Adaptive Reconstruction using Multiple Views

Morgan Ulvklo Gösta Granlund Hans Knutsson

Computer Vision Laboratory
Department of Electrical Engineering
Linköping University, SE-581 83 Linköping, Sweden
email:morgan@isy.liu.se

2 June, 1998

LiTH-ISY-R-2036, 1998-06-08, ISSN 1400-3902

Abstract

This report introduces a signal processing strategy for depth segmentation and scene reconstruction that incorporates occlusion as a natural component. The work aims to maximize the use of connectivity in the temporal domain as much as possible under the condition that the scene is static and that the camera motion is known. An object behind the foreground is reconstructed using the fact that different parts of the object have been seen in different images in the sequence. One of the main ideas in the reported work is the use of a spatiotemporal certainty volume $c(\mathbf{x})$ with the same dimension as the input spatiotemporal volume $s(\mathbf{x})$, and then use $c(\mathbf{x})$ as a "blackboard" for rejecting already segmented image structures. The segmentation starts with searching for image structures in the foreground, eliminates their occluding influence, and then proceeds. Normalized convolution, which is a Weighted Least Mean Square technique for filtering data with varying spatial reliability, is used for all filtering. High spatial resolution near object borders is achieved and only neighboring structures with similar depth supports each other.

Contents

1 Introduction **1**
1.1 Related work . . . 1

2 Scene and camera description **1**
2.1 Epipolar constraint . . . 1

3 Signal model **2**
3.1 Signal model for low-altitude surveillance . . . 2
3.2 Projection model . . . 4

4 Normalized convolution in general **5**
4.1 Historical perspective . . . 5
4.2 Basis and spaces . . . 5
4.3 Weighted Least Mean Square Method (WLMS) . . . 5
4.4 WLMS as convolutions . . . 6
4.5 Application for normalized convolution . . . 6

5 Segmentation algorithm **6**
5.1 Basis functions for projection analysis . . . 7
5.1.1 Projection vectors . . . 7
5.1.2 Basis functions . . . 8
5.2 Global models based on linear phase . . . 9
5.2.1 Hierarchical models . . . 9
5.2.2 Global projection analysis . . . 10
5.2.3 Phase-based disparity . . . 11
5.3 Spatial consistency in the $(x, y, depth)$-domain . . . 12
5.4 Reconstruction filters . . . 12

6 Results **13**

7 Applications **13**

8 Discussion **15**

1 Introduction

Recovering 3D structure from images is one of the main issues in computer vision. Structure from motion is a fragmented field with varying approaches depending on the models of camera motion and the scene under observation. Approaches can be divided into two major groups depending on the camera displacement between consecutive frames [5]. In the case of video sources with small camera displacement, one of two possible approaches uses spatiotemporal derivatives of the image brightness together with a measure of the camera velocity [2]. The other common approach dealing with small camera displacement relies on tracking of sparse image features, but both approaches are sensitive to noise and it is common to integrate structural information and camera motion over time, using Extended Kalman filtering techniques [1, 4, 16, 17]. In the case of large camera displacement, approaches are traditionally based on geometrical invariants and the extraction and matching of sparse features over successive images and then solving for camera displacement and scene structure [7].

Many of these traditional methods suffer from poor or unstable results near occlusion or motion boundaries. One reason for this is that these methods try to be general and are seldom focused on specific motion or camera models that can model phenomena from occlusion. Most of the methodology in this report has been presented in two conferences proceedings, [18] and [19], and applications are discussed in [20].

1.1 Related work

The work presented in this report aims to maximize the use of connectivity in the temporal domain as much as possible under the condition that the scene is static and that the camera motion is known. To be able to segment small details and achieve high resolution near object borders, a signal model and a projection method have been evaluated using a set of one-dimensional basis functions in the direction of the optical flow. An object behind the foreground is reconstructed using the fact that different parts of the object have been seen in different images in the sequence.

Traditional algorithms in the field of structure from motion have seldom focused on the case when the scene is static compared to the linear motion of the camera. This assumption of the scene may first be seen as a very critical limitation but it is easy to find applications that fulfill this requirement. See section 7 and [20] for examples. Most related in the area of structure from motion is the work by Bolles, Baker, and Marimont [3]. They reduced the dimension for the possible optical flow into planes and they took advantage of the long temporal connectivity of a static scene by using a line matching algorithm. Another related work is one of Wang and Adelson with their layered representation of sequences [21]. They used traditional optical flow estimates which were iteratively decomposed into spatial regions with the same affine transformation of the image velocity. Segmented areas may then be excluded in a reconstruction situation.

2 Scene and camera description

2.1 Epipolar constraint

The scene is assumed to be static compared to the linear motion $v\hat{\mathbf{v}}$ of the camera. All motion in the sequence is therefore introduced by the translating camera. Let the camera be oriented with image plane

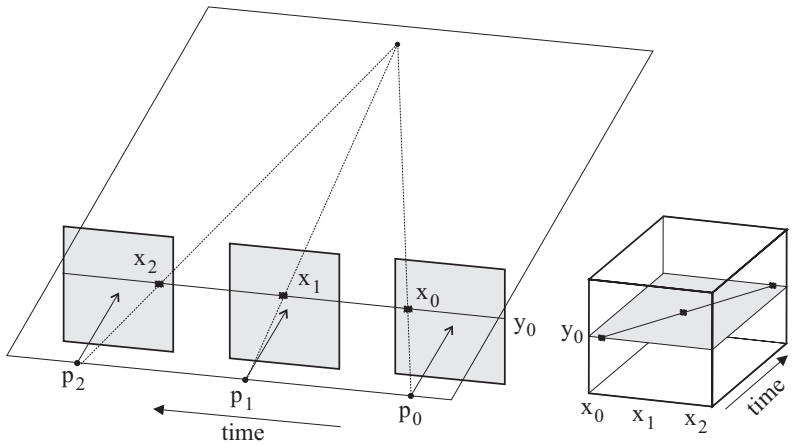

Figure 1: Epipolar geometry from linear ego-motion in a static scene leads to linear spatiotemporal structures.

axes $\hat{\mathbf{x}}_c$ and $\hat{\mathbf{y}}_c$ and assume that $\hat{\mathbf{v}}^T\hat{\mathbf{x}}_c = 1$. The epipolar constraint give that all optical flow in the spatiotemporal volume $s(\mathbf{x})$, where $\mathbf{x} = (x, y, t)^T$, will be located in planes with normals $\hat{\mathbf{n}} = \hat{\mathbf{y}}$ [3]. It is therefore possible to reduce the search for optical flow into planes with constant y-values. If the camera translates with a constant velocity v the resulting optical flow will be linear structures as illustrated in figure 1. Even if the velocity v varies in a known way it is possible to use the methodology since $s(\mathbf{x})$ can be rearranged into $s'(\mathbf{x})$ with linear structures using a global transform [3].

The orientation of the optical flow in the spatiotemporal volume corresponds to the distance to objects. The foreground is continuously described in the spatiotemporal volume but the signal from objects in the background are occluded here and there and exist only in fragmented parts of the spatiotemporal volume. A slice at a specific y-position in a typical spatiotemporal volume is shown in the upper part of figure 2.

