Wide-baseline Stereo from Multiple Views: a Probabilistic Account

Christoph Strecha, Rik Fransens, Luc Van Gool
ESAT-PSI, University of Leuven, Belgium
{christoph.strecha,rik.fransens,luc.vangool}@esat.kuleuven.ac.be

Abstract

This paper describes a method for dense depth reconstruction from a small set of wide-baseline images. In a wide-baseline setting an inherent difficulty which complicates the stereo-correspondence problem is self-occlusion. Also, we have to consider the possibility that image pixels in different images, which are projections of the same point in the scene, will have different color values due to non-Lambertian effects or discretization errors. We propose a Bayesian approach to tackle these problems. In this framework, the images are regarded as noisy measurements of an underlying 'true' image-function. Also, the image data is considered incomplete, in the sense that we do not know which pixels from a particular image are occluded in the other images. We describe an EM-algorithm, which iterates between estimating values for all hidden quantities, and optimizing the current depth estimates. The algorithm has few free parameters, displays a stable convergence behavior and generates accurate depth estimates. The approach is illustrated with several challenging real-world examples. We also show how the algorithm can generate realistic view interpolations and how it merges the information of all images into a new, synthetic view.

1. Introduction

During the last few years more and more user-friendly solutions for 3D modeling have become available. Techniques have been developed [5] to reconstruct scenes in 3D from video or images as the only input. The strength of these shape-from-video techniques lies in the flexibility of the recording, the wide variety of scenes that can be reconstructed and the ease of texture extraction.

In this paper, we present a method for dense depth reconstruction from a small set of wide-baseline images. Wide-baseline stereo has become possible thanks to recent developments in the automatic extraction of local, viewpoint invariant features [11, 7]. It is a promising avenue for 3D reconstruction for a number of reasons. First of all, modern digital cameras have very high resolutions and are capable of recording detailed, high-quality imagery. Secondly, using a limited amount of images considerably speeds up the reconstruction process. Also, the wide-baseline setting carries the promise of more accurate reconstructions, because it generates larger, hence more reliably measurable, disparities in the images. On the other hand, there is a price to pay for these advantages. Inherent to the wide-baseline setting is the problem of occlusion, which means that not all parts of the scene, which are visible in a particular image, are also visible in the other images. Also, because of the large differences in viewpoints, we have to consider the possibility that image pixels in different images, which are projections of the same point in the scene, will have different color values due to non-diffuse reflections.

Stereo matching has been studied mainly in the context of small baseline stereo and for almost fronto-parallel planes. There exist many algorithms based on a diversity of concepts, a recent comparative study is presented in [8]. Some algorithms combine multiple views, often taken from all around the object. Examples are voxel carving [6], photo hulls [9] and level sets [4]. Several of these approaches use a discretized volume and restrict possible depth values to a predefined accuracy. This is not the case for pixel-based PDE approaches [1, 10], which do not need 3D discretization and compute a continuous depth for every pixel. For large images with fine details, it is questionable if volume based algorithms can combine reasonable speed and memory requirements with high accuracy.

Here, the wide-baseline stereo problem is addressed from a probabilistic point of view. We primarily focus on the occlusion part of the problem, that is to say, we assume that we are dealing with mainly diffuse objects, and all deviations from this model are caught by a noise term. In the proposed algorithm, each input image is regarded as a noisy measurement of an unknown image irradiance or 'true' image, which is estimated as part of the optimization problem. This image combines the information from all other views and can be used as a texture map for the final reconstruction. Furthermore, because the image is in essence a model for all views, this also allows view interpolation, i.e. generate images from a camera position which is not present in the input set. View interpolation has recently been presented in a Bayesian framework [3], where a prior probability is defined by means of so-called *texture priors*. The principle difference with our approach is that we introduce prior

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)
1063-6919/04 $20.00 © 2004 **IEEE**

knowledge in terms of depth smoothness. We will discuss the merits and liabilities of both methods in section 3.

2. Problem Statement

Suppose we are given N images $\mathcal{I}_i$, $i=1\ldots N$, which associate a 2D-coordinate $\mathbf{x}$ with a color value $\mathcal{I}_i(\mathbf{x})$. If we are dealing with color images, this value is a 3-vector and for intensity images it is a scalar. The images are taken with a set of cameras of which we know the internal and external calibrations. Our aim is to estimate a set of depth-maps $\mathcal{D}_i$ which assign a depth-value $\mathcal{D}_i(\mathbf{x})$ to all pixel locations in images $\mathcal{I}_i$. These depth-maps are relative to the positions and view directions of the cameras, and can later be integrated into a single model. In this N-view stereo problem, the information from all images will contribute to the computation of each of the maps $\mathcal{D}_i$. In the remainder of the paper, we will describe how to compute $\mathcal{D}_1$ where we take $\mathcal{I}_1$ as a reference view, without loss of generality.

