Fast Joint Estimation of Silhouettes and Dense 3D Geometry from Multiple Images

Kalin Kolev, Thomas Brox, and Daniel Cremers

Abstract—We propose a probabilistic formulation of joint silhouette extraction and 3D reconstruction given a series of calibrated 2D images. Instead of segmenting each image separately in order to construct a 3D surface consistent with the estimated silhouettes, we compute the most probable 3D shape that gives rise to the observed color information. The probabilistic framework, based on Bayesian inference, enables robust 3D reconstruction by optimally taking into account the contribution of all views. We solve the arising maximum a posteriori shape inference in a globally optimal manner by convex relaxation techniques in a spatially continuous representation. For an interactively provided user input in the form of scribbles specifying foreground and background regions, we build corresponding color distributions as multivariate Gaussians and find a volume occupancy that best fits to this data in a variational sense. Compared to classical methods for silhouette-based multiview reconstruction, the proposed approach does not depend on initialization and enjoys significant resilience to violations of the model assumptions due to background clutter, specular reflections, and camera sensor perturbations. In experiments on several real-world data sets, we show that exploiting a silhouette coherency criterion in a multiview setting allows for dramatic improvements of silhouette quality over independent 2D segmentations without any significant increase of computational efforts. This results in more accurate visual hull estimation, needed by a multitude of image-based modeling approaches. We made use of recent advances in parallel computing with a GPU implementation of the proposed method generating reconstructions on volume grids of more than 20 million voxels in up to 4.41 seconds.

Index Terms—Shape from silhouettes, interactive segmentation, convex optimization.

1 Introduction

The problem of modeling 3D objects from multiple views has seen some groundbreaking advances in recent years [14], [10], [18]. Nevertheless, most methods require special settings, which allow for reliable silhouette extraction, and are computationally quite demanding due to the estimation of robust photoconsistency measures associated with each point in space. The goal of this paper is to cast multiview reconstruction as a problem of interactive joint segmentation of all images which does not require photoconsistency or silhouette information and which provides 3D models of acceptable quality in the order of 2-5 seconds. See Fig. 1 for an example.

1.1 Shape from Silhouettes

The earliest approaches for multiview 3D reconstruction, dating back to the 1970s [2], use outlines to infer geometrical structure. While silhouette-based methods are not capable of retrieving surface concavities, since these do not affect the image projections, they come along with some important advantages and are often preferred in applications like robot navigation and tracking. First, they enjoy significant stability and efficiency, which allows them to operate in challenging imaging conditions. Second, they seem to be the only reasonable alternative for recovering textureless or homogeneous objects. Third, they usually do not require exact visibility estimation. This is a great advantage over multiview/photometric stereo and shading techniques, where visibility reasoning leads to a chicken-and-egg problem. Silhouette-based methods can provide useful initial solutions that can be refined by other techniques.

Usually, silhouettes are used to infer surfaces in a two step process: An individual decision about pixel occupancy is made on a per-view basis, then geometrical structure is inferred from all estimated segmentations [2], [20], [29]. Unfortunately, the automatic segmentation of individual images is in many cases not feasible, especially in the presence of noise, illumination variations, and background clutter. To this end, researchers developed interactive approaches, where the user is required to guide the process by manually labeling image regions [5], [27], [33]. While two scribbles marking foreground and background are usually sufficient for simple images, the extent of required user interaction increases significantly in case of cluttered or camouflaged environments. This problem becomes more relevant if we consider a collection of input images where even a modest amount of user interaction on an individual image basis entails significant efforts. Applying the above interactive segmentation methods leads to a two-step silhouette fusion procedure, where binary image labelings are first computed separately and combined subsequently to build a unified 3D model. Yet, this simple scheme is suboptimal in the sense that the segmentation of each individual image does not take into account information from the remaining imagery. It is beneficial to exploit the

- *K. Kolev and D. Cremers are with the Department of Computer Science, Technical University of München, Boltzmanstrasse 3, Garching bei München 85748, Munich, Germany. E-mail: {kalin.kolev, daniel.cremers}@in.tum.de.*
- *T. Brox is with the Department of Computer Science, Albert-Ludwigs-University Freiburg, room 01-29/30, Building 052, Georges-Köhler-Allee, Freiburg 79110, Germany. E-mail: brox@informatik.uni-freiburg.de.*

Manuscript received 8 July 2010; revised 29 Dec. 2010; accepted 25 June 2011; published online 28 July 2011.
Recommended for acceptance by A. Criminisi.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-07-0517.
Digital Object Identifier no. 10.1109/TPAMI.2011.150.

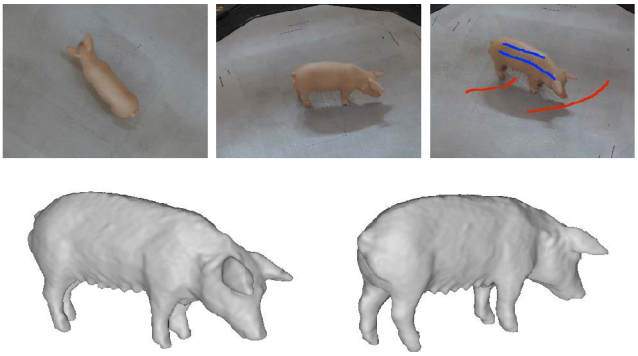

Fig. 1. The figure illustrates an example where the proposed interactive approach delivers an accurate reconstruction in only 2.08 seconds. *First row:* Three out of 27 input images and utilized user interaction (blue scribbles mark foreground and red ones background). *Second row:* Two views of the reconstructed surface. Note that the image sequence exhibits a major challenge for classical multiview stereo methods due to the homogeneous appearance of the imaged object. The figure is best viewed in color.

fact that all input views capture the same scene in order to counteract possible inaccuracies in single observations. Furthermore, for realistic inhomogeneous backgrounds, a robust segmentation will require imposing spatial regularity of the boundary. It is unclear how this additional smoothness assumption will affect the final 3D model.

1.2 Previous Work

Since the current paper touches the topic of shape from silhouettes, it is related to a vast body of prior work in multiview 3D reconstruction. Historically, the main strategy for computing a silhouette-consistent shape has been to directly implement the intersection of visual cones corresponding to different silhouettes [2]. Such techniques aim at estimating the object's *visual hull*, i.e., the largest shape that yields the same silhouettes as the observed ones [20]. An important class within this domain exhibits volumetric approaches. The key idea is to discretize the space by a fixed voxel grid and label each voxel as opaque or transparent according to its projections onto the images. An early paper reporting a volumetric approximation of the visual hull is due to Martin and Aggarwal [23]. Subsequently, octree-based representations have been employed by Potmesil [26] and Szeliski [31] in order to increase the efficiency. The viability of volumetric techniques is due to some important properties. They are not susceptible to numerical difficulties that can arise in analytic computations and enable objects of arbitrary topology and complexity to be reconstructed. Moreover, their computational time can be drastically reduced by using parallel implementations over the voxel grid. While most of the prior research has been focused on increasing the accuracy or efficiency of the process of silhouette fusion, little attention has been paid to the problem of improving the robustness and minimizing the efforts needed to achieve this goal.

Probabilistic methods for multiview silhouette fusion have previously been proposed in the context of model-free tracking [9], [13]. However, since they are based on background subtraction, they require special environmental conditions and are not directly applicable to the problem of joint color-based segmentation and reconstruction from real-world image sequences.

Along with shape from silhouette techniques, researchers have advocated the use of theoretically more transparent energy minimization methods which directly compute a 3D shape, consistent with all images [36], [29], [6]. In [36], the 3D surface sought is modeled in a variational sense by minimizing the reprojection error between estimated surface intensities and observed ones. Rather than regularizing the boundary of individual segmentations, the variational formulation allows to directly impose regularity of the estimated 3D model. One of the difficulties with such energy minimization methods is that respective functionals are not convex. Therefore, the proposed gradient descent optimization is likely to get stuck in a local minimum, especially in case of a complex object topology. The aspect that typically prevents global optimizability of respective functionals is the fact that the observed projections cannot be inverted and therefore do not allow a direct inference of voxel occupancy. They merely allow statements about the collection of voxels along respective lines of sight. This difficulty can be circumvented by measuring costs in 3D space rather than on the image plane [29], [6]. Yet, the approach in [29] requires binary view segmentations, which makes it susceptible to noise in individual observations. A more general algorithm that operates directly on image color information has independently been proposed in [6]. Although the authors report significant improvements over classical silhouette fusion techniques, the applicability of their formulation is limited. First, the method does not allow for efficient user interaction on a single image. It either runs without any user intervention, which is unreliable in many cases, or it requires interaction in all views due to the utilization of separate background models. Note that the proposed fixation arguments may fail for certain objects and camera settings. Second, the method is quite slow (computational times up to a couple of hours) and sequential in nature, which entails the lack of parallelization potential. In this respect, additional difficulties are caused by the employment of graph cut optimization, the parallelization of which is also not straightforward.