3 Signal model

3.1 Signal model for low-altitude surveillance

The main potential application is low-altitude surveillance in forest areas and the signal model is evaluated with that application in mind. A typical object in such environment might be partly observable here and there due to occluding vegetation. Signal correspondence over long time is achieved using a reconstruction method with slowly varying one-dimensional basis functions in the direction of the optical flow. This allows a signal at a specific depth to vary in a smooth way in $s((x, y, t)^T)$ due to small changes in illumination and aspect angles.

Since all analysis, except the final consistency operation described in section 5.3, is performed on planes with constant y values, the introduced signal model $f(\mathbf{x})$ is two-dimensional with $\mathbf{x} = (x, t)^T$. Writing $\mathbf{x}$ means therefore $(x, t)^T$ if it is not explicitly written as $\mathbf{x} = (x, y, t)^T$. The signal model $f(\mathbf{x}),\ \mathbf{x} \in \Omega$, has local properties in the x-dimension and consistent characteristics on projections

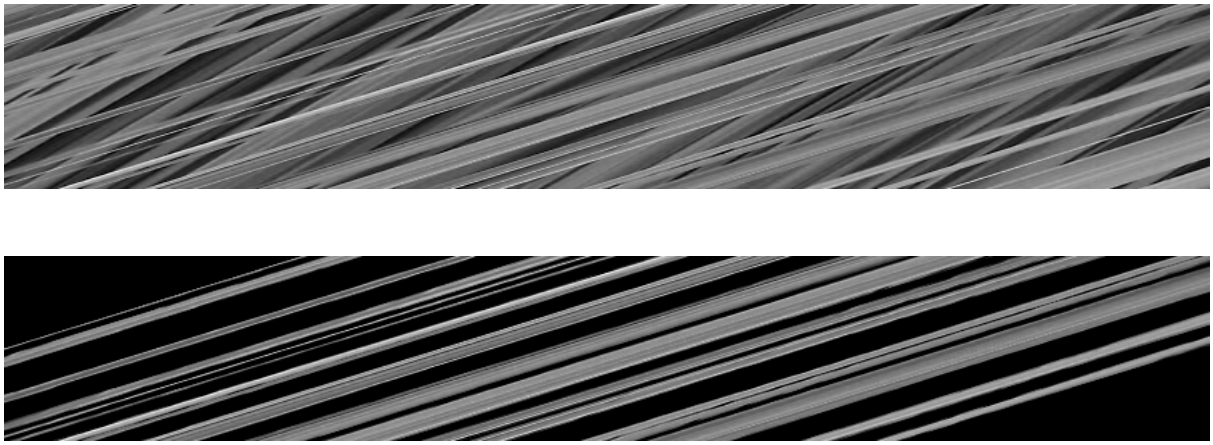

Figure 2: Top: Spatiotemporal slice for $y = y_0$. Bottom: Segmented foreground from the slice above using the presented segmentation algorithm. Top left and bottom right parts of the signal are not used.

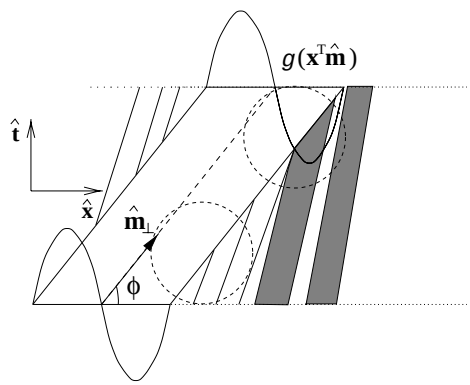

Figure 3: Local signal model in a case of occlusion.

$\mathbf{x}^T\hat{\mathbf{m}} = \alpha$ where α is a constant. An illustration is given in figure 3. A suitable signal model is

$$f(\mathbf{x}) = g(\mathbf{x}^T\hat{\mathbf{m}}) \;\text{ where }\; \hat{\mathbf{m}} = (m_x, m_t)^T. \tag{1}$$

Circles in the figure illustrates support region for traditional convolution kernels for estimation of orientation, and it is obvious that these methods will not in an optimal way utilize the temporal connectivity when the scene is static. In the inner part of an image structure which is wider than the kernel, the result from an orientation estimation will of course be the expected, but there will be great problems with interference between foreground and background at positions near object borders as shown in figure 3.

3.2 Projection model

Let us examine and reconstruct the signal $f(\mathbf{x})$ using a set of M one-dimensional basis-functions $\mathbf{b}_{ki},\ k = 1, \ldots, M$, with an orientation $\hat{\mathbf{p}}_{\perp i}$. The projection vector $\hat{\mathbf{p}}_i = (p_{x_i}, p_{t_i})^T$ is orthogonal to

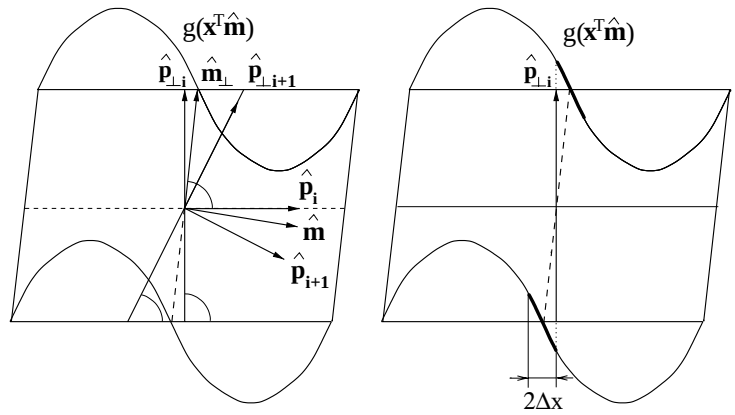

Figure 4: Shifted local signal model.

$\hat{\mathbf{p}}_{\perp i}$. Without any real restriction on the temporal interval it is assumed for simplicity that $t_{min} = -1$ and $t_{max} = 1$. This yields the following restrictions on $\mathbf{x}$

$$\mathbf{x}^T \hat{\mathbf{p}}_i = \beta, \qquad x = \frac{\beta - t\, p_{t_i}}{p_{x_i}} \qquad t \in [-1, 1]. \tag{2}$$

Using this parameterization on the signal model yields

$$\begin{aligned} f(\mathbf{x}) &= g((\frac{\beta - t\, p_{t_i}}{p_{x_i}}, t)(m_x, m_t)^T) \\ &= g(x_0 + t\Delta x) \end{aligned} \tag{3}$$

$$\text{where } \Delta x = \frac{m_t p_{x_i} - m_x p_{t_i}}{p_{x_i}} \text{ and } x_0 = \frac{\beta m_x}{p_{x_i}}.$$

An illustration of this concept of using projection vectors $\hat{\mathbf{p}}$ is given in figure 4. The signal is skewed into a position where $\hat{\mathbf{x}}_c^T \hat{\mathbf{p}}_{\perp i} = 0$ to simplify the illustration. It is now straight forward to examine the one dimensional Taylor expansion of $f(\mathbf{x})$ in equation 3 in the direction of $\hat{\mathbf{p}}_{\perp i}$ i.e. $g(x_0 + t\Delta x)$:

$$\begin{aligned} g(x_0 + t\Delta x) = g(x_0) + g'(x_0) t\Delta x + \\ \frac{g''(x_0)}{2}(t\Delta x)^2 + O((t\Delta x)^3). \end{aligned} \tag{4}$$

If $\hat{\mathbf{m}} = \hat{\mathbf{p}}_i$ it turns out that $\Delta x = 0$ and we only receive $g(x_0)$. This one-dimensional analysis is done using the projection strategy described in section 5.1.2 but first we have to introduce the filtering concept in general.

