation Eq. (2) with motion parameters **R** and **T**. A point on the object surface, called observation point $\mathbf{P}_{(k)}$, holds the surface intensity I_1, which is projected onto $\mathbf{p}_1$ in the image plane at frame k. At frame k+1 $\mathbf{P}_{(k)}$ has moved to $\mathbf{P}_{(k+1)}$, still holding I_1. In image frame k+1 the surface intensity I_1 will now be projected at $\mathbf{p}_2$, whereas the image intensity at point $\mathbf{p}_1$ has changed to I_2.

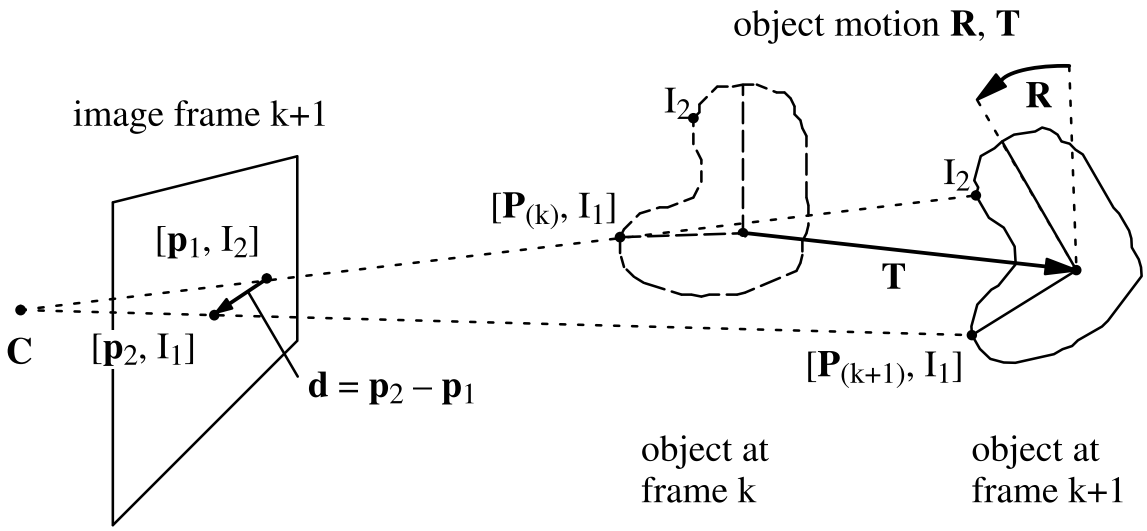


Fig. 10: Geometry for analysis by synthesis control.

The image displacement vector $\mathbf{d} = \mathbf{p}_2 - \mathbf{p}_1$ is called optical flow vector and describes the projection of the observation point displacement $\mathbf{P}_{(k+1)} - \mathbf{P}_{(k)}$ onto the image plane. When assuming a linear dependency of the surface texture between I_1 and I_2 as well as a brightness constancy constraint between frame k and k+1 it is possible to predict I_2 from I_1 and its corresponding image intensity gradients and hence to estimate $\mathbf{d}$ from the measurable difference $I_1 - I_2$.

$$I_1 - I_2 = \left(\frac{\partial I_1}{\partial x}, \frac{\partial I_1}{\partial y}\right)^T \cdot \mathbf{d} = \mathbf{g} \cdot \mathbf{d} \tag{7}$$

I_2 is measured at position of $\mathbf{p}_1$ at frame k+1, where I_1 is taken from image position $\mathbf{p}_1$ at frame k. When approximating the spatial derivatives as finite differences the optical flow vector $\mathbf{d} = (d_x, d_y)^T$ can be predicted from the image gradients $\mathbf{g} = (g_x, g_y)^T$ and the temporal image intensity difference $\Delta I_{\mathbf{p}1} = I_1 - I_2$ between frame k+1 and k at $\mathbf{p}_1$ in Eq. (8):

$$\Delta I_{\mathbf{p1}} = \mathbf{g} \cdot \mathbf{d} = g_x \cdot d_x + g_y \cdot d_y = g_x \cdot (p_{2x} - p_{1x}) + g_y \cdot (p_{2y} - p_{1y}) \tag{8}$$

In Eq. (8) $\mathbf{d}$ is related to intensity differences. Substituting the perspective projection of $\mathbf{P}_{(k)}$ and $\mathbf{P}_{(k+1)}$ for $\mathbf{p}_1$ and $\mathbf{p}_2$ in Eq. (8) yields a direct geometric to photometric transform that relates the spatial movement of $\mathbf{P}$ between frame k and k+1 to temporal intensity changes in the image sequence at $\mathbf{p}_1$.

$$\Delta I_{\mathbf{p}1} = f \cdot g_x \cdot \left(\frac{P_{(k+1)x}}{P_{(k+1)z}} - \frac{P_{(k)x}}{P_{(k)z}}\right) + f \cdot g_y \cdot \left(\frac{P_{(k+1)y}}{P_{(k+1)z}} - \frac{P_{(k)y}}{P_{(k)z}}\right) \tag{9}$$