Given the camera calibrations and a depth value $\mathcal{D}_1(\mathbf{x}_1)$ for a position $\mathbf{x}_1$ in $\mathcal{I}_1$, it is easy to compute the corresponding pixel location in the i^{th} image:

$$\lambda_i \mathbf{x}_i^h = \mathcal{D}_1(\mathbf{x}_1)\mathbf{K}_i\mathbf{R}_i^T\mathbf{R}_1\mathbf{K}_1^{-1}\mathbf{x}_1^h + \mathbf{K}_i\mathbf{R}_i^T(\mathbf{t}_1 - \mathbf{t}_i)\ , \quad (1)$$

where $\mathbf{K}_i$, $\mathbf{R}_i$ and $\mathbf{t}_i$ are the camera matrix, rotation and translation of the i^{th} camera, respectively. Superscript h denotes that the vector is expressed in homogeneous coordinates. The 2D point $\mathbf{x}_i$ is easily derived from (1) by dividing out the homogeneous factor. We will denote the overall mapping as $\mathbf{x}_i = l_i(\mathbf{x}_1, \mathcal{D}_1(\mathbf{x}_1))$, or even shorter as $\mathbf{x}_i = l_i(\mathbf{x}_1)$.

Faithful to the Bayesian philosophy, we regard each input image $\mathcal{I}_i$ as a noisy measurement of an unknown image irradiance $\mathcal{I}_i^*$. Particularly, for the chosen reference view, this allows us to write:

$$\begin{aligned} \mathcal{I}_i(l_i(\mathbf{x}_1)) &= \mathcal{I}_1^*(\mathbf{x}_1) + \epsilon \\ \epsilon &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})\ , \end{aligned} \quad (2)$$

where ϵ is image noise, which is assumed to be normally distributed with zero mean and covariance $\mathbf{\Sigma}$. Both the irradiance or 'true' image $\mathcal{I}_1^*$ and $\mathbf{\Sigma}$ are unknown, and estimating them becomes part of the optimization procedure.

A major complication in a wide-baseline setting is the occlusion problem, which arises from the fact that not all parts of the scene, which are visible in a particular image, are also visible in the other images due to occlusion. When computing image correspondences, such occluded regions must be identified and excluded from the matching procedure. This will be modeled by introducing a set of visibility maps $\mathcal{V}_i(\mathbf{x}_1)$, which signal whether a scene point X that projects onto $\mathbf{x}_1$ in $\mathcal{I}_1$ is also visible in image $\mathcal{I}_i$ or not. Every element of $\mathcal{V}_i(\mathbf{x}_1)$ is a binary random variable which is either 1 or 0, corresponding to visibility or occlusion, respectively. The set $\mathcal{V}_i$ are hidden variables, and their values must be inferred from the input images. Note that, by choice of reference, $\mathcal{V}_1(\mathbf{x}_1) = 1$.

Estimating $\mathcal{D}_1(\mathbf{x}_1)$ can now be formally stated as finding those depth values which make the image correspondences $\mathcal{I}_1^*(\mathbf{x}_1) \Leftrightarrow \mathcal{I}_i(l_i(\mathbf{x}_1))$ restricted to $\mathcal{V}_i(\mathbf{x}_1) = 1$, most probable. By definition, $l_1(\mathbf{x}_1)$ is the identity transformation which maps $\mathbf{x}_1$ back onto itself.

2.1. MAP estimation

We are now facing the hard problem of estimating the unknown quantities $\boldsymbol{\theta} = \mathcal{D}_1, \mathcal{I}_1^*$ and $\mathbf{\Sigma}$ given a collection of observable data $\mathcal{I}_1, .., \mathcal{I}_N$. Furthermore, we have introduced the unobservable or hidden variables $\mathcal{V} = \{\mathcal{V}_{i\neq 1}\}$, which must also be inferred over the course of the optimization. In a Bayesian framework, the optimal value for $\boldsymbol{\theta}$ is the one that maximizes the posterior probability $p(\boldsymbol{\theta} \mid \mathcal{I}_1, .., \mathcal{I}_N)$. According to Bayes' rule, this posterior can be written as:

$$p(\boldsymbol{\theta}\,|\,\mathcal{I}_1, .., \mathcal{I}_N) = \frac{\int p(\mathcal{I}_1, .., \mathcal{I}_N\,|\,\boldsymbol{\theta}, \mathcal{V}) p(\boldsymbol{\theta}\,|\,\mathcal{V}) p(\mathcal{V}) d\mathcal{V}}{p(\mathcal{I}_1, .., \mathcal{I}_N)}\ , \quad (3)$$

where we have conditioned the data likelihood and the prior on the hidden variables $\mathcal{V}$. The denominator or 'evidence' is merely the integral of the numerator over all possible values of $\boldsymbol{\theta}$ and can be ignored in the maximization problem. Hence, we will try to optimize the numerator only. In order to find the most probable value for $\boldsymbol{\theta}$, we need to integrate over all possible values of $\mathcal{V}$ which is computationally intractable. Instead, we assume that the probability density function (PDF) of $\mathcal{V}$ is peaked about a single value, i.e. $p(\mathcal{V})$ is a Dirac-function centered at this value. This leads to an Estimation-Maximization (EM) based solution, which iterates between (*i*) estimating values for $\mathcal{V}$, given the current estimate of $\boldsymbol{\theta}$, and (*ii*) maximizing the posterior probability of $\boldsymbol{\theta}$, given the current estimate of $\mathcal{V}$. A more detailed description of this procedure will be given later. So, given a current estimate $\hat{\mathcal{V}}$ for the hidden variables, we want to optimize:

$$q(\boldsymbol{\theta}\,|\,\mathcal{I}_1, .., \mathcal{I}_N) = p(\mathcal{I}_1, .., \mathcal{I}_N\,|\,\boldsymbol{\theta}, \hat{\mathcal{V}}) p(\boldsymbol{\theta}\,|\,\hat{\mathcal{V}}) \quad (4)$$

The a-posteriori probability of $\boldsymbol{\theta}$ is proportional to the product of two terms: the data-likelihood $p(\mathcal{I}_1, .., \mathcal{I}_N\,|\,\boldsymbol{\theta}, \hat{\mathcal{V}})$ and a prior $p(\boldsymbol{\theta} \mid \hat{\mathcal{V}})$, which we will call L and P, respectively. We now discuss both terms in turn.