4 Normalized convolution in general

4.1 Historical perspective

Normalized convolution is a general technique for filtering data with varying spatial reliability. This signal processing scheme uses a signal representation which is decomposed into a certainty part and a

signal part. A formal theory has been developed at the Computer Vision Laboratory by [23, 13, 15] with influences from earlier work on consistency operations [12]. We will now give a short review of the concept in general, based on the formulation in [11], and in section 5 show how it is included as a basic component in this new approach for depth segmentation and scene reconstruction.

4.2 Basis and spaces

Assume that we have a signal $\mathbf{g}$ of dimensionality N and a filter-set $\mathbf{b}_k$ with M filters spanning a linear space $\mathcal{B}$. If $N > M$ the general signal $\mathbf{g}$ with higher dimensionality cannot be reconstructed from its expansion in $\mathbf{b}_k$. A traditional method to find the best expansion coefficients $\mathbf{u} = (u^1, \ldots, u^M)^T$ of $\mathbf{g}$ in $\mathbf{b}_k$ is the least square method which is achieved by minimizing

$$\epsilon = ||\mathbf{g} - \tilde{\mathbf{g}}||^2 \tag{5}$$

$$\text{where } \tilde{\mathbf{g}} = \mathbf{B}\mathbf{u} \text{ and } \mathbf{B} = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_M \\ | & | & \cdots & | \end{bmatrix}.$$

4.3 Weighted Least Mean Square Method (WLMS)

Let us now introduce the matrix $\mathbf{W}$=diag($\mathbf{w}$)=diag($w_1, \ldots, w_N$) which is a diagonal matrix with the spatial reliability for different positions. This is used in the weighted least mean square method when minimizing

$$\epsilon_{\mathbf{W}} = ||\mathbf{W}(\mathbf{g} - \mathbf{B}\mathbf{u})||^2. \tag{6}$$

Minimizing equation 6 gives the following well known solution.

$$\mathbf{u} = (\mathbf{B}^*\mathbf{W}^2\mathbf{B})^{-1}\mathbf{B}^*\mathbf{W}^2\mathbf{g}. \tag{7}$$

Using the notation $<\mathbf{x}, \mathbf{y}>_{\mathbf{Q}} = \mathbf{x}^*\mathbf{Q}\mathbf{y}$ for the $\mathbf{Q}$-weighted scalar product between two signals, where $\mathbf{Q} = \text{diag}(\mathbf{q})$, and the fact that $\mathbf{W}$ is a diagonal matrix, we may rearrange equation 7 in the following way:

$$\mathbf{u} = \begin{bmatrix} <\mathbf{b}_1, \mathbf{b}_1>_{\mathbf{W}^2} \ldots & <\mathbf{b}_1, \mathbf{b}_M>_{\mathbf{W}^2} \\ \vdots & \vdots \\ <\mathbf{b}_M, \mathbf{b}_1>_{\mathbf{W}^2} \ldots & <\mathbf{b}_M, \mathbf{b}_M>_{\mathbf{W}^2} \end{bmatrix}^{-1} \begin{bmatrix} <\mathbf{b}_1, \mathbf{g}>_{\mathbf{W}^2} \\ \vdots \\ <\mathbf{b}_M, \mathbf{g}>_{\mathbf{W}^2} \end{bmatrix}. \tag{8}$$

4.4 WLMS as convolutions

From an implementation point of view, it turns out to be convenient to decompose $\mathbf{W}^2$ into $\mathbf{A}\mathbf{C}$ where $\mathbf{C}$=diag($\mathbf{c}$)=diag($c_1, \ldots, c_N$) contains the spatial certainties of the signal $\mathbf{g}$. The matrix $\mathbf{A}$=diag($\mathbf{a}$) = diag($a_1, \ldots, a_N$) holds a classical windowing function $\mathbf{a}$ in its diagonal containing reliability weights for filter positions. The function $\mathbf{a}$ is called the applicability function and it affects all basis functions in

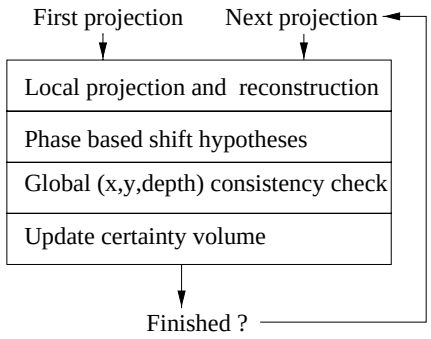

Figure 5: Flow-chart over depth segmentation algorithm.

the same way. Rewriting equation 8 using this decomposition and linear properties of the scalar product results in the following formulation in terms of standard scalar products or convolutions.

$$\mathbf{u} = \begin{bmatrix} <\mathbf{A}\mathbf{b}_1 \circ \mathbf{b}_1^*, \mathbf{c}> \ldots & <\mathbf{A}\mathbf{b}_1 \circ \mathbf{b}_M^*, \mathbf{c}> \\ \vdots & \vdots \\ <\mathbf{A}\mathbf{b}_M \circ \mathbf{b}_1^*, \mathbf{c}> \ldots & <\mathbf{A}\mathbf{b}_M \circ \mathbf{b}_M^*, \mathbf{c}> \end{bmatrix}^{-1} \begin{bmatrix} <\mathbf{A}\mathbf{b}_1, \mathbf{C}\mathbf{g}> \\ \vdots \\ <\mathbf{A}\mathbf{b}_M, \mathbf{C}\mathbf{g}> \end{bmatrix} \tag{9}$$

A convention $\mathbf{x} \circ \mathbf{y} = \mathrm{diag}(\mathbf{x})\mathbf{y}$ for point-wise vector multiplication is used. Examination of equation 9 shows that we need $M(M+1)/2$ more filters $\mathbf{A}\mathbf{b}_k \circ \mathbf{b}_l^*$ compared to standard convolution where we only need M $\mathbf{A}\mathbf{b}_k$ filters. The structure of equation 9 also shows that the filters on the left-hand side in the scalar products is unaffected of any variation in the signal $\mathbf{g}$ or the certainty $\mathbf{c}$. Implementation in terms of convolution to perform localized weighted least mean square is therefore straightforward.

4.5 Application for normalized convolution

Handling artifacts near image borders is an application where normalized convolution is preferable to traditional techniques as zero-padding or border expansion [23]. Other applications where the use of normalized convolution have been successful are interpolation [13], irregularly sampled data [15], stereo and focus-of-attention [22], frequency estimation [14] and spatiotemporal filtering [11].