As long as the motion between $\mathbf{P}_{(k)}$ and $\mathbf{P}_{(k+1)}$ can be expressed parametrically by a linear equation and assuming that the initial position of $\mathbf{P}_{(k)}$ is known, $\mathbf{P}_{(k+1)}$ can be substituted and Eq. (9) can be solved through evaluation of the spatiotemporal intensity differences at $\mathbf{p_1}$. Essentially every surface point can be taken as an observation point. It is, however, useful to restrict the number of observation points to those carrying relevant information. Eq. (9) uses image intensity as well as the spatial image gradients. Image areas with zero gradient can not be used for parameter estimation. A lower bound to the image gradient is additionally introduced to account for camera measurement noise. It is therefore necessary to impose a minimum gradient threshold when selecting observation points as described by Hötter [33]. For each observation point its initial 3D position on the object surface, image intensity and spatial image gradients are recorded. During the analysis each observation point is projected and the intensity difference to the real image is evaluated.

B. Direct estimation of 3D object motion

With the proposed approach, rigid 3D object motion can be estimated directly from the image sequence when the object shape is known. Assume an observation point with a known position $\mathbf{P}_{(k)}$ that moves in space and is observed by a camera **C** in Fig. 10. The motion is governed by the general motion equation (2). Assuming that rotation between successive images is small, the rotation matrix $[\mathbf{R_G}]$ from Eq. (3) can be linearized to Eq. (10).

$$[\mathbf{R}'] = \begin{bmatrix} 1 & -R_z & R_y \\ R_z & 1 & -R_x \\ -R_y & R_x & 1 \end{bmatrix} \qquad \begin{array}{l} \text{rotation matrix } [\mathbf{R_G}] \text{ is linearized to } [\mathbf{R}'] \\ \text{when setting } \sin\Phi \approx \Phi \text{ and } \cos\Phi = 1 \end{array} \tag{10}$$

When substituting $[\mathbf{R}']$ for $[\mathbf{R_G}]$ in Eq. (2), a linearized version for the general motion equation is found. $\mathbf{P}_{(k+1)}$ is expressed in explicit form in Eq. (11):

$$\mathbf{P}_{(k+1)} = \begin{pmatrix} P_{(k+1)x} \\ P_{(k+1)y} \\ P_{(k+1)z} \end{pmatrix} = \begin{pmatrix} P_{(k)x} - (P_{(k)y} - G_y)R_z + (P_{(k)z} - G_z)R_y + T_x \\ (P_{(k)x} - G_x)R_z + P_{(k)y} - (P_{(k)z} - G_z)R_x + T_y \\ -(P_{(k)x} - G_x)R_y + (P_{(k)y} - G_y)R_x + P_{(k)z} + T_z \end{pmatrix} \tag{11}$$

The parameters to be estimated are the translation **T** and the rotation **R**. After substitution of $\mathbf{P}_{(k+1)}$ in Eq. (9) the measurement equation for a single observation point $\mathbf{P}_{(k)}$ is found that is linear with respect to the unknown parameters **R** and **T** in Eq. (12) :

$$
\begin{aligned}
\Delta I_{\mathbf{p1}} \cdot P_z^2 = \; & f \cdot g_x \cdot P_z \cdot T_x \\
+ \; & f \cdot g_y \cdot P_z \cdot T_y \\
- \; & f \cdot (P_x g_x + P_y g_y + \Delta I_{\mathbf{p1}} \cdot P_z) \cdot T_z \\
- \; & f \cdot [P_x g_x (P_y - G_y) + P_y g_y (P_y - G_y) + P_z g_y (P_z - G_z) + \Delta I_{\mathbf{p1}} \cdot {}^{P}{}_{z}(P_y - G_y)] \cdot R_x \\
+ \; & f \cdot [P_y g_y (P_x - G_x) + P_x g_x (P_x - G_x) + P_z g_x (P_z - G_z) + \Delta I_{\mathbf{p1}} \cdot {}^{P}{}_{z}(P_x - G_x)] \cdot R_y \\
+ \; & f \cdot [g_y (P_x - G_x) - g_x (P_y - G_y)] \cdot P_z \cdot R_z
\end{aligned}
\tag{12}
$$

with $\mathbf{P}_{(k)}$ abbreviated as $(P_x, P_y, P_z)^T$.

1) Conditions for robust motion estimation

At least six distinctive observation points that lead to six linear independent equations are needed to calculate the six motion parameters **R** and **T**. In real imaging situations the measurements of the spatiotemporal derivatives are noisy and some of the observation points selected may be linear dependent of each other. To cope with those conditions an over constrained set of equations is established and a linear regression is carried out using least squares fit. All observation points of one object component are evaluated. It is important to note that we do **not** measure optical flow locally and then try to combine the flow field. Instead **all** observation points of a rigid surface are used to solve for **R** and **T**. To account for linearizing, the estimation is iterated. The position **P** of each observation point is initially determined by object shape and position. An estimate of the parameters **R** and **T** is calculated and the observation point is moved according to those parameters. The estimation is repeated with the new starting position of **P** until the parameter changes of **T** and **R** converge to zero. Because the motion update is carried out with the true rotation matrix $[\mathbf{R_G}]$, linearization errors vanish.

To improve estimation stability, dependencies between rotation and translation parameters were cancelled out through the introduction of a center of rotation **G**. The rotation of an object around an arbitrary rotation center can be separated into a rotation of the object around the object's center of gravity and an additional translation of the object. Such a decomposition leads to an independent estimation of **R** and **T** and improves convergence of the solution.