User interaction has become an established tool for segmenting real-world images. The pioneering work [5] addresses the foreground/background interactive segmentation in still images via max-flow/min-cut energy minimization. The energy balances between likelihood of pixels belonging to the foreground and the edge contrast imposing regularization. The user-provided scribbles collect statistical information on pixels and serve additionally as hard constraints. The GrabCut [3], [27] framework further simplifies the user interaction required. It allows for interactively adding scribbles to improve the initial segmentation. Full color statistics are used, modeled as mixtures of Gaussians, and these are updated as the segmentation progresses. Further developments have led to the utilization of weighted geodesic distances to the information supplied by the user [8], [1]. Recent advances in convex optimization [7] initiated the appearance of [33], where total variation minimization has been adopted to interactive image segmentation. Yet, all of these methods are restricted to individual image segmentation. They are unable to adequately process collections of images capturing the same scene, where the interdependence between different observations is crucial. This drawback motivated the development of image cosegmentation [28]—a framework exploiting the

overlapping in content of two or more images with the goal of improving the segmentation results. While the current work is inspired by a similar incentive, there is one important difference—in our case, the images are calibrated, which is an additional source of information and allows for a more accurate modeling of the connection between different observations.

Furthermore, researchers have adapted interactivity to video segmentation [35], [21]. In this context, additional improvements are obtained by imposing time coherency, under the assumption that the changes between successive frames are minor. However, the generalization of such approaches to make them applicable to the problem of segmenting collections of still photographs imaging the same scene remains an open challenge.

1.3 Contribution

In this paper, we propose a probabilistic treatment of the multiview reconstruction problem. Instead of processing the input images independently and subsequently fusing the resulting information, we compute the most probable surface that gives rise to the given observations. To this end, we adopt a volumetric approach, where we assign to each voxel probability costs for being inside or outside the imaged shape. Color distributions for foreground and background are estimated from user interactions in the form of a few scribbles in only *one* of the input images. We avoid explicit visibility reasoning by initially neglecting the interdependence of voxels and reintroducing it in a probabilistic manner at a later stage of the modeling. The consequence of this approximation is that the resulting Bayesian inference problem can be optimized *globally*. In particular, we employ established convex relaxation techniques to find the exact solution. In numerous experiments, we demonstrate that the proposed probabilistic formulation provides far more robust reconstructions than the classical silhouette fusion method [2], [20]. Furthermore, we show dramatic improvements of individual image segmentations by exploring multiview coherency criteria. Based on a volumetric parallelization, we are able to obtain reconstructions of surprising accuracy in only a few seconds by means of a GPU implementation. It is important to note that the proposed framework is general and can be used in combination with any probabilistic model for image inference.

A preliminary version of this framework was presented at a conference [17], where the probabilistic formulation was introduced for grayscale images. In the following, we summarize the novelty of the current paper over [17]:

- The iterative update of the parameters of the foreground/background color distributions is replaced by an interactive determination through user input.
- We show that the resulting functional can be minimized by means of convex relaxation. This allows us to compute globally optimal shapes independent from initialization.
- Making use of recent developments in parallel computing, we present a GPU implementation of the proposed approach which leads to a formidable reduction in runtime from above an hour to a few seconds.

The paper is laid out as follows: In the next section, we present and discuss the underlying probabilistic framework. An energy minimization formulation and a respective numerical optimization scheme are derived in Section 3. In Section 4, we show experimental results demonstrating in particular superior performance over classical silhouette fusion techniques. Finally, we conclude in Section 5.

2 Probabilistic Volume Intersection

2.1 3D Shape Modeling via Bayesian Inference

We consider the problem of probabilistic voxel labeling from a series of calibrated images of a scene. The relationship between image observations and surface estimation is established in terms of Bayesian inference, which allows to derive a MAP estimate for the sought 3D surface by modeling the process of image formation. The probabilistic framework covers a wide range of noise sources like camera sensor perturbance, surface reflections, erroneous camera calibration, etc. All these effects have as a result that observed colors deviate from the expected ones. In the following, the proposed probabilistic formulation is explained in more detail.

Let

$$V := [v_{11}, v_{12}] \times [v_{21}, v_{22}] \times [v_{31}, v_{32}] \subset \mathbb{R}^3$$

be a volume enclosing the object of interest with boundary values $v_{lm} \in \mathbb{R}$, and

$$\tilde{V} := \left\{ \begin{pmatrix} v_{11} + i \cdot \dfrac{v_{12} - v_{11}}{N_1} \\ v_{21} + j \cdot \dfrac{v_{22} - v_{21}}{N_2} \\ v_{31} + k \cdot \dfrac{v_{32} - v_{31}}{N_3} \end{pmatrix} \middle| \begin{array}{l} i = 0, \ldots, N_1 - 1 \\ j = 0, \ldots, N_2 - 1 \\ k = 0, \ldots, N_3 - 1 \end{array} \right\}$$

a discretized version of resolution $N_1 \times N_2 \times N_3$. Obviously, we have the relation $\tilde{V} \subset V$. Further, let $I_1, \ldots, I_n : \Omega \mapsto \mathbb{R}^3$ be a collection of calibrated color images with perspective projections $\pi_1, \ldots, \pi_n : V \to \Omega$, where $\Omega \subset \mathbb{R}^2$ denotes a common image domain. Given the set of views, we are looking for the most probable surface $\hat{S}$ that gives rise to them, that is,

$$\hat{S} = \arg\max_{S \in \Lambda} P(S \mid \{I_1, \ldots, I_n\}), \tag{1}$$

where $\Lambda := \{S \mid S : D \subset \mathbb{R}^2 \to V\}$ is the set of all closed surfaces lying inside the volume V. By means of the Bayes formula, we obtain

$$P(S \mid \{I_1, \ldots, I_n\}) \propto P(\{I_1, \ldots, I_n\} \mid S) \cdot P(S), \tag{2}$$

where the a priori probability $P(S)$ allows us to introduce preference to a certain class of surfaces possessing desired properties like smoothness, simple topology, etc. It should be noted that the constant term

$$\frac{1}{P(\{I_1, \ldots, I_n\})}$$

has been omitted in the above expression since it does not influence the shape retrieval process.

A crucial issue in this formulation is the modeling of the likelihood $P(\{I_1, \ldots, I_n\} \mid S)$. It reflects the image formation process in terms of the probability for observing images

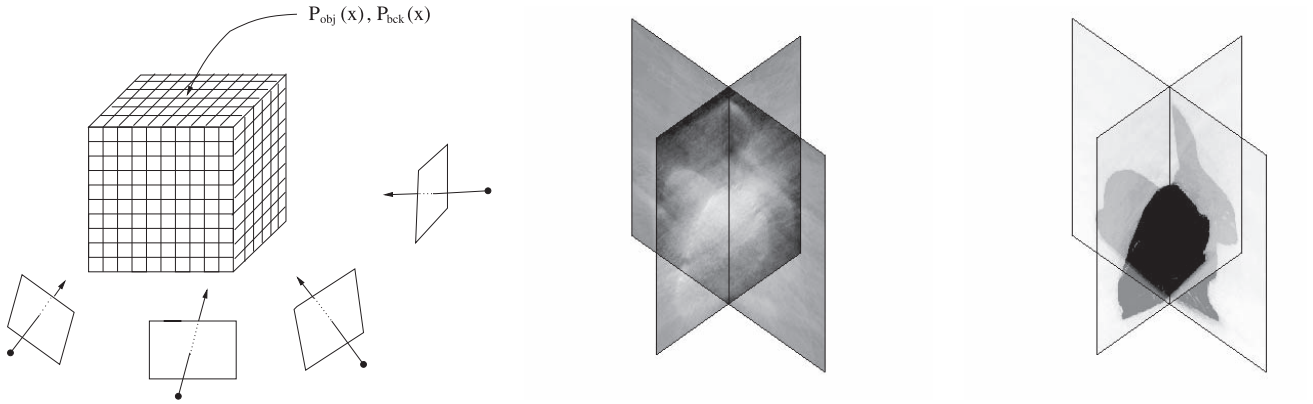

Fig. 2. Probabilistic volume intersection. *Left:* Two probabilities P_{obj}, P_{bck} are assigned to each voxel, explaining its projections onto the images with respect to the provided color models for foreground and background, respectively. *Right:* Slices through the probability maps P_{obj} and P_{bck} for the "bunny" sequence (see Fig. 6).

$I_1, \ldots, I_n$, provided a surface estimate S. To this end, we could rely on the simple and straightforward assumption that observations of separate voxels are independent of each other and only their projections onto the images influence their state. This leads to factorization over the entire volume:

$$P(\{I_1, \ldots, I_n\} \mid S) \approx \left[\prod_{x \in \tilde{V}} P(\{I_l(\pi_l(x))\}_{l=1,\ldots,n} \mid S) \right]^{dx}, \quad (3)$$

where the exponent dx denotes the discretization step and plays the role of a normalizer. It is introduced to ensure the correct continuum limit and make the expression invariant to refinement of the grid. In practice, the probability values are usually smaller than 1. Hence, the above product will tend to zero for increasing volume resolutions since the number of multipliers will grow. For example, if we double the number of voxels, the product will be generally raised to the power of 2. The effect of this modification will be neutralized by the exponent dx, which will be halved.