Under the assumption that the image noise is i.i.d. for all pixels in all views, the data likelihood L can be written as the product of all individual pixel probabilities:

$$L = \prod_{i=1}^{N} \prod_{\mathbf{x}_1} p\big(\mathcal{I}_i(l_i(\mathbf{x}_1))\,|\,\boldsymbol{\theta}\big)\ , \quad (5)$$

where the product is restricted to those terms for which $\mathcal{V}_i(\mathbf{x}_1) = 1$. Given the current estimate of the 'true' image $\mathcal{I}_1^*(\mathbf{x})$ and the noise distribution $\mathbf{\Sigma}$, we can further specify the likelihood to be:

$$L = \prod_{i=1}^{N} \prod_{\mathbf{x_1}} \mathcal{N}\big(\mathcal{I}_1^*(\mathbf{x}_1) - \mathcal{I}_i(l_i(\mathbf{x_1})); \mathbf{0}, \mathbf{\Sigma}\big) \,, \qquad (6)$$

where the normal distribution is defined by:

$$\mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\,\mathbf{\Sigma}\,|^{1/2}} \exp\big(-\frac{1}{2}\mathbf{x}^T \mathbf{\Sigma}^{-1} \mathbf{x}\big) \,. \quad (7)$$

Here, the variable d in the normalization constant denotes the dimensionality of $\mathbf{x}$, for color images this will be 3 while for intensity images d equals 1.

The formulation of an appropriate prior is slightly more complicated. We can factorize P as the product of a depth dependent and image dependent part as follows:

$$P = p\big(\mathcal{D}_1 \,|\, \mathcal{I}_1^*, \mathbf{\Sigma}\big) p\big(\mathcal{I}_1^*, \mathbf{\Sigma}\big) \,. \qquad (8)$$

Assuming we have no prior preference for the image related parameters, i.e. assuming a uniform prior over $\mathcal{I}_1^*$ and $\mathbf{\Sigma}$, this can be rewritten as:

$$P = p\big(\mathcal{D}_1 \,|\, \mathcal{I}_1^*, \mathbf{\Sigma}\big)\, c \,, \qquad (9)$$

where c is an appropriate constant. The depth prior $p(\mathcal{D}_1 \,|\, \mathcal{I}_1^*, \mathbf{\Sigma})$ will be modeled as an exponential density distribution of the form $exp\big(-R(\mathcal{I}_1^*, \mathcal{D}_1)/\lambda\big)$. Here, λ is a parameter which controls the width of the distribution, and $R(\mathcal{I}_1^*, \mathcal{D}_1)$ is a data-driven 'regularizer'. From such a regularizer we expect that it reflects our prior belief that the world is essentially simple, i.e. for a locally smooth solution $\mathcal{D}_1$ in the neighborhood of a particular point $\mathbf{x}_1$, its value should approach zero, making such a solution very likely. Vice-versa, large depth fluctuations should result in large values for the regularizer, making such solutions less likely. Furthermore, the regularizer should be data-driven: if the image $\mathcal{I}_1^*$ suggests a depth discontinuity, i.e. by the presence of a high image gradient at a particular point $\mathbf{x}_1$, a large depth discontinuity at $\mathbf{x}_1$ should not be made a-priori unlikely. Such regularizers are commonly used in the PDE-community, where they serve as *anisotropic diffusion operators* in optic flow or edge-preserving smoothing computations. In this work, we use the following regularizer [1]:

$$R(\mathcal{I}_1^*, \mathcal{D}_1) = \nabla \mathcal{D}_1^T \, T(\nabla \mathcal{I}_1^*) \, \nabla \mathcal{D}_1 \,. \qquad (10)$$

Here, $T(\nabla \mathcal{I}_1^*)$ is a diffusion tensor defined by:

$$T(\nabla \mathcal{I}_1^*) = \frac{1}{|\,\nabla \mathcal{I}_1^*\,|^2 + 2\nu^2} \Big(\nabla \mathcal{I}_1^{*\perp} \nabla \mathcal{I}_1^{*\perp T} + \nu^2 \mathbf{1} \Big) \,, \quad (11)$$

where $\mathbf{1}$ is the identity matrix, ν is a parameter controlling the degree of anisotropy and $\nabla \mathcal{I}_1^{*\perp}$ is the vector perpendicular to $\nabla \mathcal{I}_1^*$. For color images, the tensor is defined as the sum of the 3 individual color channel tensors. $R(\mathcal{I}_1^*, \mathcal{D}_1)$ is low when $\nabla \mathcal{D}_1$ is parallel to $\nabla \mathcal{I}_1^*$, which is exactly the desired behavior. Note that, by making the depth prior dependent on $\mathcal{I}_1^*$, it implicitly also makes it dependent on the original image data. While, strictly speaking, this violates the Bayesian principle that priors should not be estimated from the data, in practice it leads to more sensible solutions than setting them arbitrarily, or using so-called *conjugate* priors, whose main justification comes from computational simplicity [12].