The analysis should be robust against noisy measurements or measurements which are erroneous due to invalid model assumptions. Therefore the mean temporal intensity is computed and observation points with high intensity errors are excluded from the regression in a modified least squares fit. The measurement certainty of each parameter can be estimated through evaluation of the error covariance matrix of the regression [33]. When a parameter has an uncertain estimate it can be excluded from the regression to ensure a stable estimate for the remaining parameters.

The frame to frame estimation as described above may result in motion error accumulation because there is an error in each frame to frame motion estimate. To cope with error accumulation, observation points are created on the object surface and are tracked as long as they are visible and conform to the assumptions made above. In areas with changing image contents new observation points are created if necessary. A single observation point is created in frame k and tracked to frame k+n where n is the number of images throughout which the observation point is visible. This is equivalent to a multiple frame estimation which reduces error accumulation considerably as compared to a local frame to frame estimation.

2) Estimation of nonrigid object motion

A scene as described in Fig. 2 is constructed as a hierarchical set of flexibly interconnected components. Assume a part of the human body, consisting of the components head and shoulder. The body itself is flexible but consists of the rigid head and shoulder part. The motion of such dynamic objects can be estimated through a hierarchical coarse to fine strategy of its component motion parameters. On the first hierarchy level, the body with head and shoulders is treated as rigid and the mean object motion is determined. This object motion may be large, but due to the large measurement area available, the mean object motion can be estimated with high confidence. On the second hierarchy level, the residual motion of either head and shoulder component is estimated to compensate for nonrigid motion of the body. The residual motion is small because the mean object motion was determined before. Therefore the accurate parameter set of head and shoulder motion can be estimated. This procedure is iterated until the best motion parameter sets for the head and shoulder part are found. It allows motion estimation for nonrigid objects that are segmented into a hierarchical set of rigid components [34].

When analyzing dynamic scenes the segmentation of nonrigid objects into rigid components is not known in advance. It is therefore necessary to automatically split nonrigid objects into their rigid parts. During analysis of a moving human speaker, for example, at first the speaker is assumed to be rigid and is treated as one component. After estimating the mean body motion, residual errors will remain that are due to the nonrigid nature of the speaker and can not be explained by the rigid object model. From the residual errors a 2D optical flow vector field is computed and clustered into areas of common image plane motion. The object is then split according to the residual motion areas and components that best confirm to rigid 3D motion are formed. 3D motion compensation is then performed with the refined object structure, now consisting of hierarchically dependent, connected components [34].

C. Shape estimation

3D motion estimation as discussed above assumes that the object shape is known. During the initial estimation of the model scene, however, only an approximation of the real object shape is calculated. Throughout the analysis it may be necessary to adapt changing object shapes or to find a better approximation of the true 3D shape. When more than one view, each with a known perspective, is available the object shape can be adapted by synthesis feedback control. Images are synthesized from the initial model scene at the known viewpoints and are compared to the real images. Differences between real and synthesized images are interpreted as errors in object shape. In a shape adaptation loop the shape of the model objects is changed until image intensity differences are minimized.

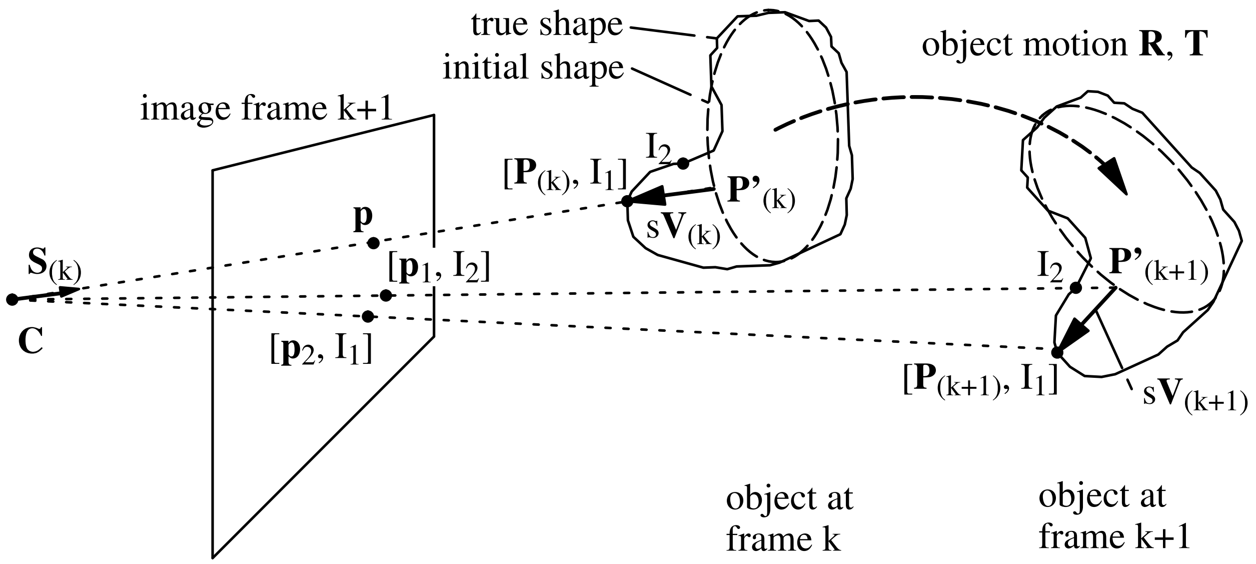


Fig. 11: Shape adaptation geometry.