In fact, the independence assumption is not fulfilled since the appearance of a voxel can be affected by other voxels in the line of sight. However, we neglect this interdependence at this point and reintroduce it at a later stage of the modeling process.[1]

According to a certain surface estimate S, the voxels can be divided into two classes: lying inside an object or belonging to the background. Hence, the volume V can be expressed as $V = R^S_{obj} \cup R^S_{bck}$, where R^S_{obj} denotes the surface interior and R^S_{bck} the exterior region, respectively. Analogously, we obtain for the discrete counterpart $\tilde{V} = R^S_{obj} \cup R^S_{bck}$, where $R^{\tilde{S}}_{obj}$ and $R^{\tilde{S}}_{bck}$ are discretized versions of R^S_{obj} and R^S_{bck}. Considering this partitioning, we can proceed with

$$\begin{aligned} &P(\{I_1, \ldots, I_n\} \mid S) \\ &\quad \approx \left[\prod_{x \in R^{\tilde{S}}_{obj}} P\big(\{I_l(\pi_l(x))\}_{l=1,\ldots,n} \mid x \in R^S_{obj}\big) \right]^{dx} \\ &\quad \cdot \left[\prod_{x \in R^{\tilde{S}}_{bck}} P\big(\{I_l(\pi_l(x))\}_{l=1,\ldots,n} \mid x \in R^S_{bck}\big) \right]^{dx}. \end{aligned} \quad (4)$$

To simplify the notation, we denote

$$\begin{aligned} P_{obj}(x) &:= P\big(\{I_l(\pi_l(x))\}_{l=1,\ldots,n} \mid x \in R^S_{obj}\big), \\ P_{bck}(x) &:= P\big(\{I_l(\pi_l(x))\}_{l=1,\ldots,n} \mid x \in R^S_{bck}\big), \end{aligned} \quad (5)$$

for $x \in V$ (see Fig. 2). Now, plugging the results in (2) and (4) into (1) gives the following expression:

$$\hat{S} = \arg\max_{S \in \Lambda} \left(\prod_{x \in R^{\tilde{S}}_{obj}} P_{obj}(x) \right)^{dx} \cdot \left(\prod_{x \in R^{\tilde{S}}_{bck}} P_{bck}(x) \right)^{dx} \cdot P(S). \quad (6)$$

Note that $P_{obj}(x)$ and $P_{bck}(x)$ defined in (5) do not represent the probability that x is part of object or background, but rather the probability for observing certain colors in respective projections given that x is part of the object or the background. In particular, this implies that for an arbitrary $x \in V$, these probabilities will generally not sum to 1. This is an important point in the modeling process since it allows us to use two different color distribution models for foreground and background instead of a single one.

2.2 Joint Probabilities

Now, we are confronted with the question of how to compute the joint probabilities given in (5). Such a computation involves fusing hypotheses stemming from different views. A straightforward way to accomplish this task is to again assume independence of the image observations. Taking visibility into account, we note that the probability of a voxel being part of the foreground is equal to the probability that all cameras observe this voxel as foreground, whereas the probability of background membership describes the probability of at least one

1. The weaker assumption of a factorization not over all voxels but merely over all lines of sight gives rise to a cost functional with integrals over all image domains, as suggested in [36]. While this approximation is more faithful, it does not lead to a globally optimizable cost functional and does not entail uniqueness of solutions.

camera seeing background. This formulation can be regarded as the probabilistic analog to classical silhouette carving techniques, where a voxel is set transparent if it projects on background in at least one of the input images. Note that this is a conceptual difference to explicit visibility estimation, where the current surface determines the state of each voxel [36]. Following this train of thoughts, we obtain the formulation

$$\begin{aligned} P_{obj}(x) &= \prod_{i=1}^{n} P\big(I_i(\pi_i(x))|x \in R^S_{obj}\big), \\ P_{bck}(x) &= 1 - \prod_{i=1}^{n} \big[1 - P\big(I_i(\pi_i(x))|x \in R^S_{bck}\big)\big]. \end{aligned} \tag{7}$$

The asymmetry in both expressions is due to the fact that they describe different types of events. The expression for $P_{obj}(x)$ relies on the assumption that the observed object is completely visible in all of the images, i.e., no obstacles block the field of view of the cameras to it. The overall foreground score can then be obtained by simple multiplication of all image votes. The term $P_{bck}(x)$ requires more care regarding the fact that a background voxel could be occluded in some of the images by the object itself. Hence, a simple multiplication of the single probabilities will not work. Instead, we revert the foreground evidence of the individual image responses with respect to the background model. In this sense, the interdependence of voxels neglected in (3) is now reintroduced.

A closer look at (7) reveals that it contains a bias with respect to the number n of images. Since the individual observation probabilities $P(I_i(\pi_i(x)) \mid x \in R^S_{obj})$ and $P(I_i(\pi_i(x)) \mid x \in R^S_{bck})$ are both bounded by 1 and typically smaller than 1 for realistic scenarios, $P_{obj}(x)$ (and $P_{bck}(x)$) would tend to zero (or one) for $n \to \infty$. This bias disappears if we consider each camera separately to approximate $P_{obj}(x)$ and $P_{bck}(x)$:

$$\begin{aligned} P_{obj}(x) &\approx P\big(I_i(\pi_i(x)) \mid x \in R^S_{obj}\big) \quad \forall\, i, \\ 1 - P_{bck}(x) &\approx 1 - P\big(I_i(\pi_i(x)) \mid x \in R^S_{bck}\big) \quad \forall\, i, \end{aligned} \tag{8}$$

and subsequently compute the geometric mean as an average score, yielding

$$\begin{aligned} P_{obj}(x) &= \sqrt[n]{\prod_{i=1}^{n} P\big(I_i(\pi_i(x)) \mid x \in R^S_{obj}\big)}, \\ P_{bck}(x) &= 1 - \sqrt[n]{\prod_{i=1}^{n} \big[1 - P\big(I_i(\pi_i(x)) \mid x \in R^S_{bck}\big)\big]}. \end{aligned} \tag{9}$$

A direct comparison to (7) shows that the proposed model results in a normalizing root being introduced, which makes both expressions invariant to the number of cameras. The use of geometric mean is motivated by the nature of the fusion process. For example, if one camera supplies weak evidence for foreground membership (i.e., $P(I_i(\pi_i(x)) \mid x \in R^S_{obj}) \approx 0$), this will immediately decrease the overall product and therewith the final value for P_{obj}. Analogously, a strong background response (i.e., $P(I_i(\pi_i(x)) \mid x \in R^S_{bck}) \approx 1$) of one of the cameras will drastically bring the value of P_{bck} closer to 1. This coincides with the classical visual hull computation, where a voxel is classified as background if at least one of its projections is inside a background region.

The probability of observing a certain color value in a given image can be modeled by a parametric distribution such as multivariate Gaussian:

$$\begin{aligned} P(I_i(\pi_i(x)) \mid x \in R^S_{obj}) &\sim \mathcal{N}(\mu_{obj}, \Sigma_{obj}), \\ P(I_i(\pi_i(x)) \mid x \in R^S_{bck}) &\sim \mathcal{N}(\mu_{bck}, \Sigma_{bck}), \end{aligned} \tag{10}$$

in sRGB color space. Here, μ_{obj}, μ_{bck} denote the mean vectors and Σ_{obj}, Σ_{bck} the covariance matrices of both regions. As previously mentioned, the parameters of the color distributions are determined interactively by requiring the user to mark object and background regions via scribbles in one of the input images (see Section 4). Note that $\mathcal{N}(\mu_{obj}, \Sigma_{obj})$ and $\mathcal{N}(\mu_{bck}, \Sigma_{bck})$ stand for continuous density functions. In order to derive corresponding probability values, a normalization over the entire discretized color space has to be performed. This step is important since it guarantees that the values for $P(I_i(\pi_i(x)) \mid x \in R^S_{obj})$ and $P(I_i(\pi_i(x)) \mid x \in R^S_{bck})$ are within the unit interval $[0, 1]$ and validates the formulation in (9). Example probability maps P_{obj} and P_{bck} are depicted in Fig. 2. Note that the probability for foreground evidence is quite blurry, while that of the background region is more distinct. This is due to the nature of the silhouette fusion scheme (see (9)). As the foreground probability map P_{obj} is estimated by simply averaging the single observation probabilities, the obtained values are diluted. In contrast, one high observation probability with respect to the background model would immediately result in a high value for P_{bck}.

It should be noted that the proposed probabilistic framework is quite general and does not involve any inherent assumptions about particular modeling of the observation probabilities in (10). Alternatively to the proposed formulation, the method in [12] or [32] could be used instead to derive respective probability maps.

3 MAP Estimation via Energy Minimization

3.1 Variational Formulation

Now, we come to the question of how the MAP estimation problem in (6) can be solved. It can be converted to an equivalent energy minimization problem that can be solved exactly by means of established convex relaxation techniques.