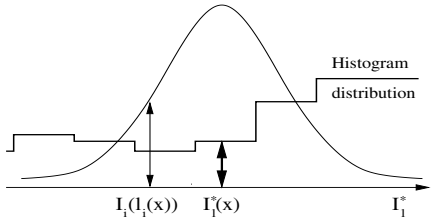

Figure 2: **Visibility estimation**: *For the estimation of visibility, we have to evaluate (i) the noise distribution on the value of the color-difference* $\mathbf{m}_i(\mathbf{x}_1) = \mathcal{I}_1^*(\mathbf{x}_1) - \mathcal{I}_i(l_i(\mathbf{x}_1))$ *and (ii) the histogram on the value of* $\mathcal{I}_1^*(\mathbf{x}_1)$.

We can now turn back to the optimization of $\boldsymbol{\theta}$. Instead of maximizing the posterior in (4), we minimize its negative logarithm. This leads (up to a constant) to the following energy formulation:

$$E(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^{N} \sum_{\mathbf{x_1}} \mathcal{V}_i \Big(\mathbf{m}_i^T \mathbf{\Sigma}^{-1} \mathbf{m}_i + \log\big((2\pi)^{\frac{d}{2}} |\mathbf{\Sigma}|\big) \Big) + \frac{1}{\lambda} R(\mathcal{I}_1^*, \mathcal{D}_1), \qquad (12)$$

where $\mathbf{m}_i = \mathcal{I}_1^*(\mathbf{x}_1) - \mathcal{I}_i(l_i(\mathbf{x}_1))$. Interestingly, by incorporating a noise-model in the image formation, there is an automatic balancing between the matching term and regularization term. When the norm of $\mathbf{\Sigma}$ is high, the influence of the matching term decreases, which leads to a higher degree of regularization, i.e. smoother depth solutions. Vice-versa, when optimizing (12) converges to a stable point, the norm of $\mathbf{\Sigma}$ decreases, which in turn puts more emphasis on the matching term.

2.2. An EM solution

In the previous paragraph, an energy equation w.r.t. the unknown quantities $\boldsymbol{\theta}$ was derived. This energy corresponds to the negative logarithm of the posterior distribution of $\boldsymbol{\theta}$, given the current estimate of the hidden variables $\mathcal{V}$. Now we will derive the EM-equations, which iterate between the estimation of $\mathcal{V}$ and the minimization of $E(\boldsymbol{\theta})$.

Figure 1: **Bookshelf scene**: *The top row shows the 4 original input images. In this experiment, we zoomed in a virtual camera on the first image ($\mathcal{I}_1$), to visualize the integration effects of $\mathcal{I}^*$. The second row displays the output of the algorithm. Here, the 1^{th} image is the depth map $\mathcal{D}$ computed from the new camera position. The last 2 images show a detail of $\mathcal{I}_1$ and $\mathcal{I}^*$, respectively. By integrating the information of all 4 views, the discretization errors in $\mathcal{I}_1$ have largely disappeared.*

E-step On the $(k+1)^{th}$ iteration, the hidden variables $\mathcal{V}_i(\mathbf{x}_1)$, are replaced by their conditional expectation given the data, where we use the current estimates $\boldsymbol{\theta}^{(k)}$ for $\boldsymbol{\theta}$. The expected value for the visibility is given by $E[\mathcal{V}_i|\mathcal{I}_1^*, \boldsymbol{\Sigma}, \mathcal{D}_1] \equiv \Pr(\mathcal{V}_i=1|\mathcal{I}_1^*, \boldsymbol{\Sigma}, \mathcal{D}_1)$. According to Bayes' rule, the latter probability can be expressed as:

$$\Pr(\mathcal{V}_i=1|\mathcal{I}_1^*, \boldsymbol{\Sigma}, \mathcal{D}_1) = \frac{p(\mathcal{I}_1^*|\mathcal{V}_i=1, \mathcal{D}_1, \boldsymbol{\Sigma})}{p(\mathcal{I}_1^*|\mathcal{V}_i=1, \mathcal{D}_1, \boldsymbol{\Sigma}) + p(\mathcal{I}_1^*|\mathcal{V}_i=0, \mathcal{D}_1, \boldsymbol{\Sigma})}\,, \quad (13)$$

where we have assumed equal priors on the probability of a pixel being visible or not. Given the current estimate of $\boldsymbol{\theta}$, the PDF $p(\mathcal{I}_1^*|\mathcal{V}_i = 1, \mathcal{D}_1, \boldsymbol{\Sigma})$ is given by the value of the noise distribution evaluated on the color-difference $\mathbf{m}_i$ between $\mathcal{I}_1^*(\mathbf{x}_1)$ and $\mathcal{I}_i(l_i(\mathbf{x}_1))$:

$$p(\mathcal{I}_1^*|\mathcal{V}_i=1, \mathcal{D}_1, \boldsymbol{\Sigma}) = \mathcal{N}\big(\mathbf{m}_i; \mathbf{0}, \boldsymbol{\Sigma}\big)\,. \quad (14)$$