Fig. 11 displays the shape adaptation geometry. The assumed initial shape is drawn as dotted oval and the true shape as solid closed line. During frame k the true observation point position $\mathbf{P}_{(k)}$ lies somewhere on the line of sight $\mathbf{S}_{(k)}$ between the assumed observation point $\mathbf{P'}_{(k)}$ and its projection onto the image plane at $\mathbf{p}$. The true position of $\mathbf{P}_{(k)}$ is defined as

$$\mathbf{P}_{(k)} = \mathbf{P'}_{(k)} + s \cdot \mathbf{V}_{(k)} \tag{13}$$

where $\mathbf{P}_{(k)}$ defines the assumed observation point position, $\mathbf{V}_{(k)}$ is a unit vector pointing along the line of sight $\mathbf{S}_{(k)}$ and s is the shape parameter that controls the deviation of the true shape from the assumed shape. After motion compensation between frame k and k+1 the observation point $\mathbf{P}_{(k+1)}$ with surface intensity I_1 is projected onto the image plane at $\mathbf{p}_2$, whereas the projection of the assumed observation point $\mathbf{P'}_{(k+1)}$ at $\mathbf{p}_1$ holds the surface intensity I_2. Shape vector $\mathbf{V}_{(k)}$ is rotated to $\mathbf{V}_{(k+1)}$ by the object rotation $[\mathbf{R}]$ to

$$\mathbf{V}_{(k+1)} = [\mathbf{R}] \cdot \mathbf{V}_{(k)} \tag{14}$$

which is projected onto the image plane at $\mathbf{p}_2 - \mathbf{p}_1$. The intensity difference $I_2 - I_1$, caused by the object shape error between **P’** and **P** at frame k+1, will be minimized. It should be noted that shape measurement is possible only as long as the direction of $\mathbf{V}_{(k+1)}$ deviates from the line of sight to $\mathbf{P'}_{(k+1)}$, otherwise $\mathbf{p_2}$ equals $\mathbf{p_1}$ and the image difference vanishes.

Following the approach of Eq. (9) it is necessary to find a linear dependency between $\mathbf{P}_{(k+1)}$ and $\mathbf{P'}_{(k+1)}$ at frame k+1. The shape error as given in Eq. (13) is governed by the shape parameter s. Substituting $\mathbf{P}_{(k+1)} = \mathbf{P'}_{(k+1)} + s\mathbf{V}_{(k+1)}$ as derived in Eq. (13) and inserting $\mathbf{P'}_{(k+1)}$ for $\mathbf{P}_{(k)}$ in Eq. (9), a relationship between the surface intensity difference $\Delta I_{\mathbf{p1}} = I_1 - I_2$ and the shape parameter s at frame k+1 can be found.

$$\Delta I_{\mathbf{p1}} = fg_x \cdot \left(\frac{P'_x + sV_x}{P'_z + sV_z} - \frac{P'_x}{P'_z}\right) + fg_y \cdot \left(\frac{P'_y + sV_y}{P'_z + sV_z} - \frac{P'_y}{P'_z}\right) \tag{15}$$

From Eq. (15) the shape parameter s can be calculated as

$$s = \frac{\Delta I_{\mathbf{p1}} \cdot P'^2_z}{fg_x \cdot (V_x P'_z - V_z P'_x) + fg_y \cdot (V_y P'_z - V_z P'_y) - \Delta I_{\mathbf{p1}} \cdot V_z P'_z} \tag{16}$$

The shape parameter s can be estimated for each observation point separately. In real image sequences a single measurement of s will give unreliable results because of observation noise and erroneous model assumptions. More reliable results are obtained when several observation points in a local environment are used to determine the best value of s through linear regression.

In a similar fashion the flexible deformation of surface texture can be estimated. According to Fig. 3b flexible surface deformations can be measured in a plane perpendicular to the line of sight $\mathbf{S}_{(k)}$. Instead of a single shape parameter s for the observation point shift along $\mathbf{V_2}$ there will be two parameters (s_H, s_V) along the coordinates of the plane spanning the surface $\mathbf{S} \cdot \mathbf{V_2} = 0$. The plane is defined by the two vectors $\mathbf{V_V}$ and $\mathbf{V_H}$ with $\mathbf{V_V} \cdot \mathbf{V_H} = 0$. They substitute $\mathbf{V_2}$ through Eq. (17)

$$s \cdot \mathbf{V}_2 = s_H \mathbf{V_H} + s_V \mathbf{V_V} \tag{17}$$

The flexible surface shift is then defined as

$$\mathbf{P} = \mathbf{P}' + s_H \mathbf{V_H} + s_V \mathbf{V_V} \tag{18}$$

Substituting Eq. (18) into Eq. (9) we find the equation for estimating the parameters for a flexible surface shift that can be solved using least squares fit

$$\begin{aligned} \Delta I_{\mathbf{P}_1} = & f\left\{ g_x\left(\frac{V_{Hx}}{P'_z} - \frac{P'_x V_{Hz}}{P'^2_z}\right) + g_y\left(\frac{V_{Hy}}{P'_z} - \frac{P'_y V_{Hz}}{P'^2_z}\right)\right\} s_H + \\ & f\left\{ g_x\left(\frac{V_{Vx}}{P'_z} - \frac{P'_x V_{Vz}}{P'^2_z}\right) + g_y\left(\frac{V_{Vy}}{P'_z} - \frac{P'_y V_{Vz}}{P'^2_z}\right)\right\} s_V \end{aligned} \tag{19}$$