A standard approach to achieve that is to apply the negative logarithm, which converts the maximization problem in (6) to a minimization one. In a continuous setting, this yields

$$\begin{aligned} E(S) = &- \int_{R^S_{obj}} \log P_{obj}(x)\, dx \\ &- \int_{R^S_{bck}} \log P_{bck}(x)\, dx - \log P(S), \\ \hat{S} &= \arg\min_{S \in \Lambda} E(S). \end{aligned} \tag{11}$$

Minimizing the above energy functional is equivalent to maximizing the total a posteriori probability of all voxel assignments. In the spirit of energy minimization, the first

two terms can be interpreted as external costs and measure the discrepancy between image observations and projections predicted by the model. The last term exhibits internal energy costs and summarizes prior knowledge on the surface geometry. In order to handle image perturbances like sensor noise, imprecise camera calibration, and background clutter, this term is usually used to impose spatial smoothness of the recovered surface. From a theoretical point of view, a regularization term is often needed to guarantee uniqueness of solutions [24]. This can be achieved by setting

$$P(S) = e^{-\nu|S|}, \tag{12}$$

where ν is a weighting constant and $|S|$ denotes the euclidean surface area. The euclidean metric could be replaced by a more general Riemannian metric so as to impose image edge alignment, for example, [15], [6]. Yet, edge responses are provided separately by individual observations and could degrade the reconstructions, especially in case of noisy or cluttered image data. For that reason, we relied on the simple euclidean metric in our regularization criterion. It should be mentioned that the a priori model in (12) introduces a minimal surface bias, even though it achieves a high degree of smoothness. This could cause problems with thin or elongated structures, depending on the value of the parameter ν. Alternatively, higher order shape characteristics like curvature could be used instead, but this would make the optimization much more challenging. To the best of our knowledge, to date there is no approach allowing global minimization of curvature in 3D. The choice of the minimal surface model in (12) is motivated by its simplicity and global optimizability, as well as its high efficiency in suppressing noise. By plugging (12) into the functional in (11), we finally obtain

$$\begin{aligned} E(S) = &- \int_{R^S_{obj}} \log P_{obj}(x)\, dx \\ &- \int_{R^S_{bck}} \log P_{bck}(x)\, dx + \nu|S|. \end{aligned} \tag{13}$$

Our goal is to minimize this functional.

3.2 Numerical Optimization

Following recent advances in convex optimization [7], we observe that our functional at hand (13) is amenable to global optimization. Important advantages of continuous minimal surface optimization methods over graph cuts are their straightforward parallelizability and accurate regularization scheme [16], [19]. While a lot of efforts have been made to parallelize graph cut algorithms [30], [22], [34], there is usually no theoretical guarantee that the parallelization will be faster for every problem instance [11]. Moreover, graph cut optimization entails metrication errors, which can be resolved by increasing the grid connectivity [4] but at the expense of considerably higher memory requirements. All these difficulties can naturally be circumvented by utilizing continuous alternatives. Indeed, continuous optimization techniques are particularly suitable for our framework and justify the switch to a continuous setting in the modeling process.

The first step is to represent the surface S implicitly by the characteristic function $u : V \to \{0, 1\}$ of R^S_{obj}, i.e., $u = \mathbf{1}_{R^S_{obj}}$ and $1 - u = \mathbf{1}_{R^S_{bck}}$. A known advantage of this representation is that changes in the topology of S are handled automatically without reparametrization. Now, we obtain the following constrained nonconvex energy minimization problem corresponding to (13):

$$\begin{aligned} E(u) = &\int_V \log \frac{P_{bck}(x)}{P_{obj}(x)}\, u(x)\, dx + \nu \int_V |\nabla u|\, dx \to \min, \\ &\text{s.t.} u \in \{0, 1\}. \end{aligned} \tag{14}$$

The minimization problem stated in (14) is nonconvex since the optimization is carried out over a nonconvex set of binary functions. This difficulty can be circumvented by relaxing the set of binary labeling functions to $u \in [0, 1]$. This leads to a typical constrained convex minimization problem for which a globally optimal solution can easily be obtained. A key observation which makes such relaxation techniques interesting is that thresholding its global minimizer at some value within $(0, 1)$ gives a global minimizer of the original nonconvex problem (14) (see [7] for more details). The relaxed problem can be solved globally by any iterative local optimization procedure. Even though the particular choice of minimization method will not affect the final result, it influences the speed of convergence. In our implementation, we apply the primal-dual method proposed in [25], explained in the following in more detail.

One can notice that the energy functional in (14) can be written in the form

$$E(u) = \int_V f u\, dx + \nu \int_V |\nabla u|\, dx, \tag{15}$$

where $f : V \to \mathbb{R}$ summarizes the constant part not dependent on u, i.e.,

$$f := \log \frac{P_{bck}(x)}{P_{obj}(x)}.$$

We proceed by switching to a dual formulation of the total variation regularizer by means of an auxiliary variable $\xi : V \to \mathbb{R}^3$, which allows for the following conversion:

$$E(u) = \int_V f u\, dx + \nu \left(\sup_{|\xi| \leq 1} \int_V \langle \xi, \nabla u \rangle\, dx \right). \tag{16}$$

Now, we obtain a new functional

$$E(u, \xi) = \int_V f u\, dx + \nu \int_V \langle \xi, \nabla u \rangle\, dx \tag{17}$$

that should be minimized with respect to u and maximized with respect to ξ under the constraints $u \in [0, 1]$ and $|\xi| \leq 1$. This states a typical saddle point problem that can be solved by a projected gradient descent/ascent strategy. Denoting by $C_{rel} := \{u \mid u : V \to [0, 1]\}$ the set of relaxed labeling functions and by $K := \{\xi \in \mathbb{R}^3 \mid |\xi| \leq 1\}$ the unit ball, the primal-dual optimization scheme can be described as follows: We choose $(u^0, \xi^0) \in C_{rel} \times K$ and let $\bar{u}^0 = u^0$. We choose two time steps $\tau, \sigma > 0$. Then, we iterate for $n \geq 0$:

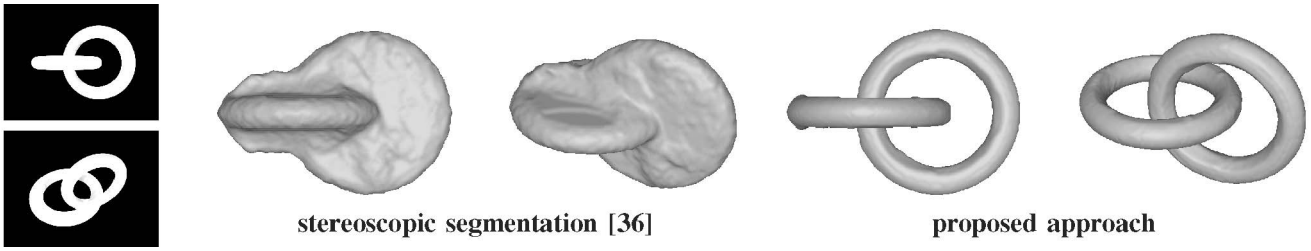

Fig. 3. Tori sequence. *Left:* Two out of 20 synthetic input images of resolution 640×480. *Middle:* Multiple views of the reconstruction with stereoscopic segmentation [36]. *Right:* Corresponding views of the reconstruction obtained by the proposed approach. Computation time: 3.54 seconds. Both methods were initialized with a sphere enclosing the objects. While stereoscopic segmentation gets stuck in a local minimum and completely fails to capture the correct topology, the presented probabilistic fusion scheme accurately recovers the imaged geometry.

$$\begin{aligned} \xi^{n+1} &= \Pi_K(\xi^n + \sigma\, \nabla \bar{u}^n), \\ u^{n+1} &= \Pi_{C_{rel}}(u^n + \tau(\nu\, \mathrm{div}(\xi^{n+1}) - f)), \\ \bar{u}^{n+1} &= 2u^{n+1} - u^n, \end{aligned} \tag{18}$$

where Π_K and $\Pi_{C_{rel}}$ denote projections onto the corresponding sets. Both projections can be easily realized by simple normalization and clipping, respectively. The ∇-operator was discretized by means of forward differences on $\tilde{V}$ and the div-operator—with backward differences so as to ensure correct integration by parts.

For sufficiently small time-step parameters, convergence of the above iterative procedure can be proven [25]. In our experiments, we observed stable behavior for $\tau = \sigma = 0.1$. Moreover, the parameter ν balancing the weighting between data fidelity term and smoothness was fixed to 1.8 throughout all our experiments.

4 Experiments

We demonstrate the viability of the proposed approach on multiple challenging synthetic and real-world image sequences. In particular, we show that the suggested probabilistic fusion scheme can handle objects of arbitrary topology independent of initialization and offers significant robustness to shading effects, camera sensor noise, and background clutter, as frequently encountered in real scenarios.

4.1 Insensitivity to Object Topology

To validate the importance of global optimization, we start with a synthetic image sequence of two coupled tori (see Fig. 3). Although the data set is not interesting from a photometric point of view due to the contrasting appearance of objects and background, it is intriguing from a geometric point of view due to the complex topology of the objects. We compare the proposed approach to stereoscopic segmentation [36], which is an alternative local multiview fusion scheme. As can be expected, the local optimization procedure in [36], involving surface evolution at current contour generators only, is highly sensitive to initialization. As stated in [36], the method requires the initial surface to intersect each of the holes of the final one in order to converge to an accurate result. However, finding such an initialization is not a trivial task since it implies knowledge of the imaged objects. It is not surprising that stereoscopic segmentation completely fails to recover the correct topology starting from a sphere enclosing the two tori. In contrast, the proposed approach, which does not depend on initialization and always guarantees convergence to a global optimum for the provided user input, quite accurately captures the imaged geometry.

4.2 Robustness to Shading Effects and Camera Sensor Noise

The next two experiments, illustrated in Figs. 4 and 6, show the effect of shading effects like shadows and illumination highlights on the 3D reconstruction process.