We provide a *global* estimate for the second PDF $p(\mathcal{I}_1^*|\mathcal{V}_i = 0, \mathcal{D}_1, \boldsymbol{\Sigma})$ by building a histogram of the color-values in $\mathcal{I}_1^*$ which are currently invisible. This is merely the histogram of $\mathcal{I}_1^*$ where the contribution of each pixel is weighted by $(1-\mathcal{V}_i(\mathbf{x}_1))$. Note that, if a particular pixel in $\mathcal{I}_1^*$ is marked as not-visible, in the next iterations this will automatically decrease the visibility estimates of all similarly colored pixels. This makes sense from a perceptual point of view, and has a regularizing effect on the visibility maps. Both PDFs are shown in fig.2. The update equations for $\mathcal{V}_i(\mathbf{x}_1)$ are:

$$\mathcal{V}_{i\neq 1} \leftarrow \frac{\mathcal{N}\big(\mathbf{m}_i; \mathbf{0}, \boldsymbol{\Sigma}\big)}{\mathcal{N}\big(\mathbf{m}_i; \mathbf{0}, \boldsymbol{\Sigma}\big) \;+\; \mathrm{HIST}_{\mathcal{I}_1^*, (1-\mathcal{V}_i)}(\mathcal{I}_1^*)}\,. \quad (15)$$

Note that visibility is computed purely photometrically, and is in essence a measure for how well the pixel colors are explained by the model in eq.(2), given the current estimate of the parameters $\boldsymbol{\theta}$. Obviously, geometrical information, derived from the evolving depth-estimates, could be included in the procedure. Given the available camera calibrations and the current estimate of $\mathcal{D}_1$, it is possible to compute hidden surfaces w.r.t. all cameras and set the visibilities of the occluded parts to zero. However, for several reasons we chose not to do so. First of all, this would significantly increase the computational cost of the algorithm. Secondly, the proposed approach generates, as will be shown in the next section, accurate visibility maps. Finally, because 'visibility' measures deviation from the image model, it does not only signal geometric occlusion, but also detects outlier pixels. These outliers typically occur at positions with strong specular reflections, and cause, when not excluded from the computations, spurious 3D-effects in the final reconstruction.

M-step At the M-step, the intent is to compute values for $\boldsymbol{\theta}$ that minimize (12), given the current estimates of $\mathcal{V}_i$. This is achieved by setting the parameters $\boldsymbol{\theta}$ to the appropriate root of the derivative equation, $\partial E(\boldsymbol{\theta})/\partial\theta = \mathbf{0}$. For the image related parameters $\mathcal{I}_1^*$ and $\boldsymbol{\Sigma}$, a closed form expressions for the roots can be derived and the update equations are:

$$\mathcal{I}_1^* \leftarrow \frac{\sum_i \mathcal{V}_i\, \mathcal{I}_i(l_i)}{\sum_i \mathcal{V}_i}\,,$$

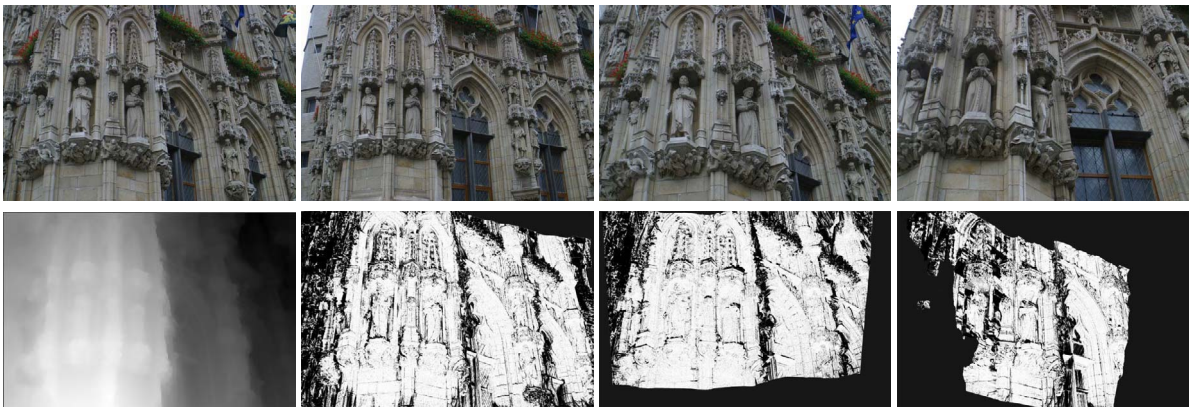

Figure 3: **Cityhall scene**: *The top row shows the 4 original input images. The bottom displays some of the output of the algorithm, where the 1^{st} image of the sequence ($\mathcal{I}_1$) was taken as the reference view. The left-most image is the depth map $\mathcal{D}_1$ of $\mathcal{I}_1$. The last 3 images are the visibility maps $\mathcal{V}_2$, $\mathcal{V}_3$ and $\mathcal{V}_4$, which signal which pixels in $\mathcal{I}_1$ are visible in the other images.*

$$\mathbf{\Sigma} \leftarrow \frac{\sum_i \sum_x \mathcal{V}_i \left(\mathcal{I}_1^* - \mathcal{I}_i(l_i)\right)\left(\mathcal{I}_1^* - \mathcal{I}_i(l_i)\right)^T}{\sum_i \sum_x \mathcal{V}_i} . \quad (16)$$