5 Segmentation algorithm

From equation 6 we observed that normalized convolution was equivalent to finding the best weighted projection $\tilde{\mathbf{g}}$ of a signal $\mathbf{g}$ in a specific base $\mathbf{B}$ using a spatial certainty weight $\mathbf{c}$. One of the main ideas in this report is the use of a spatiotemporal certainty volume $c(\mathbf{x})$ with the same dimensions as the input spatiotemporal volume $s(\mathbf{x})$, and then use $c(\mathbf{x})$ as a "blackboard" for remembering already segmented image structures.

The basic parts of the segmentation algorithm are illustrated in figure 5. A spatiotemporal slice $s((x, y_0, t)^T)$ is first decomposed into a number of local overlapping temporal windows s_w, $w = 1, \ldots, N$ as shown in figure 6. A typical size of a temporal local window is from 7 to 21 frames.

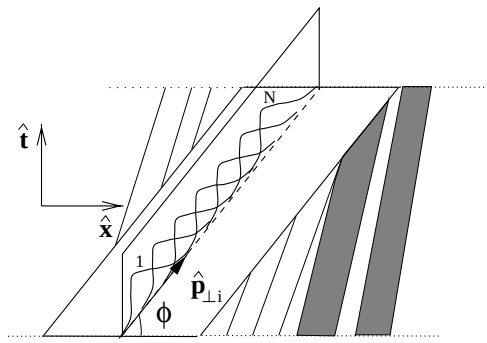

Figure 6: Local models in different time slices with projection orientation $\hat{\mathbf{p}}_{\perp i}$ are connected to one global model.

One temporal slice from the same sequence as figure 2 is shown at top in figure 8. Each local window generates a local model which is used for global signal correspondence.

The certainty volume $c(\mathbf{x})$ is initiated to one but when a region $\mathbf{x} \in R_0$ is confirmed as foreground, it is receded from the sequence by setting its certainty value $c(\mathbf{x}) = 0$ for $\mathbf{x} \in R_0$. This strategy minimizes the occluding influence of the foreground when analyzing underlying image structures in the background of the scene. A natural approach is therefore a batch procedure where we start with searching for image structures in the foreground, eliminate their occluding influence, and then proceed with the search for correct depth for structure on larger distances. The use of normalized convolution is therefore well motivated since the dependence between basis functions is continuously changing under the segmentation process due to variations in the certainty volume $c(\mathbf{x})$.

If the reconstruction error is low we are probably near the correct orientation. The next step is therefore to fuze neighboring structures with low reconstruction error into a coherent signal representation on a global level. Only structures with neighbouring depth-values are connected using this strategy and this makes it possible to track very narrow image structures which can be seen in the segmented spatiotemporal slice in the bottom part of figure 2. The global signal is evaluated using the phase-based method described in section 5.2.1. Since all processing is performed in 2D with constant y -values, a final consistency operation in the $(x, y, depth)$-domain is then performed using a 3D-smoothing filter as described in section 5.3.

5.1 Basis functions for projection analysis

5.1.1 Projection vectors

The number of needed projection vectors will depend on the number of frames, the minimum and maximum distances in the scene, and the highest frequency component in the scene. A large number of projection vectors will be needed if the highest frequency component is the Nyqvist frequency.

5.1.2 Basis functions

Assume for the moment that we are analyzing a signal $f(\mathbf{x})$ in one temporal slice s_w using a set of one-dimensional basis functions $\mathbf{b}_{ki},\ k = 1, \ldots, M$, with an orientation $\hat{\mathbf{p}}_{\perp i}$. Three basis functions are used

for reconstruction; a constant function $\mathbf{b}_{0i}$, a linear function $\mathbf{b}_{1i}$, and a quadratic function $\mathbf{b}_{2i}$ without DC-term. The applicability equals to one in the central part of the window but are smoothly going to zero near window borders using a quadratic spline function. Since these local models are overlapping the applicability for all models should sum to a constant. Some temporal positions will otherwise be in favour and all positions are supposed to be equally important. Using the same parameterization as in equation 2 the scalar-product can be written as

$$\begin{aligned} &< \mathbf{b}_{ki}, \mathbf{b}_{li} >_{\mathbf{A}_i\mathbf{C}_i} \quad \text{where} \;\; \mathbf{C}_i = \text{diag}(\mathbf{c}_i) \\ &\text{and} \;\; \mathbf{c}_i = c(\frac{\beta - t\; p_{t_i}}{p_{x_i}}, y_0, t) \;\; t \in [-1, 1]^{\cdot} \end{aligned} \tag{10}$$

The filter-set is orthogonal when $\mathbf{C}_i = \mathbf{I}$ since the quadratic function $\mathbf{b}_{2i}$ has zero DC-level. This can be written as $< \mathbf{b}_{ki}, \mathbf{b}_{li} >= \delta^{kl}$. Let us expand equation 4 in our chosen basis $\mathbf{B}_i$ as illustrated in figure 7.

$$\mathbf{B}_i = \begin{bmatrix} | & | & | \\ \mathbf{b}_{0i} & \mathbf{b}_{1i} & \mathbf{b}_{2i} \\ | & | & | \end{bmatrix}$$

$$\begin{aligned} < \mathbf{b}_{0i}, g(x_0 + t\Delta x) >_{\mathbf{A}_i\mathbf{C}_i} = g(x_0) < \mathbf{b}_{0i}, 1 >_{\mathbf{A}_i\mathbf{C}_i} + \\ \frac{g''(x_0)}{2}\Delta x^2 < \mathbf{b}_{0i}, t^2 >_{\mathbf{A}_i\mathbf{C}_i} \end{aligned} \tag{11}$$

$$< \mathbf{b}_{1i}, g(x_0 + t\Delta x) >_{\mathbf{A}_i\mathbf{C}_i} = \; g'(x_0)\Delta x < \mathbf{b}_{1i}, t >_{\mathbf{A}_i\mathbf{C}_i} \tag{12}$$

$$< \mathbf{b}_{2i}, g(x_0 + t\Delta x) >_{\mathbf{A}_i\mathbf{C}_i} = \; \frac{g''(x_0)}{2}\Delta x^2 < \mathbf{b}_{2i}, t^2 >_{\mathbf{A}_i\mathbf{C}_i} \tag{13}$$

There are three observations that should be mentioned here. Primarily, the scalar-products with $\mathbf{b}_{1i}$ and $\mathbf{b}_{2i}$ are both small when $\hat{\mathbf{p}}_i$ is close to $\hat{\mathbf{m}}$. Secondly, one is odd and the other is even and thirdly that these scalar-products complement each other since $g'(x_0)$ and $g''(x_0)$ do not usually share the same zero-crossings. These properties are useful for interpolation between fixed projections.
Given a projection vector $\hat{\mathbf{p}}_i$ we will receive a hypothetical signal

$$\tilde{g}_i = \mathbf{B}_i\mathbf{u}_i$$

using the weighted LMS solution

$$\mathbf{u}_i = (\mathbf{B}_i^*\mathbf{A}_i\mathbf{C}_i\mathbf{B}_i)^{-1}\mathbf{B}_i^*\mathbf{A}_i\mathbf{C}_i g$$

with $\mathbf{C}_i$ defined in equation 10. The corresponding local reconstruction error is ϵ_i

$$\epsilon_i = ||\mathbf{A}_i\mathbf{C}_i(\mathbf{g}_i - \mathbf{B}_i\mathbf{u}_i)||^2. \tag{14}$$

Only projections with low reconstruction error ϵ_i will be accepted as segmented regions if they also fulfill the requirements of linear phase. Reconstruction of a signal using different numbers of basis functions is illustrated in figure 8.