The first image sequence, depicted in Fig. 4, captures a sow figurine.[2] The data set is relatively challenging, even though it does not create such an impression at first glance. While the figurine is rosy and well distinguishable from the surrounding gray environment, the numerous shading effects like shadows and light reflections adulterate the color and significantly diminish this discrepance. Furthermore, the images exhibit relatively bad color calibration since they have been acquired by different camera devices. Such effects usually cause misclassification of the respective foreground pixels when performing individual image segmentation (see Fig. 5), which in turn leads to overcarving of the subsequently computed visual hull. The proposed probabilistic fusion scheme, which avoids premature hard labeling decisions by exploiting the entire amount of available image information, is designed as a remedy to similar frequently appearing difficulties. We emphasize the benefits of the utilized outline coherency criteria by showing a direct comparison to the classical two-step silhouette integration method [2], [20] (see Fig. 4). In particular, we used the approach in [33] to perform individual image segmentations. It should be mentioned that in addition to regional color cues, this method relies on image edge information to increase the precision of the segmentations. In both cases, we used the same user input in *one* of the images, displayed in Fig. 4, to build the underlying color models. Moreover, in both cases foreground/background distributions were modeled by multivariate Gaussians. Note that even though individual user interaction per view helps to overcome color calibration problems, it doesn't give any substantial improvements in the case of shading effects and considerably increases the interactive efforts required. Expectedly, the independent silhouette fusion technique produces a rather poor reconstruction. This is confirmed by individual image segmentations (see Fig. 10 for an example). In contrast, the proposed probabilistic fusion method produces a quite accurate 3D

2. The data set is publicly available at http://cvpr.in.tum.de/data/datasets/3dreconstruction.

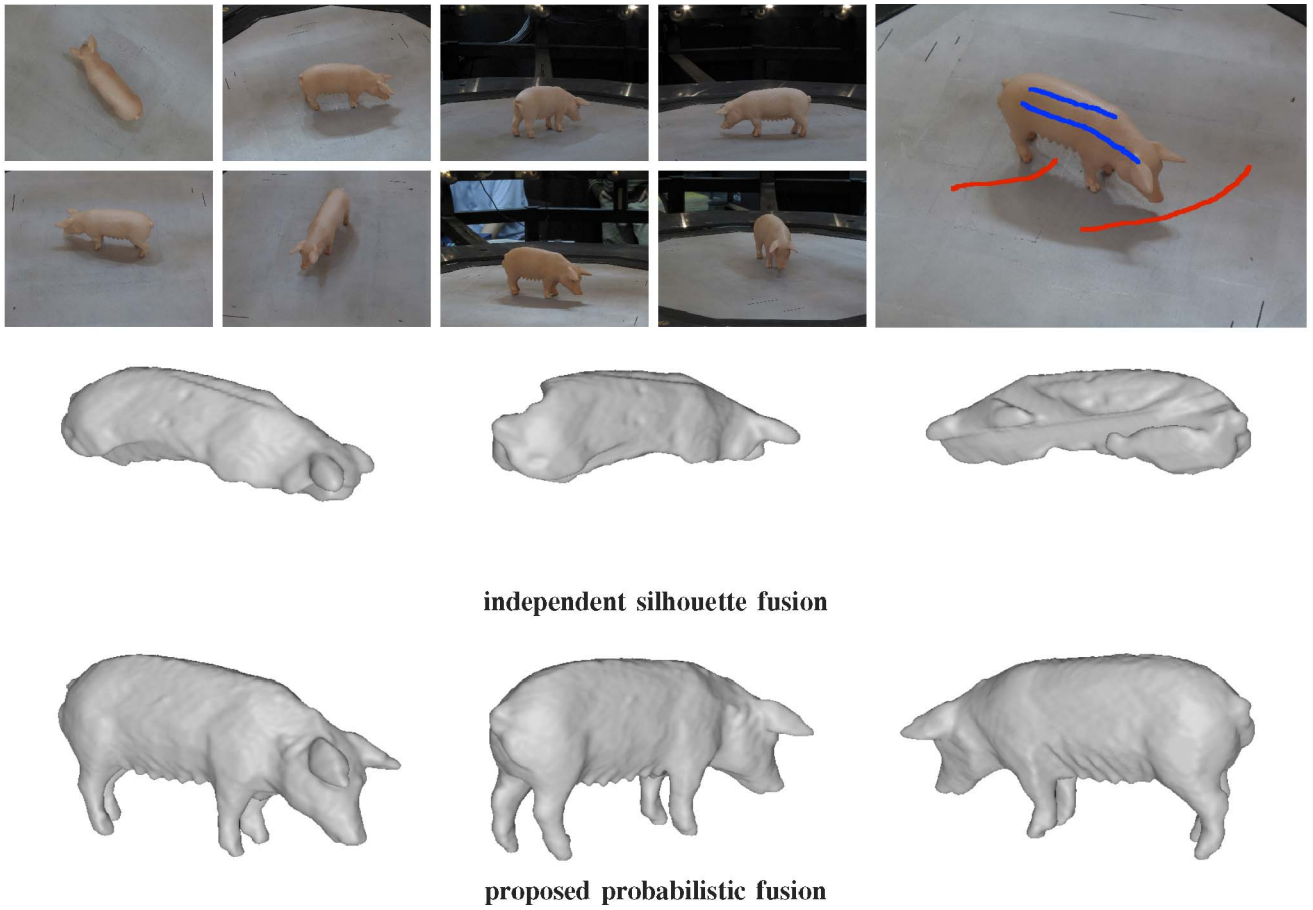

Fig. 4. Sow sequence. *First row:* Nine out of 27 input images of resolution $1{,}024 \times 768$. The utilized user input is highlighted, whereas blue scribbles mark foreground and red—background. *Second row:* Multiple views of the visual hull obtained with the classical independent silhouette fusion technique. *Third row:* Corresponding views of the reconstruction result produced by the proposed approach. Computation time: 2.08 seconds. The numerous shading effects, like shadows and light reflections on the object's surface, as well as the bad color calibration of the cameras lead to relatively poor independent segmentations of individual images (see Fig. 5). This, in turn, results in overcarving of the subsequently computed visual hull. In contrast, the proposed probabilistic fusion method produces a very accurate 3D model under these challenging conditions. The figure is best viewed in color.

model under these challenging conditions. Even some of the small-scale surface details are recognizable.

Similar conclusions can be drawn from the experiment depicted in Fig. 6. The image sequence displays a red ceramic bunny figurine. This time, illumination variations cause less problems due to the diffuse reflectance properties of the material. For that reason, the independent silhouette fusion approach already gives a satisfactory result (see Fig. 6). Most of the small inaccuracies are due to unclean locations on the figurine. We use this data set to investigate the behavior of both methods (the proposed probabilistic fusion and the independent silhouette fusion) in case of camera sensor noise. To this end, noise within a certain range was added randomly to the image color data. We plotted the deviation of the computed 3D model from a given ground truth at increasing noise range (see Fig. 7), measured in units of sRGB color space (color values are within $[0, 255]$). As a ground truth, we used a visual hull of the object computed from manually obtained segmentations. Note that the visual hull is only an approximation of the physical object. Yet, it serves as a ground truth in this case since it exhibits the case of perfect data. If $u_{gt} : V \rightarrow \{0, 1\}$ denotes an implicit labeling representing this ground truth surface (being 1 within the interior region and 0 within the exterior) and $u : V \rightarrow \{0, 1\}$ the obtained 3D labeling, we measure the misalignment between them as

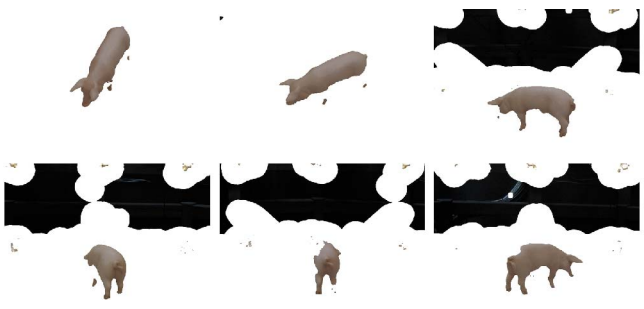

Fig. 5. Independent image segmentations for the sow sequence (6 out of 27) and the user interaction in Fig. 4. False negatives are mainly caused by shading effects like shadows and light reflections, whereas false positives are due to variations in the background color. Expectedly, the poor segmentation results produce a poor 3D model (see Fig. 4). Note, however, that while false positives do not lead to reconstruction inaccuracies in most cases, false negatives have a direct influence due to overcarving along the respective viewing rays. The figure is best viewed in color.

$$\epsilon = \frac{\int_V |u_{gt}(x) - u(x)| \, dx}{\int_V u_{gt}(x) \, dx + \int_V u(x) \, dx}. \tag{19}$$