In order to arrive at these closed-form expressions, we ignored the effects of these variables on the regularization term. This is admissible because their influence on the depth regularizer R is small compared to their influence on the matching term. $\mathbf{\Sigma}$ is only indirectly related to R by way of computation of the visibility maps, which have an effect on R via the computation of $\mathcal{I}_1^*$. The image $\mathcal{I}_1^*$ has an effect on R via its gradient, which is used to define a quadratic norm on the depth gradient (10). Changes of $\mathcal{I}_1^*$ will therefore only exert a minor influence on R. However, for the update of the depth map $\mathcal{D}_1$ we are not so lucky, because $\mathcal{D}_1$ strongly influences both the matching and the regularization term. To minimize E w.r.t. $\mathcal{D}_1$, we therefore follow a gradient descent approach. By applying the Euler-Lagrange formalism, we get:

$$\begin{aligned} \frac{\partial E}{\partial \mathcal{D}_1} &= \sum_{i=2}^{N} -2\mathcal{V}_i(\mathcal{I}_1^* - \mathcal{I}_i(l_i))^T \mathbf{\Sigma}^{-1} \nabla \mathcal{I}_i(l_i)\, \partial l_i \\ &\quad + \frac{1}{\lambda} \mathrm{div}\big(T(\nabla \mathcal{I}_1^*)\, \nabla \mathcal{D}_1\big) . \end{aligned} \quad (17)$$

Here, we ignored the $\mathcal{D}_1$-dependencies of $\mathbf{\Sigma}$ and $\mathcal{I}_1^*$, which are small compared to the $\mathcal{D}_1$-dependencies of $\mathcal{I}_i(l_i)$ and the regularizer R. Image $\mathcal{I}_1$ is excluded from the sum, because $l_1(\mathbf{x}_1)$ is the identity transformation, i.e. changing $\mathcal{D}_1$ will not change the influence of $\mathcal{I}_1$ on the matching term. The derivative ∂l_i is a 2-vector, whose expression is easily derived from (1).

2.3. Novel view synthesis

In the previous discussion, the first image of the sequence, $\mathcal{I}_1$, was taken as a reference view, and we described how to compute $\mathcal{D}_1$ in the reference frame attached to this camera. Obviously, any image from the input set could have been chosen as a reference. Now, we will leave the set of input camera's and describe how to compute a synthetic view from a virtual camera.

Central to the ongoing discussion is the image $\mathcal{I}^*$, which is a model for the unknown image irradiance. When we choose one of the input images as a reference view, $\mathcal{I}^*$ will evolve towards what could have been observed from this point of view, if image formation were perfect and all objects in the scene were well-behaved (i.e. perfect diffuse reflectors). There is nothing that prevents us from applying the procedure to a virtual camera position, where the hope is that $\mathcal{I}^*$ will evolve towards what would have been observed, were a real camera put at this location.

The challenge of such *novel view synthesis* is that, initially, we have no certain visual information to hold on to. Therefore, we will rely on the image correspondences which remain from the calibration procedure. They constitute a small set of initial 3D-points which pull the depth solution in the right direction. To this end, we implemented an inhomogeneous time diffusion as described in [10], which slows down the depth evolution of the calibration points.

2.4. Algorithmic issues

On the first iteration of the EM-algorithm, we have to provide an initial estimate for the hidden variables $\mathcal{V}$. When we use the 1^{th} image as a reference, $\mathcal{V}_1$ is fixed to 1.0, and all other visibility maps are initialized with 0.5, to reflect

Figure 4: **Cityhall scene**: *Evolution of a detail of $\mathcal{I}_1^*$ during subsequent EM iterations. Note how $\mathcal{I}_1^*$ gradually sharpens, due to the ongoing convergence of the depth estimation.*

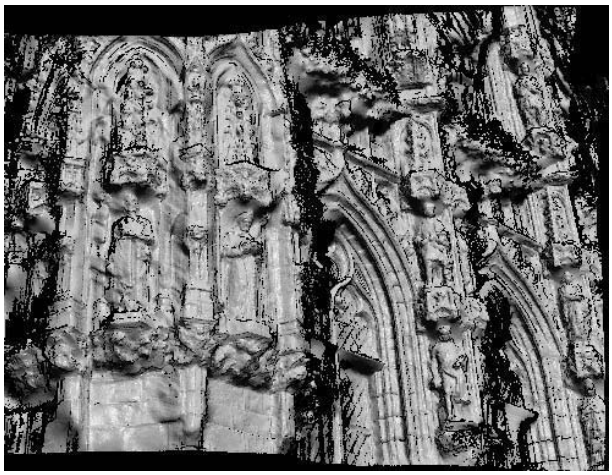

Figure 5: **Cityhall scene**: *A rendering of the (untextured) 3D-model of the first image from the sequence.*

our initial uncertainty about visibility or occlusion. When we compute synthetic views, on the other hand, all maps are initialized with 0.5. From the update equations (16), it can be seen that $\mathcal{I}_1^*$ is an average of all images, where the contribution of each pixel is weighted by its estimated visibility. Because, by choice of reference, $\mathcal{V}_1(\mathbf{x}_1) = 1$, $\mathcal{I}_1$ keeps on pulling $\mathcal{I}_1^*$ towards itself. In the first iterations of the algorithm when the depth estimates have not yet converged, $\mathcal{I}_1^*$ will be a smooth image, because pixel values which do not correspond to the same point in the scene are averaged. This leads to large values for $\mathbf{\Sigma}$, which in turn puts more emphasis on the regularizer. This type of *soft decision making* is typical for EM algorithms.