D. Elimination of residual errors

The image analysis described so far is based on a parametric model that attempts to estimate the parameters of the real scene. In complex scenes, however, there will always be areas that do not comply to the model which is a simplified assumption of the real world. In the analysis through synthesis system the success of modelling is measured by comparison between synthesized and real image sequence as well as through evaluation of subjective image quality. Areas where the model assumptions fail and which will cause annoying visual distortions in the synthesized image are termed 'model failures' [35]. The opening of a person's mouth, for example, can not be modelled because the teeth will appear in the image. Typical areas of model failure are therefore eyes and mouth in a person's face. In these areas the surface texture changes and has to be updated from the current image frame. The detection of model failure areas is performed in two steps as displayed in Fig. 12 for one frame of the video phone test sequence CLAIRE. Fig. 12a shows one frame of the original image sequence and Fig. 12b displays the motion compensated difference image between the real and synthesized image for that frame. A threshold is applied to the difference image to detect areas with a synthesis error being significantly higher than image background noise as shown in Fig. 12c. From the synthesis error areas only those areas that exhibit visual distortions are selected as model failure. Subjective tests have shown that small and line–shaped structures that are due to slight object displacements do not disturb the subjective image quality. Only large and solid areas have to be exchanged. These areas, that comply with model failures, are marked black in Fig. 12d.

Fig. 12: a) original image from sequence CLAIRE, b) difference image between real and synthesized image, c) areas of residual synthesis error, d) areas of model failure.

VI. Modelling of a Speaker in a Video Phone Scene

The scene analysis approach presented in this paper has been applied to different imaging situations. Special emphasis was put on applications for data compression and visual simulations. Very high data compression rates are needed for transmission of time variant TV signals via telephone lines. These high compression rates can only be achieved in certain restricted imaging situations where the scene content is rather static and only few dominant objects are visible. This situation is given when transmitting video phone scenes. In a typical video phone scene the head and shoulder part of a person is visible. The person itself, however, may exhibit complex nonrigid motion and may have flexibly deformed surfaces like the facial mimic of the head. Even time variant surface texture must be considered when the person closes the eyes. The goal of model based analysis in this case is to find a very compact description of the relevant contents of the image sequence through a parametric scene description that is transmitted to the receiver. The optimality criterion is to generate the most compact description of the parameters needed at the receiver to synthesize an image that best resembles the original input image but is computed from the parametric model scene.

A. Modelling strategy

The concept of controlled analysis through synthesized image feedback is specially suited for this approach because the model scene is constantly adapted to the new scene contents through comparison with the real image sequence. The task is difficult even for simple scenes because all 3D information needs to be extracted from a single camera viewpoint. To completely determine all model parameters from the image sequence a strategy was developed that sequentially exploits scene specific knowledge to guide the analysis process [36]. This knowledge is formulated as a set of specific hypotheses that are verified throughout the analysis. Starting from very general assumptions, each hypothesis introduces new model assumptions and guides a special part of the analysis. The assumptions and the actions taken accordingly are summarized in the following.

- Hypothesis 1: The scene consists of a moving person in front of a static background.
  Hypothesis 1 governs the initial model generation. An initial 3D object is modelled based on the change detection silhouette.

- Hypothesis 2: The object exhibits rigid 3D motion.

  It is assumed that the object consists of one rigid component and the mean 3D object motion is estimated. This global motion compensation reduces the mean estimation error.

- Hypothesis 3: The object may consist of a set of flexibly connected rigid components. The residual error is analyzed in order to search for clusters of rigid motion components. The object is split into a connected set of rigid components and the 3D motion of each component is estimated.

- Hypothesis 4: The object shape may be erroneous.

  Residual errors after motion compensation are attributed to wrong object shape and flexible texture deformations. To adapt the object surface each visible control point **P** of the object surface is analyzed locally and is shifted if necessary according to the shape equations. To distinguish between wrong 3D shape and flexible surface texture sequence information over time is employed. It is assumed that the true 3D shape of the person is unknown but time invariant and only the facial expression changes with time. This allows to attribute fast temporal shape changes to flexible texture deformations and the long term shape deviation is interpreted as wrong 3D shape. The shape changes are recorded over time, the mean shape is adapted and fast flexible distortions are corrected from frame to frame.

Hypotheses 1 to 4 are concerned with parametric modelling of the scene. In complex scenes there will always be image areas that do not comply with the underlying model assumption and hence can not be modelled. These areas are termed ’model failure areas’ that have to be detected and replaced directly from the original image sequence. Hypothesis 5 deals with this kind of errors.