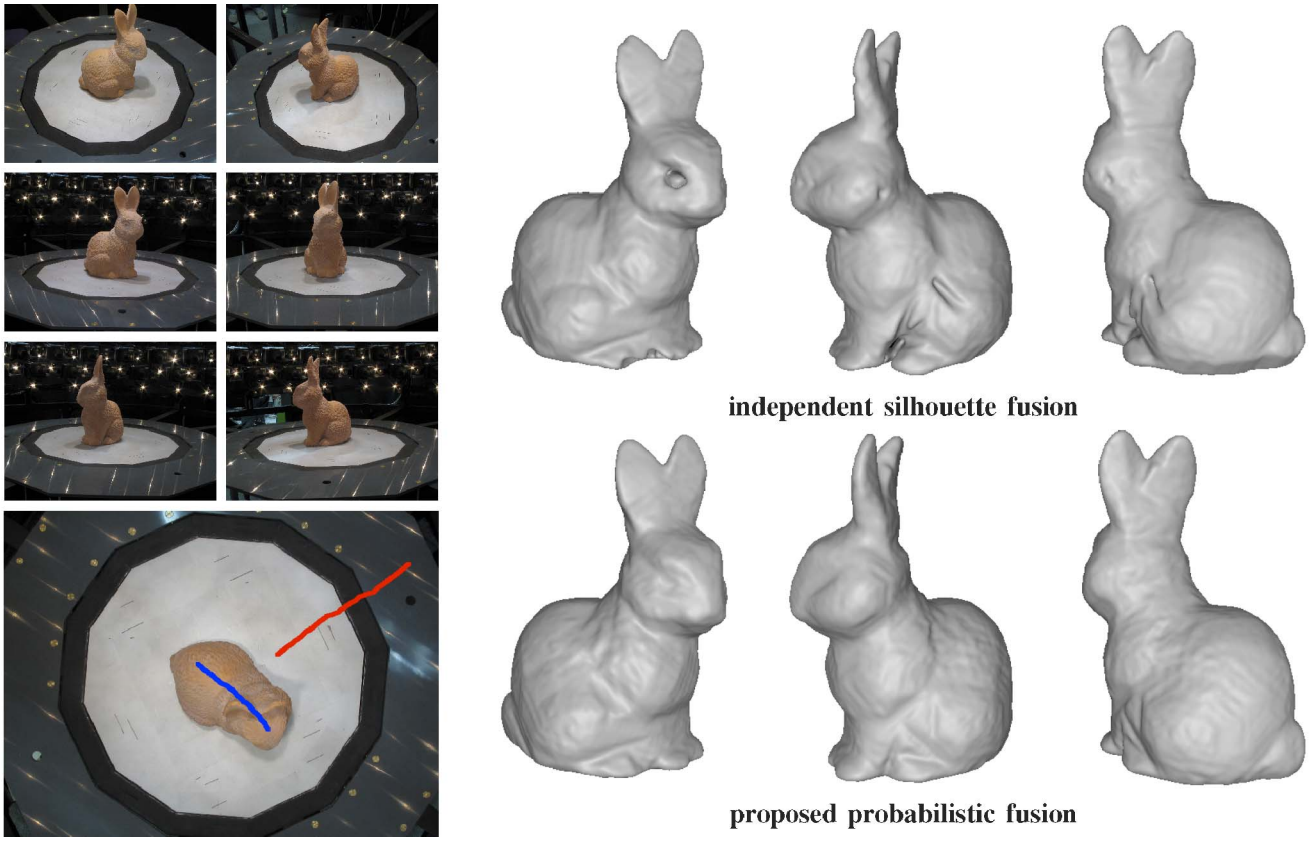

Fig. 6. Bunny sequence. *Left:* Seven out of 36 input images of resolution 640×480. The utilized user input is highlighted, whereas blue scribbles mark the foreground and red the background. *Right, first row:* Multiple views of the visual hull obtained with the classical independent silhouette fusion technique. *Right, second row:* Corresponding views of the reconstruction result produced by the proposed approach. Computation time: 4.41 seconds. Although the traditional silhouette fusion approach gives a relatively accurate reconstruction in this case, some small imprecisions are still notable, in particular caused by unclean locations on the figurine. The concurrent probabilistic fusion scheme produces an impeccable 3D model. The figure is best viewed in color.

In particular, we have $\epsilon \in [0, 1]$ with $\epsilon = 0$ if and only if both reconstructions are identical and $\epsilon = 1$ if u is the empty set. Two important observations can be made when analyzing the graphs in Fig. 7. First, it is evident that the noise levels at which both compared approaches start to degrade are quite different. While the independent silhouette fusion method shows a notable deviation at noise range of 20 color space units, the accuracy of the probabilistic one is unaffected up to noise range of 50 units. The superior resilience to camera sensor noise of the proposed probabilistic formulation is additionally emphasized by its generally smooth behavior for ascending noise levels, which is in contrast to the jumpy performance of its opponent.

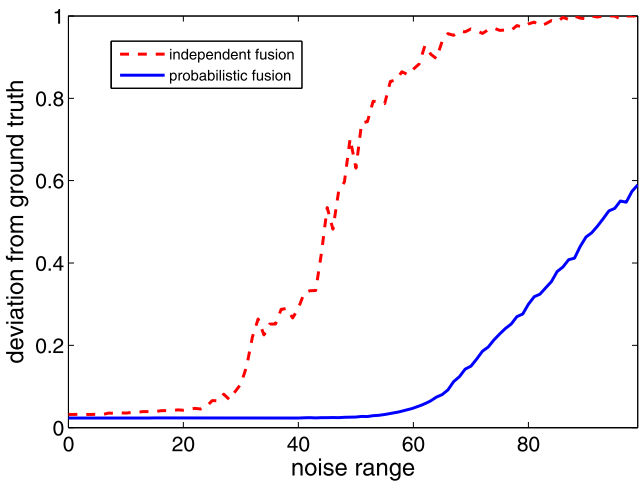

Fig. 7. Robustness to camera sensor noise. The accuracy of the proposed probabilistic approach and the traditional independent silhouette fusion procedure for the image sequence in Fig. 6 is investigated for ascending levels of image noise. The noise is added randomly and measured in terms of its application range in units of sRGB color space. The precision of the reconstruction is computed as the deviation from a provided ground truth surface. See text for more details.

4.3 Robustness to Background Clutter

While the image sequences considered so far capture a more or less homogeneous background, the next two data sets take a further step and increase the degree of difficulty by picturing typical real-world backgrounds spanning a wide range of colors.

The first sequence, depicted in Fig. 8, illustrates a statue, imaged in front of a blue poster, in the Academic Art Museum in Bonn, Germany. Although the poster helps to separate the captured statue from the others in the background, the object is not completely separable in color space due to its similarity to the pedestal. Since the goal was the precise reconstruction of the statue, the pedestal underneath was marked as background by the provided user interaction (see Fig. 8). Expectedly, this diminishes the discriminative power of both color distributions. An

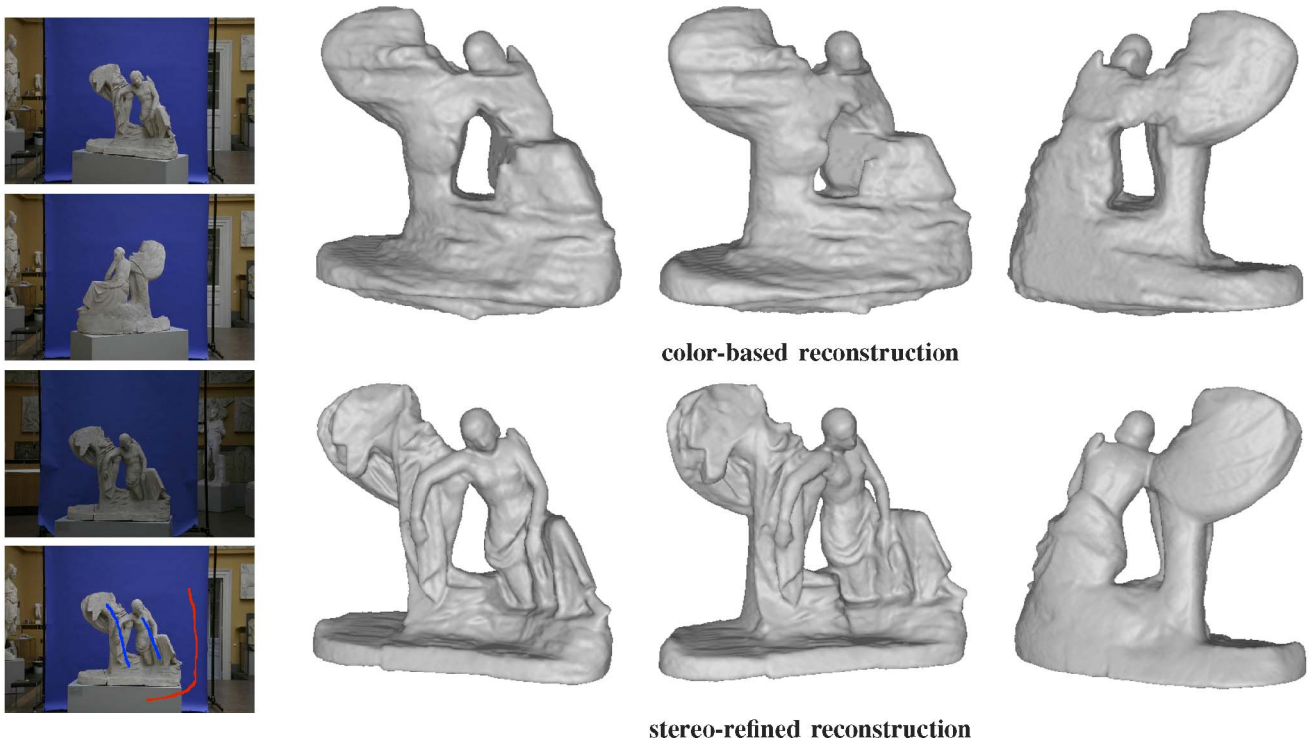

Fig. 8. Statue sequence. *Left:* Four out of 36 images of resolution $1{,}536 \times 1{,}024$. The utilized user interaction is superimposed in the last image (blue scribbles mark the foreground and red the background). *Right, first row:* Multiple views of the estimated color-based reconstruction. *Right, second row:* Stereo-refined reconstruction by the method in [19], initialized with the above result. Note the color similarity between the object and the pedestal as well as the severe intensity variations. The figure is best viewed in color.

additional challenge is posed by the severe intensity variations, which are due to the fact that the photographs were taken at different times of the day. Despite all of these difficulties, the proposed approach produces a relatively accurate reconstruction result, even though it exhibits some small imprecisions (e.g., at the basement). In fact, the estimated 3D model turns out to be precise enough to initialize a stereo-refinement process and obtain a highly accurate 3D model (see Fig. 8). Thereby, for the stereo-based reconstruction, we used the method in [19].