The presented algorithm has only 2 free parameters. They are ν, which controls the degree of anisotropy in (11), and λ, which controls the width (hence the importance) of regularization prior. Not surprisingly, both parameters originate from our prior beliefs which we incorporated into the equations, as such they can be considered unavoidable.

In the M-step of the algorithm, the current depth estimate is refined through a gradient descent procedure. This is implemented as a diffusion equation on $\mathcal{D}_1$. The search for correct depth estimates is guided by a sparse set of initial point correspondences which originate from the camera calibration procedure. Implicit discretization [13] was used for the sake of computational speed. During optimization, integral positions $\mathbf{x}_1$ in $\mathcal{I}_1^*$ will, by applying $l_i()$, in general not map onto integral positions in $\mathcal{I}_i$, and we use bilinear interpolation to sample pixel and gradient values from $\mathcal{I}_i$. To cope with large baselines between the images, a pyramidal coarse-to-fine strategy is followed. We will end with a final note on the convergence properties of the algorithm. Dempster *et al.* [2] have shown that, for Maximum-Likelihood (ML) estimation, each iteration of EM is guaranteed to increase the data-likelihood, which drives the parameters $\boldsymbol{\theta}$ to a local optimum of L. In this work, we have included a prior on the unknown variables, so for the moment we can not make such strong claims. However, various trials on different data sets have confirmed the robust behavior of the proposed algorithm.

3. Experiments

We tested the proposed algorithm on 3 different data sets. The first one is the so-called *bookshelf scene* [7], which presents a case with strong wide-baseline and relatively large image discretization effects. The second one is the *cityhall scene* [10], which is a case with large depth discontinuities. Finally, the third data set is the so-called *monkey scene* [3], where we have to deal with mixed pixels and non-Lambertian lighting effects.

Bookshelf scene In this experiment, we use 4 input images of size 640x480 pixels. The images have relatively strong discretization effects, which are particularly obvious in the neighborhood of the printings on the books. To test the integration effects of $\mathcal{I}_1^*$, we chose a virtual camera position which has the same direction as the camera of the first image from the input sequence, but which is moved closer to the scene. This allows us to visually compare the final image model $\mathcal{I}^*$ with the original image $\mathcal{I}_1$ at a higher pixel rate. The results are shown in figure 1. They show that, by integrating the information of all 4 views, the discretization effects of the original images have largely disappeared.

Cityhall scene In this experiment, we use 4 input images of size 3072x2048 pixels. This scene presents a case

Figure 6: **Monkey scene**: *The left image of the top row is the novel view which was synthesized from a virtual camera position. The right image represents the groundtruth, i.e. the image from the input sequence which was taken from the chosen camera position. The bottom row shows* 2 *corresponding image regions from the novel view (left) and the groundtruth (right). The correspondence is almost perfect and high-frequency detail has been largely preserved. However, where foreground and background meet there are tiny defects, due to single depth assignment.*

with a complicated 3D-structure and large depth discontinuities. We choose the 1^{th} image of the sequence as the reference image. It takes 5 minutes for the algorithm to converge. Figure 3 shows the input images and the final depth map $\mathcal{D}_1$. It also shows the visibility maps $\mathcal{V}_2$, $\mathcal{V}_3$ and $\mathcal{V}_4$, which signal for every pixel of $\mathcal{I}_1$ whether or not the corresponding scene point is visible in $\mathcal{I}_2$, $\mathcal{I}_3$ and $\mathcal{I}_4$, respectively. Because visibility is computed photometrically, all pixels which are not well explained by the image formation model are excluded from the computations. These are not only geometrically occluded pixels, but also pixels with a high degree of specular reflections (e.g. reflections in the window in $\mathcal{V}_4$), or mixed pixels in high frequency regions (e.g. flower beds in $\mathcal{V}_2$ and $\mathcal{V}_3$). Figure 4 displays the evolution of a detail of $\mathcal{I}_1^*$ during subsequent EM-iterations. Note that, while $\mathcal{D}_1$ converges to the final solution, $\mathcal{I}_1^*$ gradually sharpens because corresponding pixels are brought into alignment. Finally, textured and untextured renderings of the 3D-model are shown in figures 5 and 7.

Monkey scene This sequence consists of 30 consecutive images of size 640x480 pixels, which are taken along a camera path around the object. We use these images to test novel-view synthesis. The experiment goes as follows: from the overall set, we take the camera position of one of the images as our virtual camera. Next, 6 images, 3 from the left ($\mathcal{I}_1$, $\mathcal{I}_2$, $\mathcal{I}_3$) and 3 from the right ($\mathcal{I}_5$, $\mathcal{I}_6$, $\mathcal{I}_7$) are chosen as the input for our algorithm. Note that the image of the central camera, $\mathcal{I}_4$, is not part of the input set. This image will serve as groundtruth with which we compare our synthesis. The result is shown in figure 6. As can be observed from this figure, the results are very good. Noticeably, the high-frequency details have been largely preserved, which is not obvious given the irregular 3D-details on the surface of the animal. The reason for this is that, when computing image correspondences, the algorithm attaches to strong features (e.g. strong specularities on the hairs) which are local to particular points of the scene and well preserved over the sequence. As a result, at these positions depth is computed well, and the features are not smoothed out in I^*. On the other hand, we have difficulties in handling mixed pixels which combine visual information of scene points with very different depths. The visually observable errors are therefore mainly situated at the hull of the object. In [3] this is solved by introducing image-based priors. This is sensible if one is interested in faithful renderings, however, such an approach requires more input data and takes much longer to compute.