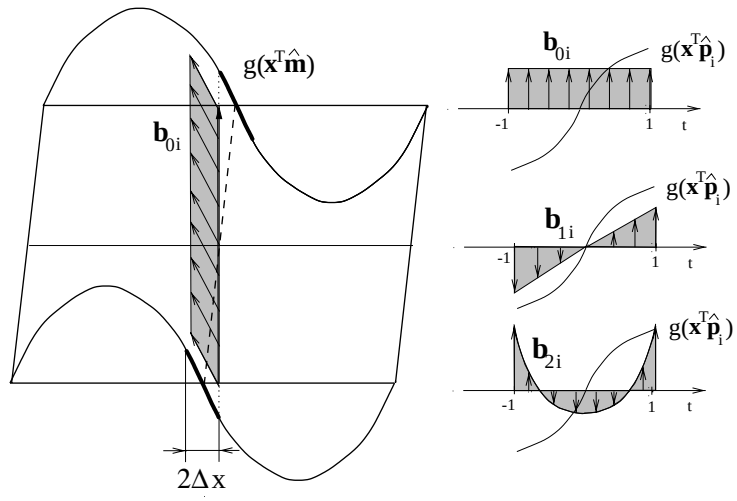

Figure 7: Basis functions and scalar-products for depth channel i.

5.2 Global models based on linear phase

5.2.1 Hierarchical models

Neighboring structures with low local reconstruction error ϵ_i are fused into a coherent signal representation on a global level. Only structures with neighboring depth-values are connected using this strategy. This global signal is evaluated using a phase-based method with much in common with phase-based disparity estimation in stereo vision [22].There are three major reasons why a phase-based method is used in this part of the segmentation. Primarily, phase is not so sensitive to variations in illumination when the camera moves, since it is invariant to signal energy. Secondly, non-frontally viewed surfaces are scaled and compressed as the perspective changes when the camera moves and it is well-known that phase has desirable properties in scale-space [8, 9]. Thirdly, phase based approaches give sub-pixel properties of shift estimates since local phase is equivariant with spatial position. The influence of spatial positions with high reconstruction errors on the phase estimate is rejected by using normalized convolution where the certainties $\mathbf{c}^{ph}$ for these positions is zero. High spatial resolution near object borders is therefore achieved and only neighboring structures with similar depth will support each other.

5.2.2 Global projection analysis

In order to achieve this degree of resolution we first have to combine local models into global models. Each temporal slice $s_w(\mathbf{x})$ generates a model $\tilde{g}_{iw}$ for a hypothetical orientation $\hat{\mathbf{p}}_{\perp i}$ of optical flow. The same projection strategy is now used on all center-positions of local models in order to generate a global model in a hierarchical way. This is illustrated in figure 6. Given a set of center-positions of local models $\tilde{g}_{iw}, w = 1, \ldots, N$, we will receive a hypothetical global signal

$$\tilde{g}^g_{i_j} = \mathbf{B}_{i_j} \mathbf{u}^g_{i_j} \tag{15}$$

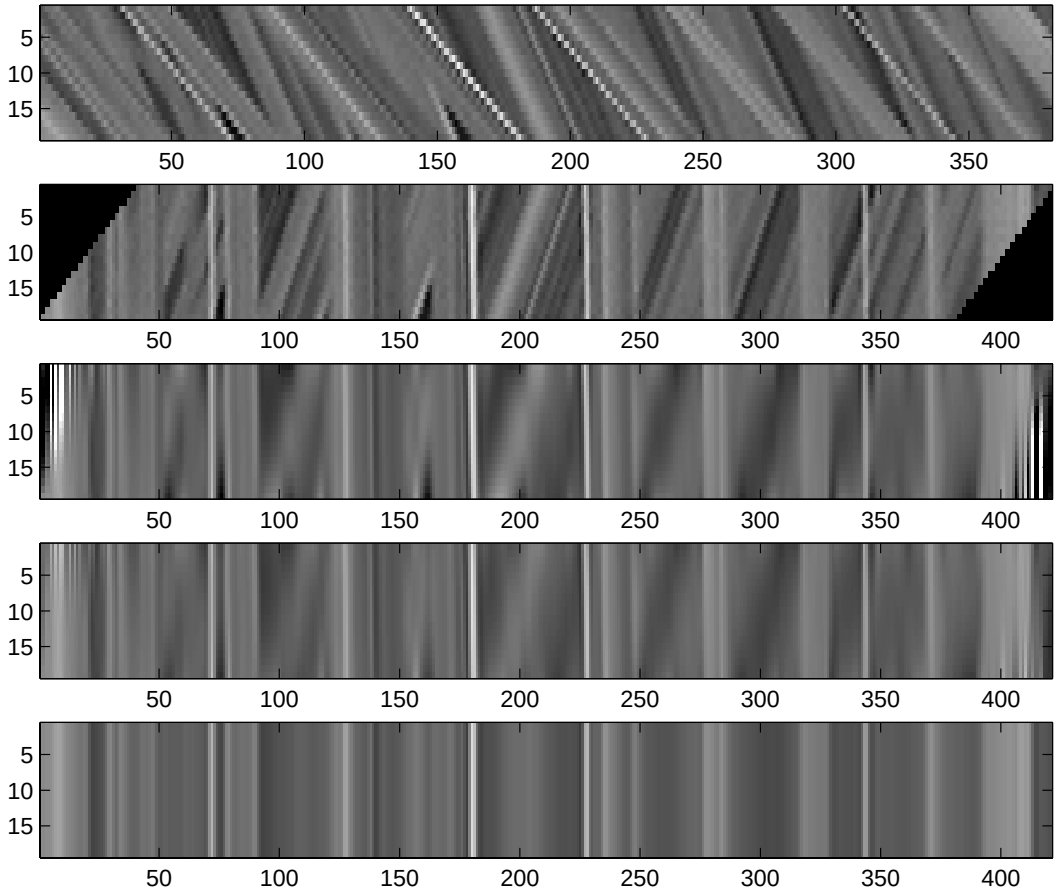

Figure 8: Local reconstruction using normalized convolution. Top image: Input x-t slice. Second image: Input slice shifted corresponding to depth channel i. Third image: Local reconstruction using all three basis functions $\{\mathbf{b}_{0i}, \mathbf{b}_{1i}, \mathbf{b}_{2i}\}$. Fourth image: Local reconstruction using only basis functions $\{\mathbf{b}_{0i}, \mathbf{b}_{1i}\}$. Bottom image: Local reconstruction using only the DC base function $\mathbf{b}_{0i}$.