The second image sequence, depicted in Fig. 9, displays a bronze bust sculpture of Robert Sauer. As can be seen from the example pictures, the background continually changes including the surrounding building interior and hence a very wide range of colors. This significantly exacerbates the separability of the sculpture in individual images, even though most of the objects in the background are relatively far apart from it. Additional difficulties are caused by the complex reflectance properties of the material. Once again, the proposed approach produces a quite accurate result under these challenging conditions. Even though the reconstruction exhibits some small-scale artifacts (e.g., at the basement) and some oversmoothing effects (e.g., the spectacle frame), the shape of the bust is clearly recognizable.

The accuracy of the computed 3D models is confirmed by the image segmentations obtained by projecting them onto the input views (see Fig. 10). In case of background clutter, this leads to dramatic improvements over the naive

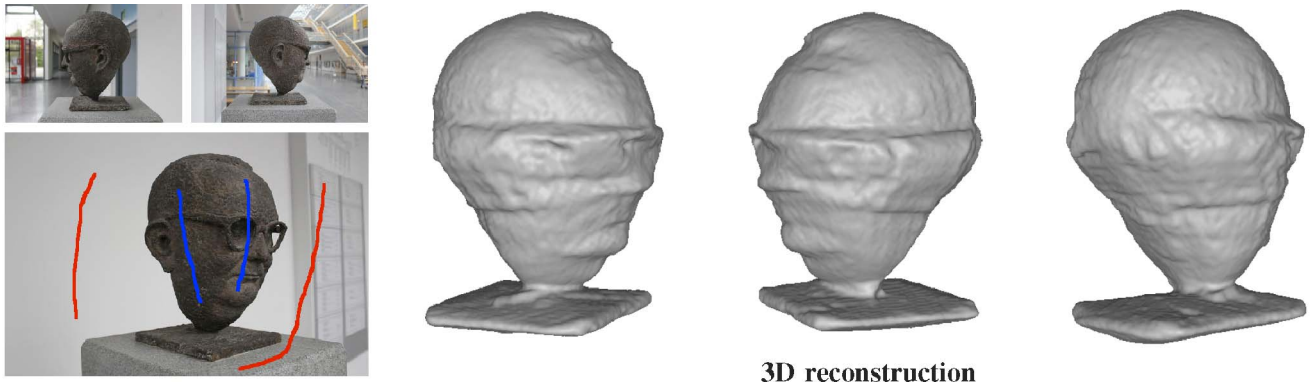

Fig. 9. Bust sequence. *Left:* Three out of 36 input images of resolution $1{,}296 \times 864$ and superimposed user interaction (blue scribbles mark the foreground and red the background). *Right:* Multiple views of the 3D reconstruction obtained with the proposed approach. Computation time: 4.26 seconds. Note the wide range of background colors as well as the complex reflectance properties of the material. The figure is best viewed in color.

Fig. 10. Segmentation of individual images of the sequences in Figs. 4, 6, 8, and 9. *First column:* One of the input images. *Second column:* Interactive segmentation with the method in [33]. *Third column:* Interactive segmentation with the proposed approach, obtained by projecting the computed 3D model onto the image. Even though the estimated silhouette-coherent segmentations are not pixel precise due to the use of 3D regularization and the discrepancy between image resolution and volumetric resolution, the silhouettes are registered accurately and offer dramatic improvements over independent 2D segmentations.

isolated segmentation approach and clearly demonstrates the potential of the proposed probabilistic silhouette coherency criteria. This observation is additionally emphasized by a quantitative evaluation over the entire image sequences, shown in Fig. 11. To this end, ground truth segmentations were obtained by labeling the images manually. The segmentation error was computed as

$$err = \frac{p_{false}}{p_{true} + p_{false}}, \tag{20}$$

where p_{true} and p_{false} denote the number of correctly classified and misclassified pixels in all views, respectively. Note that $err \in [0, 1]$. The independent segmentation method demonstrates poor performance for all data sets except for the "bunny" sequence due to shading effects, illumination variations, and background clutter. In contrast, the proposed probabilistic fusion approach shows clear superiority and gives accuracy improvements ranging from factor 3 (for the "bunny" sequence) to factor 46 (for the "statue" sequence). Note that while the segmentation error is negligible for the "sow" and "bunny" sequences, acquired in lab conditions, it increases for the "statue" and "bust" sequences, generated in more complex environments, but to an acceptable extent. These results provide an explicit justification for the exploration of various coherency criteria in the context of multiview segmentation.

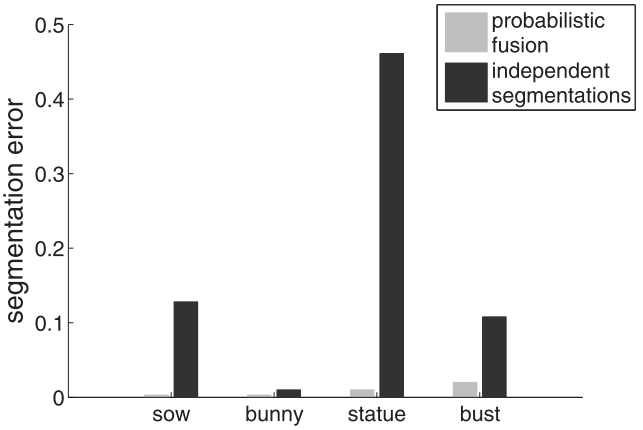

Fig. 11. Accuracy of individual image segmentations. The proposed probabilistic approach is compared to the naive independent segmentation method. The probabilistic fusion scheme was evaluated by projecting the final 3D model onto the image planes. Ground truth segmentations were obtained by labeling the images manually. Note the tremendous improvement in segmentation accuracy, achieved by exploiting probabilistic silhouette coherency criteria.

4.4 User Interaction

The provided user interaction is visualized for all real experiments in Figs. 4, 6, 8, and 9. It is evident that the proposed approach gets by with only a few roughly specified scribbles in *one* of the input images. This suggests that the method is not very sensitive to user intervention, which has been confirmed in our experiments.

As previously mentioned, we relied on single Gaussians in our modeling since all of our test objects are single colored. Multivariate Gaussians minimize the interaction efforts while achieving a substantial degree of robustness to model deviations. We also experimented with Gaussian mixture models. However, we observed that the results gradually degrade for more than two mixture modes due to overfitting effects. Note that the user-specified scribbles occupy only a small portion compared to the entire amount of pixel data. Yet, Gaussian mixture models or kernel density estimation can still be preferable in case of multicolored objects.

It should be emphasized that all demonstrated data sets could successfully be handled, e.g., by the independent silhouette fusion scheme or the method in [6] with the appropriate amount of user interaction on a per-view basis. In contrast, the proposed approach stands out by its capability of producing an accurate reconstruction from only a few scribbles in *one* of the input images. This property reveals its high practical value, especially in case of long sequences containing multiple hundreds or thousands of photographs.

4.5 Computational Time

As previously mentioned, the proposed approach was designed with focus on not only robustness but also computational efficiency. In particular, we make use of recent progress in parallel computing with a GPU implementation of the method. Note that its ingredients enable parallelization over the volume grid since all involved computations are at a voxel basis. Moreover, it can be observed that the overall computational time scales linearly with both the number of input images and the volume resolution. Runtimes for all demonstrated

TABLE 1
Data Sets and Runtimes
for All Demonstrated Experiments

	# images	image resolution	runtime GPU
tori	20	640 × 480	3.54 s
sow	27	1024 × 768	2.08 s
bunny	36	640 × 480	4.41 s
statue	36	1536 × 1024	3.95 s
bust	36	1296 × 864	4.26 s

The computational times were measured on an NVIDIA Tesla C2070 graphics card.

experiments, measured on a NVIDIA Tesla C2070, can be found in Table 1. In our GPU implementation, we exclusively used global memory to store all input images and volumetric data. Even though we tried to employ shared memory in the optimization step, exploiting the neighboring structure of the underlying PDEs (18), this didn't lead to a notable runtime reduction. Note that the computational time of the presented method does not depend on the image resolution (ignoring the time for loading the images) but only on the number of views. In all test cases, volumetric resolution was in the range between 8 and 21 million voxels.

It should be recalled that the input of the proposed approach consists not only of the image sequence and the provided user interaction but also of a specification of a bounding box containing the object of interest. Although a tight specification is not necessary for the method to work, it influences the precision of the computed 3D model and hence the computational time (a loose bounding box requires a high volume resolution). One way to obtain a bounding box estimate is to use the 3D point cloud, produced by classical structure-from-motion techniques which are needed to calibrate the input views.