4. Conclusion

In this paper, we presented a method for dense depth reconstruction and view interpolation from a small set of wide-baseline images, where the problem is addressed from a probabilistic point of view. One of the advantages of such an analysis is that it makes the tactile or implicit assumptions underlying a particular algorithm explicit. In our approach, the main assumptions are dominant diffuse reflection and i.i.d. pixel-color distributions. A smoothness regularizer was introduced to give shape to our prior beliefs about the world. The key result of this paper is that energy minimization, which is the cornerstone of PDE-based methods, is strongly related to MAP-estimation. More specifically, in terms of our notation, the typical PDE-functional is a special case of eq.(12), in which $\mathcal{I}^*$ is defined to be the reference image and noise is supposed to have unit strength.

In the presented framework, images are modeled as noisy measurements of an unknown irradiance or 'true' image-function. This has three principal advantages. First of all, it brings about an automatic balancing between matching and smoothness. In early stages of the optimization, more emphasis is put on regularization, whereas in the convergence stage, the matching term will gain importance. Secondly, because the true image is a learned model of image irradiance, we are able to leave the input camera positions, which in turn allows to compute view interpolations. Finally, the resulting model integrates all available image information, and can as such be used as a texture map for the final 3D-reconstruction.

Figure 7: **Cityhall scene**: *Three detail views of the textured 3D-model of the first image from the sequence. The black areas (e.g. behind the statue in the* 1^{st} *model) correspond to pixels which are not visible in the other images.*

In the experiments on the *monkey scene*, we showed that our algorithm can generate realistic view interpolations from a small set of images, without loss of high frequency information. On the other hand, we have difficulties in handling mixed pixels, which combine visual information of scene points with very different depths. To deal with this problem, we would have to assign multiple depths to a particular pixel, and at this stage it is not obvious how that could be done.

A strong emphasis was put on the computation of visibility. The visibility or occlusion of a particular pixel is modeled as a mixture problem, and we introduced a set of hidden variables, the so-called visibility maps, which are sequentially updated in the EM algorithm. These estimates are a measure for how well the images are explained by the model, given the current estimates of depth, noise-level and true image. We can therefore frame our algorithm as a (regularized) iteratively reweighted least squares algorithm. We only used color information to estimate visibility, but other information, e.g. derived from the evolving depth estimates, can be easily included.

In the current version of algorithm, small deviations from the Lambertian assumptions are caught by the noise term. Obviously, if the objects in the scene display a high degree of specularity, such an approach is not sufficient, and light sources and surface properties should be explicitly modeled. This fits well in the EM-scheme, and will be the object of our future research.

References

[1] L. Alvarez, R. Deriche, J. Sánchez, J. Weickert, "Dense disparity map estimation respecting image derivatives: a PDE and scale-space based approach", *JVCIR*, Vol. 13, pp. 3-21, 2002.

[2] A. P. Dempster, N. M. Laird, D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *J. R. Statist. Soc. B*, Vol. 39, pp. 1-38, 1977.

[3] A. Fitzgibbon, Y. Wexler, A. Zisserman, "Image-based rendering using image-based priors", *ICCV*, pp. 1176-1183, 2003.

[4] O. Faugeras, R. Keriven, "Complete dense stereovision using level set methods", *Proc. ECCV*, pp. 379-393, 1998.

[5] R. I. Hartley, A. Zisserman, *Multiple View Geometry in Computer Vision,* Cambridge University Press, 1998.

[6] K. N. Kutulakos, S. M. Seitz, "A theory of shape by space carving", *IJCV*, Vol. 38(3), pp. 197-216, 2000.

[7] J. Matas, O. Chum, M. Urban, T. Pajdla, "Robust Wide Baseline Stereo for Maximally Stable External Regions", *Proc. BMVC*, pp. 414-431, 2002.

[8] D. Scharstein, R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms", *IJCV*, Vol. 47, pp. 7-42, 2002.

[9] G. Slabaugh, R. Schafer, M. Hans, "Image-Based Photo Hulls", *1st Int. Symp. of 3D Data Processing Visualization and Transmission*, pp. 704-707, 2002.

[10] C. Strecha, T. Tuytelaars, L. Van Gool, "Dense Matching of Multiple Wide-baseline Views", *ICCV*, pp. 1194-1201, 2003.

[11] T. Tuytelaars, L. Van Gool, "Wide baseline stereo matching based on local, affinely invariant regions", *Proc. BMVC*, pp. 412-422, 2000.

[12] N. Vasconceles, A. Lippman, "Empirical Bayesian EM-based Motion Segmentation", *CVPR*, pp. 527-532, 1997.

[13] J. Weickert, B.M. ter Haar Romeny, M.A. Viergever, "Efficient and Reliable Schemes for Nonlinear Diffusion Filtering," *IEEE Transactions on Image Processing*, Vol. 7(3), pp 398-410, 1998.