using the weighted LMS solution from equation 7

$$\mathbf{u}^g_{i_j} = (\mathbf{B}^*_{i_j}\mathbf{A}_{i_j}\mathbf{C}_{i_j}\mathbf{B}_{i_j})^{-1}\mathbf{B}^*_{i_j}\mathbf{A}_{i_j}\mathbf{C}_{i_j}[\tilde{g}_{i1}\dots\tilde{g}_{iN}]^T \tag{16}$$

where $\mathbf{C}_{i_j}$ is defined as $\mathbf{C}_{i_j} = \text{diag}(\mathbf{c}_{iw})$. Sub-index i_j indicates that a higher angular resolution between consecutive projection-orientations is needed on the global level. The local certainty $\mathbf{c}_{iw}$ is proportional to the spatial support for a local model, i.e. the tendency to occlusion of already segmented structures. The corresponding global reconstruction error is

$$\epsilon^g_{i_j} = ||\mathbf{A}_{i_j}\mathbf{C}_{i_j}([\tilde{g}_{i1}\dots\tilde{g}_{iN}]^T - \mathbf{B}_{i_j}\mathbf{u}^g_{i_j})||^2. \tag{17}$$

The global certainty $\mathbf{c}^{ph}_{i_j}$ for a global projection i_j is based on a non-linear mapping of the sum of local reconstruction errors ϵ_{iw} and the global reconstruction error $\epsilon^g_{i_j}$. Since these errors will depend

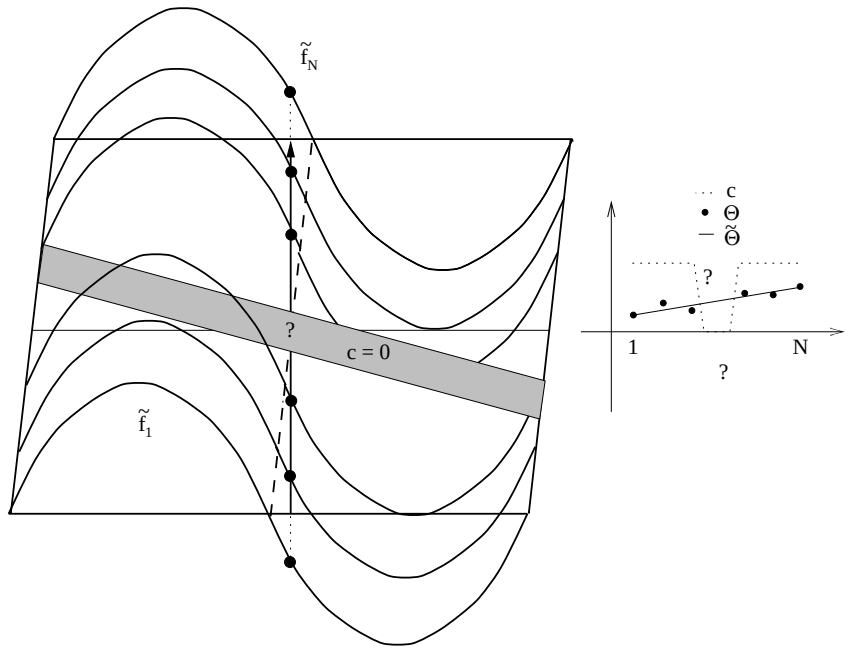

Figure 9: Linear phase model $\tilde{\Theta}$ using normalized convolution.

on the amplitude of the signal and increase with higher frequency, a normalized error is used which is compensated for the local spatial derivative in the x-dimension of the hypothetical signal. Limited variations in signal energy due to changes in illumination and specular reflections are modulated in the analysis since the linear and quadratic functions in the global base $\mathbf{B}_{i_j}$ allow the signal to vary in a smooth way over time.

5.2.3 Phase-based disparity

The analytical signal of the center of each local model is approximated in the x-dimension with $\mathbf{q}_k$, responses of a set of one-dimensional quadrature-filters with different center frequencies ρ_k. The theoretical background of this model is that a spatial shift of a signal corresponds to a phase shift of its analytical signal [10, 22]. The strategy has much in common with multi-baseline stereo approaches. Phase Θ and instant frequency $\dot{\Theta} = \frac{d\Theta}{dx}$ are estimated from each $\mathbf{q}_k$ and a linear phase model $\tilde{\Theta}_{ki_j}$ is evaluated using the weighted LMS solution from equation 7. A global base $\mathbf{D}$ composed of a linear and a constant function and an constant applicability function are used for this purpose.

$$\tilde{\Theta}_{ki_j} = \mathbf{D}_{i_j}\mathbf{p}_{ki_j} \qquad \mathbf{D}_{i_j} = \begin{bmatrix} | & | \\ \mathbf{b}_{0i_j} & \mathbf{b}_{1i_j} \\ | & | \end{bmatrix} \tag{18}$$

$$\mathbf{p}_{ki_j} = (\mathbf{D}^*_{i_j}\mathbf{C}^{ph}_{i_j}\mathbf{D}_{i_j})^{-1}\mathbf{D}^*_{i_j}\mathbf{C}^{ph}_{i_j}\Theta_{ki_j} \tag{19}$$

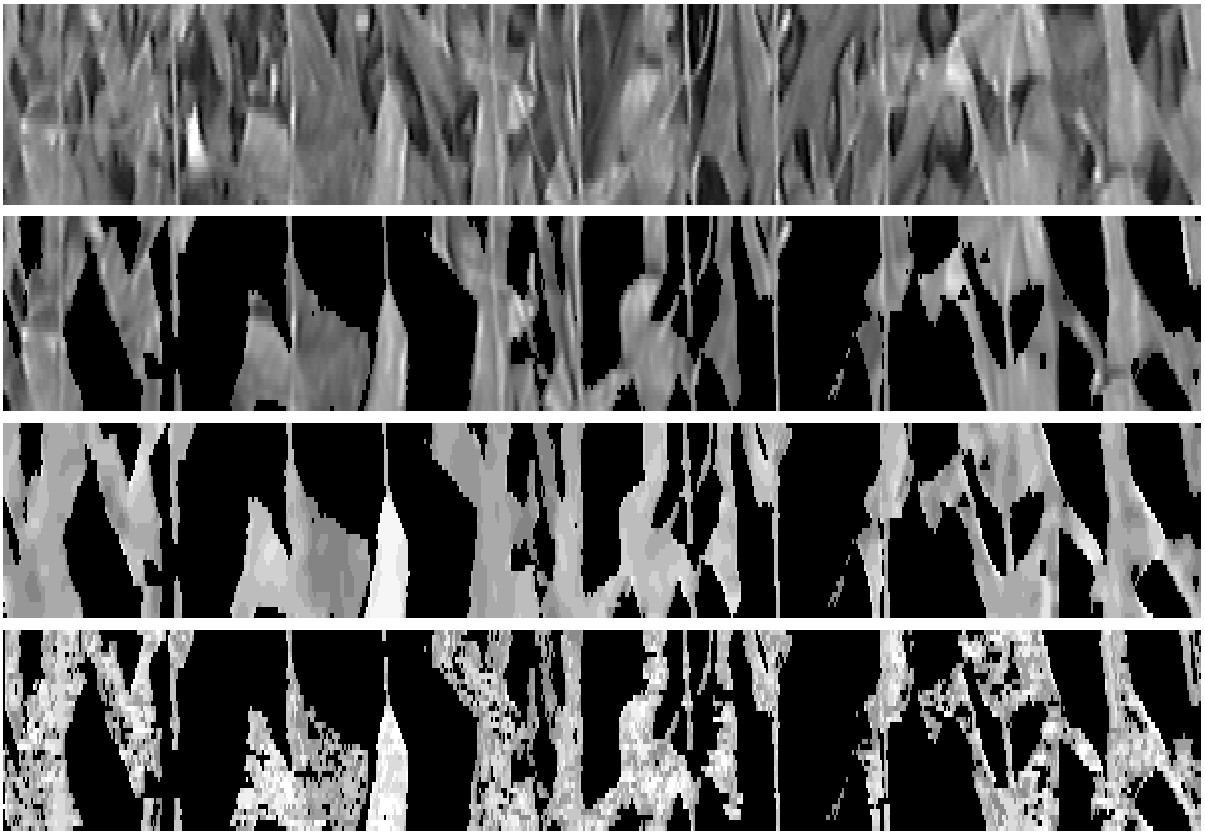

Figure 10: Top: Part from one frame in the test sequence. Second: Reconstructed foreground using local consistency operation. Third: Grey-scale values indicate depth variations in the reconstructed foreground using local consistency operation. Bottom: Depth variations without local consistency operation.