5 Conclusion

We presented a novel energy minimization approach for interactive joint silhouette extraction and 3D reconstruction from a number of calibrated 2D camera views. The energy model is derived from a probabilistic setting via Bayesian inference and is optimized globally using convex relaxation. The probabilistic formulation avoids making hard decisions about silhouette occupancy based on single views and allows us to optimally take into account color information from all input images. In addition, it provides a novel decoupling scheme to account for the interdependence between voxels, which gives rise to a Bayesian inference problem and allows to compute the globally optimal reconstruction. We experimentally demonstrated that the proposed method compares favorably to state-of-the-art silhouette-based reconstruction methods in that it is more robust to noise, background clutter, shading effects, and camera sensor perturbations. Moreover, it does not require initialization and therefore easily handles 3D shapes of complex topology. Making use of a GPU implementation, robust interactive reconstructions were computed with runtimes of up to 4.41 seconds.

References

[1] X. Bai and G. Sapiro, "A Geodesic Framework for Fast Interactive Image and Video Segmentation and Matting," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[2] B. Baumgart, "Geometric Modeling for Computer Vision," PhD thesis, Dept. of Computer Science, Stanford Univ., 1974.

[3] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr, "Interactive Image Segmentation Using an Adaptive GMMRF Model," *Proc. European Conf. Computer Vision,* pp. 428-441, 2004.

[4] Y. Boykov and V. Kolmogorov, "Computing Geodesics and Minimal Surfaces via Graph Cuts," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 26-33, 2003.

[5] Y.Y. Boykov and M.P. Jolly, "Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D Images," *Proc. IEEE Int'l Conf. Computer Vision,* vol. 1, pp. 105-112, 2001.

[6] N.D.F. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, "Automatic 3D Object Segmentation in Multiple Views Using Volumetric Graph-Cuts," *Proc. 18th British Machine Vision Conf.,* vol. 1, pp. 530-539, 2007.

[7] T. Chan, S. Esedolu, and M. Nikolova, "Algorithms for Finding Global Minimizers of Image Segmentation and Denoising Models," *SIAM J. Applied Math.,* vol. 66, no. 5, pp. 1632-1648, 2006.

[8] A.X. Falcao, J. Stolfi, and R.A. Lotufo, "The Image Foresting Transform: Theory, Algorithms, and Applications," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 1, pp. 19-29, Jan. 2004.

[9] J.-S. Franco and E. Boyer, "Fusion of Multi-View Silhouette Cues Using a Space Occupancy Grid," *Proc. IEEE Int'l Conf. Computer Vision,* 2005.

[10] Y. Furukawa and J. Ponce, "Accurate, Dense, and Robust Multi-View Stereopsis," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* June 2007.

[11] L.M. Goldschlager, R.A. Shaw, and J. Staples, "The Maximum Flow Problem Is Log Space Complete for p," *Theoretical Computer Science,* vol. 21, pp. 105-111, 1982.

[12] L. Grady, "Random Walks for Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 11, pp. 1768-1783, Nov. 2006.

[13] L. Guan, J.S. Franco, and M. Pollefeys, "3D Occlusion Inference from Silhouette Cues," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2007.

[14] C. Hernandez and F. Schmitt, "Silhouette and Stereo Fusion for 3D Object Modeling," *Computer Vision and Image Understanding,* vol. 96, no. 3, pp. 367-392, 2004.

[15] R. Keriven, "A Variational Framework to Shape from Contours," Technical Report 2002-221, CERMICS, 2002.

[16] M. Klodt, T. Schoenemann, K. Kolev, M. Schikora, and D. Cremers, "An Experimental Comparison of Discrete and Continuous Shape Optimization Methods," *Proc. European Conf. Computer Vision,* Oct. 2008.

[17] K. Kolev, T. Brox, and D. Cremers, "Robust Variational Segmentation of 3D Objects from Multiple Views," *Proc. DAGM Symp. Pattern Recognition,* K. Franke et al., eds., pp. 688-697, Sept. 2006.

[18] K. Kolev and D. Cremers, "Integration of Multiview Stereo and Silhouettes via Convex Functionals on Convex Domains," *Proc. European Conf. Computer Vision,* Oct. 2008.

[19] K. Kolev, M. Klodt, T. Brox, and D. Cremers, "Continuous Global Optimization in Multiview 3D Reconstruction," *Int'l J. Computer Vision,* vol. 84, no. 1, pp. 80-96, Aug. 2009.

[20] A. Laurentini, "The Visual Hull Concept for Visual-Based Image Understanding," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 16, no. 2, pp. 150-162, Feb. 1994.

[21] Y. Li, J. Sun, and H.-Y. Shum, "Video Object Cut and Paste," *ACM Trans. Graphics,* vol. 24, no. 3, pp. 595-600, 2005.

[22] J. Liu and J. Sun, "Parallel Graph-Cuts by Adaptive Bottom-Up Merging," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2010.

[23] W.N. Martin and J.K. Aggarwal, "Volumetric Descriptions of Objects from Multiple Views," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 5, no. 2, pp. 150-158, Mar. 1983.

[24] D. Mumford and J. Shah, "Optimal Approximations by Piecewise Smooth Functions and Associated Variational Problems," *Comm. Pure and Applied Math.,* vol. 42, pp. 577-685, 1989.

[25] T. Pock, D. Cremers, H. Bischof, and A. Chambolle, "An Algorithm for Minimizing the Piecewise Smooth Mumford-Shah Functional," *Proc. IEEE Int'l Conf. Computer Vision,* 2009.

[26] M. Potmesil, "Generating Octree Models of 3D Objects from Their Silhouettes from a Sequence of Images," *Computer Vision, Graphics, and Image Processing*, vol. 40, no. 1, pp. 1-29, 1987.
[27] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive Foreground Extraction Using Iterated Graph Cuts," *ACM Trans. Graphics*, vol. 23, no. 3, pp. 309-314, 2004.
[28] C. Rother, V. Kolmogorov, T. Minka, and A. Blake, "Cosegmentation of Image Pairs by Histogram Matching—Incorporating a Global Constraint into MRFs," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pp. 993-1000, 2006.
[29] D. Snow, P. Viola, and R. Zabih, "Exact Voxel Occupancy with Graph Cuts," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 345-353, 2000.
[30] P. Strandmark and F. Kahl, "Parallel and Distributed Graph Cuts by Dual Decomposition," *Proc. Int'l Conf. Computer Vision and Pattern Recognition*, 2010.
[31] R. Szeliski, "Rapid Octree Construction from Image Sequences," *Computer Vision, Graphics, and Image Processing*, vol. 58, no. 1, pp. 23-32, 1993.
[32] J.K. Udupa and P.K. Saha, "Fuzzy Connectedness and Image Segmentation," *Proc. IEEE*, vol. 91, no. 10, pp. 1649-1669, Oct. 2003.
[33] M. Unger, T. Pock, D. Cremers, and H. Bischof, "TVSeg—Interactive Total Variation Based Image Segmentation," *Proc. British Machine Vision Conf.*, Sept. 2008.
[34] V. Vineet and P.J. Narayanan, "Cuda Cuts: Fast Graph Cuts on the Gpu," *Proc. Computer Vision and Pattern Recognition Workshop*, pp. 1-8, 2008.
[35] J. Wang, P. Bhat, R.A. Colburn, M. Agrawala, and M.F. Cohen, "Interactive Video Cutout," *ACM Trans. Graphics*, vol. 24, no. 3, pp. 585-594, 2005.
[36] A. Yezzi and S. Soatto, "Stereoscopic Segmentation," *Proc. Eighth IEEE Int'l Conf. Computer Vision*, vol. 1, pp. 59-66, July 2001.

Kalin Kolev received the BS and MS (Diplom) degrees in computer science from the University of Bonn, Germany, in 2002 and 2005, respectively. Since January 2006, he has been working toward the PhD degree in the Computer Vision Group at the University of Bonn (until November 2009) and TU München (since December 2009). His research interests include multiview 3D reconstruction, statistical approaches, and continuous optimization.

Thomas Brox received the PhD degree in computer science from Saarland University, Germany, in 2005. Subsequently, he spent two years as a postdoctoral researcher at the University of Bonn, Germany, and one year as a temporary faculty member at the University of Dresden, Germany. He was a postdoctoral fellow in the Computer Vision Group of Jitendra Malik at the University of California, Berkeley, for two years. Since 2010, he has headed the Computer Vision Group at the Albert-Ludwigs-University Freiburg. His research interest is in computer vision with special focus on video analysis, particularly optical flow estimation, motion segmentation, learning and detection in videos. In 2004, he received the Longuet-Higgins Best Paper Award at ECCV for his work on optical flow estimation.

Daniel Cremers received the MS (Diplom) degree in theoretical physics (1997) from the University of Heidelberg and the PhD degree in computer science (2002) from the University of Mannheim, Germany. Subsequently, he spent two years as a postdoctoral researcher at the University of California, Los Angeles, and one year as a permanent researcher at Siemens Corporate Research in Princeton, New Jersey. From 2005 until 2009, he headed the Computer Vision Group at the University of Bonn, Germany. Since 2009, he has been a full professor at TU München. He has received several awards, in particular the Best Paper of the Year 2003 by the Pattern Recognition Society, the Olympus Award 2004, and the 2005 UCLA Chancellor's Award.