- Hypothesis 5: The object has a time variant surface texture (model failure areas).

  The remaining intensity errors between synthesized and real image sequence are attributed to changes in surface texture and failure of model assumptions. Those areas that exhibit annoying visual distortions in the image are detected as model failure areas and surface texture from the real image is inserted into the model through texture mapping.

B. Results

To demonstrate the results achieved by the presented algorithms, two experiments were carried out for modelling a human speaker in a monocular video phone sequence, the first focussing on the implications for efficient coding in video phone transmission and the second discussing the visible modelling results achievable for 3D motion tracking and shape adaptation.

1) model based image coding

A fully automatic video phone coding system for 64 kbit/s transmission channels based on the modelling of rigid 3D objects was developed by Ostermann and demonstrated at the Picture Coding Symposium 1991 [37]. It complies to the modelling hypotheses 1, 2, 3, and 5. The object is initialized by silhouette change detection and consists of hierarchically connected, rigid components. A hierarchical 3D motion compensation of the components is performed but no shape adaptation according to hypothesis 4 is carried out. Model failures are detected and replaced throughout the sequence and the resulting model data stream is coded and transmitted to the receiver where it is used to synthesize an output sequence from the transmitted model scene. The coding system has been tested with the 10 frames/s video phone test sequences CLAIRE and MISS AMERICA. It was shown that the model based coder is capable to synthesize on the average 96% of the image area solely from the model scene, needing only shape and motion information. The shape and motion information, however, can be transmitted with 2000 bit/frame, leaving the remaining 4400 bit/frame to code the 4% model failure area in the image. The image quality synthesized with this simplified model is similar to the image quality achieved with the standard H261 coding scheme. Subjective image perception, however, is improved because the model failure areas that are transmitted with high quality are concentrated in the relevant areas of eyes and mouth. Image quality will be enhanced further once shape adaptation and modelling of flexible facial mimic according to hypothesis 4 is included. An additional advantage of the 3D modelling approach is the possibility to include even more specific models like 3D facial masks that allow a parametric interpretation of the facial mimic [29].

2) visual demonstration of 3D scene modelling

A monocular 90 frame video phone sequence with 10 frames/s frame rate displaying a moving person in front of static background is analyzed. Fig. 13 depicts the results of the model based analysis applied to the video phone sequence after modelling of 10 and 50 input images. All modelling hypotheses were employed to analyze object motion, shape, and time varying surface texture. The left part displays frame 10 and 50 of the output sequence synthesized from the model scene, the middle and right parts show shaded surface representations of the analyzed 3D object shape in front and side views. It can be seen that the 3D motion of the person is tracked correctly and that the shape analysis adapts the facial area. The person rotates the head only about 15 degrees, therefore little shape information can be extracted from the sequence. Modelling errors occur at regions where the surface texture is time variant (mouth area) or position variant (specular reflections on the forehead). To model these areas properly an extended illumination and surface reflectance model must be included in the analysis. Due to the strictly local shape adaptation the model surface tends to become bumpy over time. A global shape interpolation or some smoothness constraints could be helpful to overcome that problem.