$$\epsilon^{ph}_{ki_j} = ||\mathbf{C}^{ph}_{i_j}(\Theta_{ki_j} - \mathbf{D}_{i_j}\mathbf{p}_{ki_j})||^2 \tag{20}$$

The global certainty $\mathbf{C}^{ph}_{i_j}$ is defined as $\mathbf{C}^{ph}_{i_j} = \text{diag}(\mathbf{c}^{ph}_{i_j})$. The certainty approach will eliminate the influence of an occluding foreground and the linear phase model relies on image components beside the already segmented foreground. This is illustrated in figure 9. The corresponding residual $\epsilon^{ph}_{ki_j}$ is used to evaluate the consistency of linear phase models for each spatial position x, phase filter k, and temporal projection channel i_j. Three quadrature filters with center frequencies $\rho_1 = \pi/2$, $\rho_2 = \pi/4$, and $\rho_3 = \pi/8$ and a relative bandwidth of $\sqrt{2}$ have been used in all experiments in this report.

5.3 Spatial consistency in the $(x, y, depth)$-domain

Each quadrature filter k generates a shift hypothesis $\Delta\xi_{ki_j}$. With every hypothesis follows a certainty $c^{hyp}_{ki_j}$, a "probability density", which combines spatial covering for the model, tendency to linear phase,

and low reconstruction error. These densities are smoothed in the $(x, y, depth)$-domain using a three-dimensional Gaussian with $\sigma = 1.5$. A hypothesis has only support in regions with low reconstruction error and the traditional tendency to enlarge foreground regions in x- or y -dimension will therefore be small, see figure 11 top right. Only image parts close in $(x, y, depth)$-domain will have influence on each other. This averaging of hypotheses in $(x, y, depth)$-domain results in high spatial resolution and smoother surfaces which is illustrated in figure 10. The second image shows the segmented and reconstructed foreground of the first image using the consistency operation. Third and bottom image visualize the distribution of depth-values within the foreground distance and show that this smoothing operation is an indispensable part to the overall performance.

5.4 Reconstruction filters

Different shapes of the applicability functions are preferable for global reconstruction of occluded image-parts depending on the application and noise-level. The use of an applicability function equal to one, which is equivalent with letting all non-occluded frames be equal important, will be very fragile to small errors in the estimates depth but will on the other hand decrease the noise level to a minimum if the depth estimate is correct. On the other hand, if the applicability is very narrow, the result will not be so sensitive to small errors in depth estimate. The results in figure 11 are achieved with a constant applicability function on the global level.

6 Results

The test-sequence consists of 135 frames with a camouflage net as foreground, another net as background, and a soldier doll in between, being observable only here and there. One frame in the middle of the sequence is shown in figure 11 (top). The segmented and reconstructed foreground is shown in figure 11 (third). It is then possible to manually reconstruct the depth channel corresponding to the doll without the hiding influence of the foreground (second). Depth variations in the segmented foreground are also shown (bottom). More applications and results are presented in [20].

7 Applications

The main potential application is low-altitude surveillance in forest areas and the signal model is evaluated with that application in mind. More applications and scenarios is discusesed in [20] and some of them are summerized here:

- Reconnaissance from low altitude with a traditional aeroplane or an UAV (Unmanned Aerial Vehicle) as platform.
- Passive range measurement using egomotion.
- Superresolution, i.e. increasing the detail resolution of the processed image compared to the images in the time sequence.
- Noise reduction using depth information, see section 5.4.
- Increasing the effective fill-factor for infrared focal plane array cameras (staring IR-cameras)
- Multi-baseline stereo, i.e. fusion of images from different view-points.

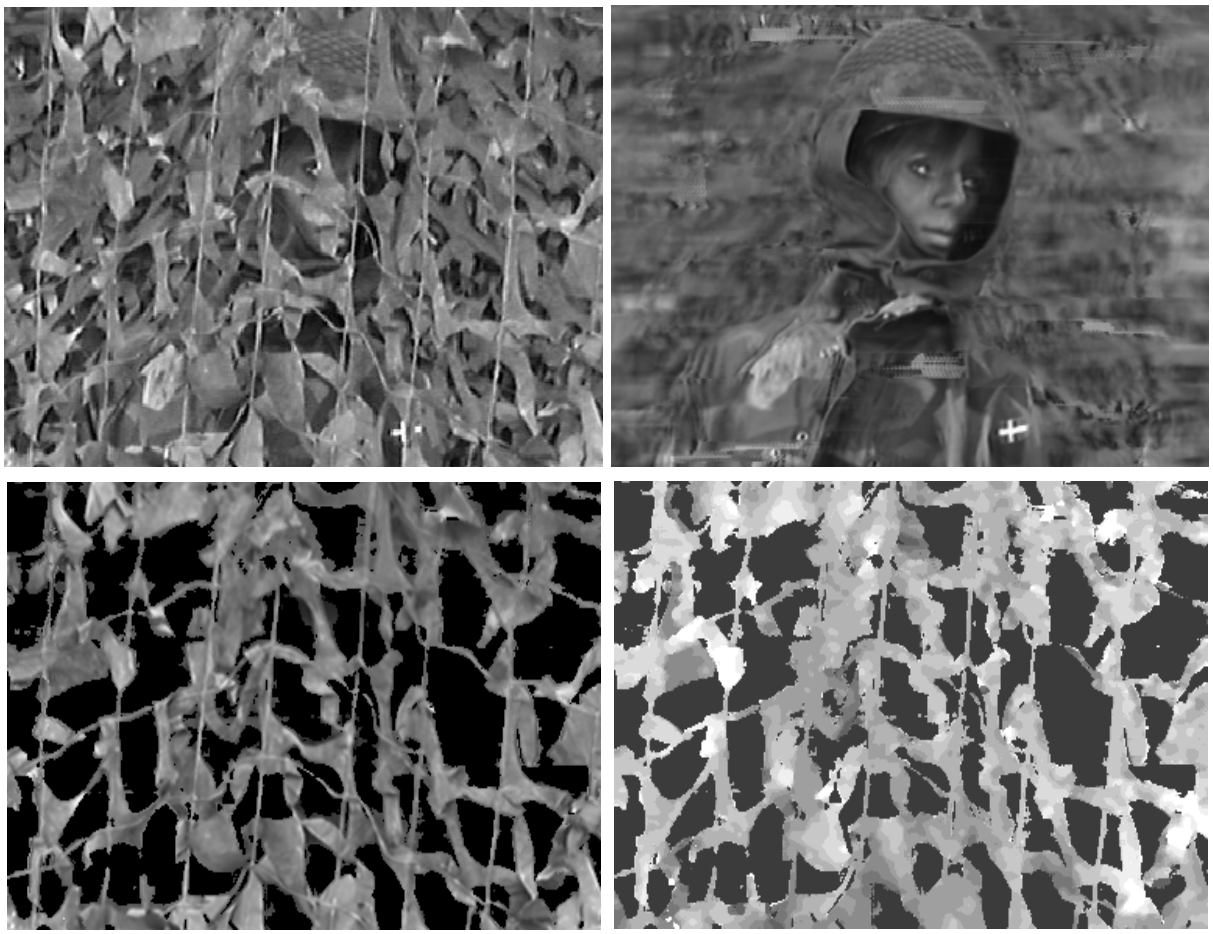

Figure 11: Top left: Frame in the middle of the sequence. Top right: Reconstructed image for the depth of interest. Bottom left: Segmented foreground. Bottom right: Depth variations in the segmented foreground.

8 Discussion

Recovering 3-dimensional structure from image sequences is one of the main issues in the field of computer vision. Structure from ego-motion in a dynamic environment is a very difficult task. Constraints on the scene or on the camera motion are needed to simplify the analysis. Many traditional structure from motion algorithms and spatiotemporal segmentation algorithms rely on the assumption that the scene is composed of planar surfaces and use spatial grouping algorithms searching for affine deformable velocity fields [21, 6]. There is no such assumption of the scene structure made in this work. The only constrain on the scene in this report is that it should be static compared to the motion of the sensor. All grouping mechanisms are focused on temporal connectivity, and correct information of the camera motion is therefore critical. The implemented algorithm is focused on a specific simple camera motion, a translation in x-direction of the image-plane.

A specific feature to mention of the presented algorithm is the possibility to "see behind" the foreground and receive a reconstructed non-occluded image. It has however turned out to be very complicated to get reliable depth estimates from image parts behind the occluding foreground in the test sequence. This is probably depending on the fact that only seven local models were used. A higher number of local models with shorter time support is probably a better solution since we will then always achieve at least a few non occluded models to match. A future research project will include a motion calibrating strategy within the segmentation concept.

Acknowledgments

This work was sponsored by Innovationsfonden, Defence Research Establishment, Sweden. The authors want to thank all people at Innovationsfonden for their belief in and support to the project "Adaptive Reconstruction using Multiple Views".

References

[1] Ali Azerbayejani and Alex P. Pentland. Recursive estimation of motion, structure. and focal length. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-17(6), 1995.

[2] J.L. Barron, A.D Jepson, and J.K. Tsotsos. The feasibility of motion and structure from noisy time-varying image velocity information. *International Journal of Computer Vision*, 3(5):239–269, 1990.

[3] R.C. Bolles, H.H. Baker, and D.H. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision*, 1(1):7–55, 1987.

[4] T.J. Broida, S. Chandrashekhar, and R. Chellappa. Recursive 3-d motion estimation from a monocular image sequence. *IEEE Transactions on Aerospace and Electronic Systems*, 26(4), 1990.

[5] Francois Chaumette, Samia Boukir, Patrick Bouthemy, and Didier Juvin. Structure from controlled motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-18(5), 1996.

[6] G. Farnebäck. Motion-based segmentation of image sequences using orientation tensors. In *Proceedings of the SSAB Symposium on Image Analysis*, Stockholm, March 1997. SSAB.

[7] O. D. Faugeras and S. Maybank. Motion from point matches: Multiplicity of solutions. *International Journal of Computer Vision*, 3(4):225–246, 1990.

[8] D. J. Fleet and A. D. Jepson. Stability of phase information. In *Proceedings of IEEE Workshop on Visual Motion*, pages 52–60, Princeton, USA, October 1991. IEEE, IEEE Society Press.

[9] D. J. Fleet, A. D. Jepson, and M. R. M. Jenkin. Phase-based disparity measurement. *CVGIP Image Understanding*, 53(2):198–210, March 1991.

[10] G. H. Granlund and H. Knutsson. *Signal Processing for Computer Vision*. Kluwer Academic Publishers, 1995. ISBN 0-7923-9530-1.

[11] J. Karlholm. Efficient spatiotemporal filtering and modelling, June 1996. Thesis No. 562, ISBN 91–7871–741–8.

[12] H. Knutsson, G. H. Granlund, and J. Bigun. Apparatus for detecting sudden changes of a feature in a region of an image that is divided into discrete picture elements. US-Patent 4.747.150, 1988, 1988.

[13] H. Knutsson and C-F. Westin. Normalized and Differential Convolution: Methods for Interpolation and Filtering of Incomplete and Uncertain Data. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, New York City, USA, June 1993. IEEE.

[14] H. Knutsson, C-F. Westin, and G. H. Granlund. Local Multiscale Frequency and Bandwidth Estimation. In *Proceedings of IEEE International Conference on Image Processing*, Austin, Texas, November 1994. IEEE. Cited in Science: Vol. 269, 29 Sept. 1995.

[15] H. Knutsson, C-F. Westin, and C-J. Westelius. Filtering of Uncertain Irregularly Sampled Multidimensional Data. In *Twenty-seventh Asilomar Conf. on Signals, Systems & Computers*, Pacific Grove, California, USA, November 1993. IEEE.

[16] L. Matthies, T. Kanade, and R. Szeliski. Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision*, 3:210–238, 1989.

[17] Philip F. McLauchlan and David W. Murray. Active camera calibration for a head-eye platform using the variable state-dimension filter. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-18(1), 1996.

[18] M. Ulvklo, H. Knutsson, and G. Granlund. Adaptive reconstruction using multiple views. In *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation*, Tucson, Arizona, USA, April 1998. IEEE.

[19] M. Ulvklo, H. Knutsson, and G. Granlund. Depth segmentation and occluded scene reconstruction using ego-motion. In *Proceedings of the SPIE Conference on Visual Information Processing*, Orlando, Florida, USA, April 1998. SPIE.

[20] M. Ulvklo and M. Uppsäll. Adaptive reconstruction using multiple views - results and applications. User Report FOA-R-97-00571-408–SE ISSN 1104-9154, Defense Research Establishment, Departement of Sensor Technology, S-581 11 Linköping, Sweden, November 1997.

[21] J. Y. A. Wang and E. H. Adelson. Layered representation for motion analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 361–366, June 1993.

[22] C-J. Westelius. *Focus of Attention and Gaze Control for Robot Vision*. PhD thesis, Linköping University, Sweden, S–581 83 Linköping, Sweden, 1995. Dissertation No 379, ISBN 91–7871–530–X.

[23] C-F. Westin. *A Tensor Framework for Multidimensional Signal Processing*. PhD thesis, Linköping University, Sweden, S–581 83 Linköping, Sweden, 1994. Dissertation No 348, ISBN 91–7871–421–4.