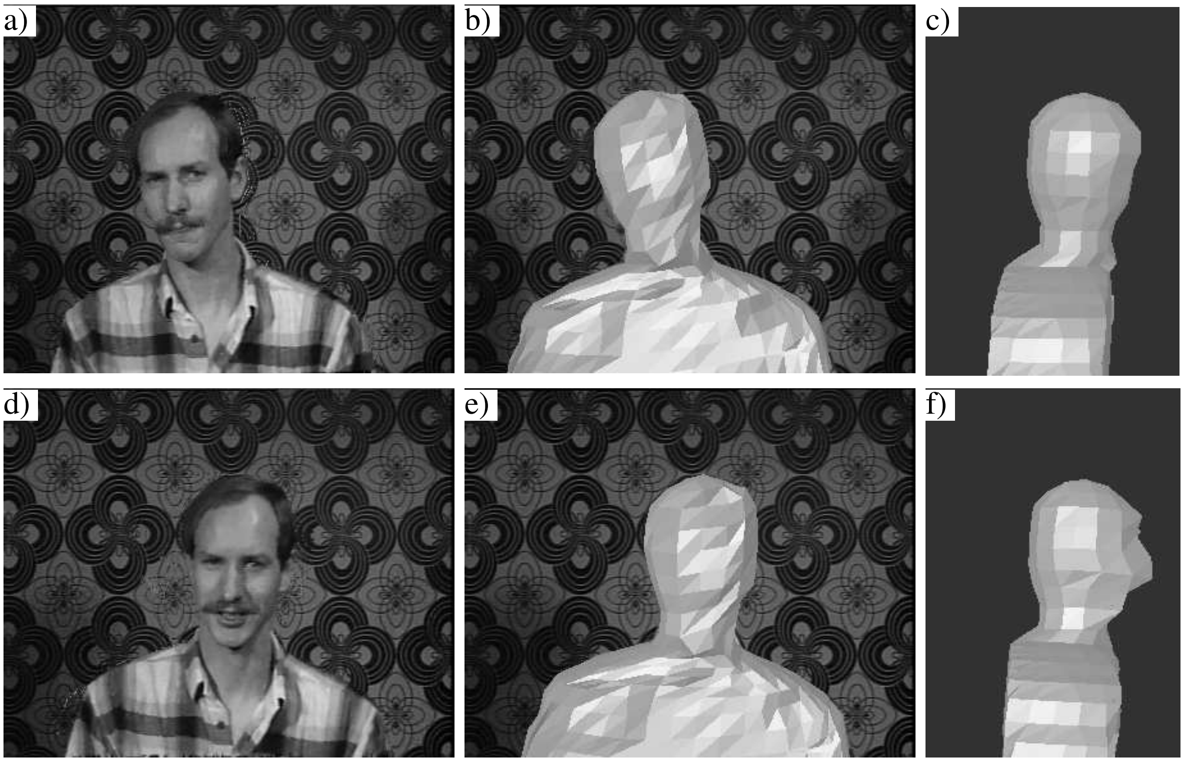


Fig. 13: Modelling results. (a) Synthesized model image and (b) front view and (c) side view of object shape after processing of 10 input images;
(d) synthesized model image and (e) front view and (f) side view of object shape after processing of 50 input images.

C. Conclusion

In this contribution an approach to 3D scene modelling from 2D image sequences was presented and the special topic of analysis by synthesis for data compression purposes in video phone scenes was discussed.

In a video phone scene the speaker is modelled as a dynamic flexible object with arbitrary 3D motion. Scene specific knowledge is applied to govern the analysis process and motion, shape, and texture parameters are extracted from the image sequence through image synthesis feedback. All information available originates from a monocular image sequence and the modelling hypotheses are processed in sequence. Motion and shape analysis, however, are not independent of each other. 3D motion can only be estimated when the object shape is known and the shape can only be adapted from known viewpoints. This dependency will affect the analysis when the initial shape approximation is insufficient. Solutions to the problem could be to estimate motion and shape parameters simultaneously through iterating or combined methods, or to introduce additional sensors that allow an independent measurement of shape and motion.

The problem of simultaneous estimation of motion and shape leads to solving systems of equations with a large number of parameters and additional constraints are necessary to govern this approach. We are currently investigating solutions to this problem. Improved modelling results are also expected when independent measurements are available for shape and motion analysis. Such independent measurements can be obtained when using stereoscopic image sequences as input. From the binocular image pair at a certain time instant, 3D shape parameters are extracted without any motion influence. 3D motion between successive image pairs is then estimated with the measured object shape and the motion equations can be extended to fully exploit the stereoscopic sequence information. With this approach we expect to be able to model more complex scenes and to apply the algorithms to the emerging field of stereoscopic television.

ACKNOWLEDGEMENT

The work presented in this paper is based on previous research carried out by F. Kappei and H. Busch. J. Ostermann provided the image coding results. The research was made possible by a grant from the Research Institute of the German TELEKOM.

REFERENCES

[1] J.K. Aggarwal, N. Nandhakumar, ”On the Computation of Motion from Sequences of Images –A Review,” *Proc. of the IEEE*, Vol. 76 (8), pp. 917–935, Aug. 1988.

[2] A. Blake, A. Zissermann, ”Visual Reconstruction,” *MIT–Press*, Cambridge, 1987.

[3] D. Marr, K. H. Nishihara, ”Representation and recognition of the spatial organisation of three-dimensional shapes,” *Proc. Royal. Soc. Lond. B*, vol. 200, pp. 269 –294, 1978.

[4] H.G. Musmann, M. Hötter, J. Ostermann, ”Object–oriented analysis–synthesis coding of moving images,” *Signal Processing: Image Communication*, Vol. 1 (2), pp. 117–138, Nov. 1989.

[5] R. Forchheimer, T. Kronander, ”Image Coding– from Waveforms to Animation,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 37 (12), pp. 2008–2033, Dec. 1989.

[6] H. Harashima, F. Kishino, ”Intelligent Image Coding and Communications with Realistic Sensations – Recent Trends,”, *IEICE Transactions*, Vol. E 74 (6), pp. 1582–1592, June 1991.

[7] R. A. Jarvis, ”A Perspective on Range Finding Techniques for Computer Vision,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 5 (2), pp. 122–139, March 1983.

[8] H.H. Baker, T.O. Binford: ”Depth from edge and intensity based stereo,” *Proc. seventh Int. joint Conf. Artif. Intell*. pp. 632–636, 1981.

[9] J. Aloimonos, D. Shulman, ”Integration of Visual Modules,” *Academic Press*, San Diego, USA, 1989.

[10] B.K.P. Horn, ”Robot Vision,” *McGraw–Hill*, New York, 1986.

[11] A.N. Netravali, J. Salz, ”Algorithms for Estimation of Three–Dimensional Motion,” *AT&T Technical Journal*, Vol. 64 (2), 1985.

[12] R.Y. Tsai, T.S. Huang, ”Uniqueness and estimation of three–dimensional motion parameters of rigid objects with curved surfaces,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 6 (1), pp. 13–26, 1984.

[13] D. Terzopoulos , D. Metaxas, ”Dynamic 3D Models with Local and Global Deformations: Deformable superquadratics,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13 (7), pp. 703–714, July 1991.

[14] A. Pentland, B. Horowitz, ”Recovery of Nonrigid Motion and Structure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13 (7), pp. 730–742, July 1991.

[15] M. Rydfalk, ”CANDIDE, a parametrized face,” *Internal Report LITH–ISY–I–0866*, Linköping University, 1987.

[16] P. Ekman, W. Friesen, ”Facial Action Coding System,” *Consulting Psychologists Press*, Palo Alto, CA., 1977.

[17] P. Roivainen, ”Motion estimation in model based coding of human faces,” *Thesis 225*, Linköping University, May 1990.

[18] W. Welsh, ”Model–based coding of images ,” *Ph.D. Thesis*, University of Essex, Jan. 1991.

[19] M. Buck, N. Diehl, ”Segmentation and modelling of head and shoulder scenes,” *3rd Int. Workshop on 64 kbit/s Coding of Moving Video*, Rotterdam, Sept. 1990.

[20] K. Aizawa, H. Harashima, T. Saito, ”Model–based analysis synthesis coding for a person’s face,” *Signal Processing: Image Communication*, Vol. 1 (2), pp. 139–152, Oct. 1989.

[21] F. Kappei, C.–E. Liedtke, ”Modelling of a natural 3D scene consisting of moving objects from a sequence of monocular TV images,” *Proc. SPIE,* Vol. 860, pp. 126ff, 1987.

[22] F. Kappei, ”Modellierung und Rekonstruktion bewegter dreidimensionaler Objekte in einer Fernsehbildfolge,” *Ph.D. Thesis*, University of Hannover, Nov. 1988.

[23] H. Busch, ”Ein Verfahren zur Oberflächenmodellierung dreidimensionaler Objekte aus Kamerabildfolgen,” *Ph.D. Thesis*, University of Hannover, April 1991.

[24] C.–E. Liedtke, H. Busch, R. Koch, ”Automatic Modelling of 3D Moving Objects from a TV Image Sequence,” *SPIE Conf. on Sensing and Reconstruction of 3D–Objects and Scenes*, Vol. 1260, pp. 230–239, Santa Clara, USA, Feb. 1990.

[25] H. Busch, ”Automatic Modelling of Rigid 3D Objects Using an Analysis by Synthesis System,” *SPIE Conf. on Visual Communications and Image Processing IV,* Vol. 1199, pp. 356ff, Philadelphia, USA, 1989.

[26] R. Koch, ”Automatic Modelling of Natural Scenes for Generating Synthetic Movies*,” Proc. of European Computer Graphics Conference Eurographics ’90*, North Holland Publ., pp. 215–224, Montreux, Switzerland, Sept. 1990.

[27] R. Koch, C.–E. Liedtke, ”3D–Modelling of Natural Scenes from Image Sequences for Stereoscopic Reconstruction,” *Proc. of 2nd Workshop on Stereoscopic Television*, Darmstadt, Germany, Sept. 1990.

[28] R. Koch, ”Model–based 3D Scene Analysis from Stereoscopic Image Sequences”, *Proc. of the International Society of Photogrammetry and Remote Sensing*, ISPRS–Conference, Washington, D.C., USA, Aug. 1992.

[29] R. Koch, ”Adaptation of a 3D Facial Mask to Human Faces in Videophone Sequences using Model Based Image Analysis,” *Proc. of Picture Coding Symposium*, pp. 285–288, Tokyo, Japan, Sept. 1991.

[30] Yakimovski, Y., R. Cunningham, ”A System for Extracting 3D Measurements from a Stereo Pair of TV Cameras,” *Computer Graphics and Image Processing,* Vol 7, pp. 195–210, 1978.

[31] J. Ostermann, ”Modelling of 3D Moving Objects for an Analysis–Synthesis Coder,” *SPIE Conf. on Sensing and Reconstruction of 3D–Objects and Scenes*, Vol. 1260, pp. 240–249, Santa Clara, USA, Feb. 1990.

[32] T. Aach, A. Kaup, ”Partitioning of Stereoscopic Sequences by Evaluation of Stereo Disparity and Temporal Change Detection,” *Proceedings of Picture Coding Symposium*, pp. 443–446, Tokyo, Japan, Sept. 1991.

[33] M. Hötter, R Thoma, ”Image segmentation based on object oriented mapping parameter estimation,” *Signal Processing*, Vol. 15 (3), pp. 315–334, Oct. 1988.

[34] H. Busch, ”Subdividing nonrigid 3D objects into quasi rigid parts,” *IEE 3rd Intl. Conf. on Image Processing and Applications*, IEE Publ. 307, pp. 1–4,Warwick, UK, July 1989.

[35] J. Ostermann, ”Detection and coding of model failures in an analysis–synthesis coder based on moving 3D objects,” *3rd Int. Workshop on 64 kbit/s Coding of Moving Video*, Rotterdam, Sept. 1990.

[36] R. Koch, ”A hierarchical strategy for analyzing motion, shape, and color of three dimensional objects from an image sequence,” *Proc. of 2nd International Workshop on 64 Kbit/s Coding of Moving Video*, Hannover, Oct. 1989.

[37] J. Ostermann, H. Li,” Coding of color parameters in an analysis–synthesis coder based on moving 3D–objects”, *Proc. of Picture Coding Symposium*, pp. 315–318, Tokyo, Japan, Sept. 1991.