Chapter 9
Variational Multiview Reconstruction

Multiple View Geometry
Summer 2024

Prof. Daniel Cremers
Chair of Computer Vision and Artificial Intelligence
School of Computation, Information and Technology
Technical University of Munich

Overview

1. Shape Representation and Optimization
2. Variational Multiview Reconstruction
3. Super-resolution Texture Reconstruction
4. Space-Time Reconstruction from Multiview Video

Variational Methods for Dense Reconstruction

The key idea of variational methods is to cast a given problem as a problem of optimization over a continuous space of variables. They are called variational because one determines the solution by varying an initial estimate of the solution so as to reduce the cost or loss function.

The simplest example of an infinite-dimensional variational method is a denoising technique where for a given input image $f : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$, the denoised image is given by the minimizer of a functional of the form

$$E(u) = \int_\Omega \left(f(x) - u(x)\right)^2 + \lambda \left|\nabla u(x)\right|^2 d^2x.$$

It corresponds to a function that is both close to the input image f, but also spatially smooth (weighted by $\lambda > 0$) in the sense that the spatial gradient $\nabla u(x)$ is small.

In this chapter, we will discuss how dense multi-view reconstruction can be cast as a variational method.

Shape Optimization

Shape optimization is a field of mathematics that is focused on formulating the estimation of geometric structures by means of optimization methods.

Among the major challenges in this context is the question how to mathematically represent shape. The choice of representation entails a number of consequences, in particular regarding the question of how efficiently one can store geometric structures and how efficiently one can compute optimal geometry.

There exist numerous representations of shape which can loosely be grouped into two classes:

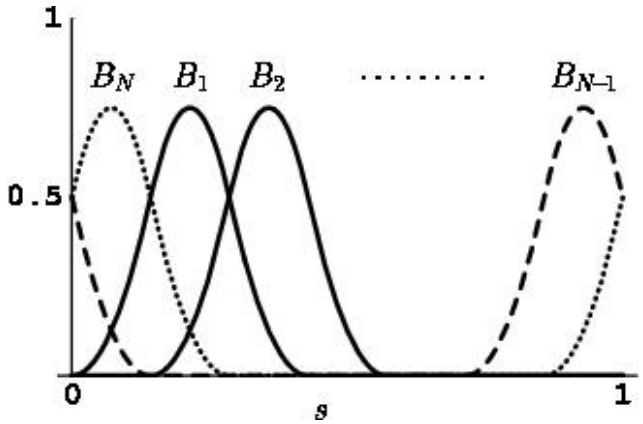
- Explicit representations: The points of a surface are represented explicitly (directly), either as a set of points, a polyhedron or a parameterized surface.
- Implicit representations: The surface is represented implicity by specifying the parts of ambient space that are inside and outside a given surface.

Explicit Shape Representations

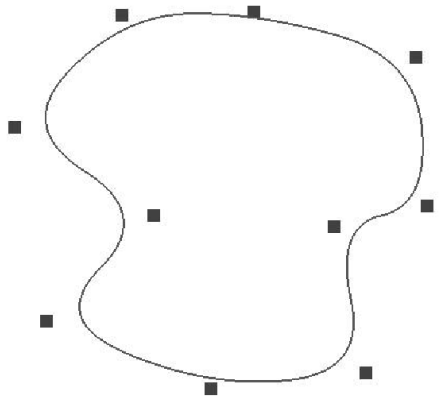
An explicit representations of a closed curve $C \subset \mathbb{R}^d$ is a mapping $C : \mathbb{S}^1 \to \mathbb{R}^d$ from the circle $\mathbb{S}^1$ to $\mathbb{R}^d$. Examples are polygons or – more generally – spline curves:

$$C(s) = \sum_{i=1}^{N} C_i \, B_i(s),$$

where $C_1, \ldots, C_N \in \mathbb{R}^d$ denote control points and $B_1, \ldots, B_N : \mathbb{S}^1 \to \mathbb{R}$ denote a set of spline basis functions:



basis functions



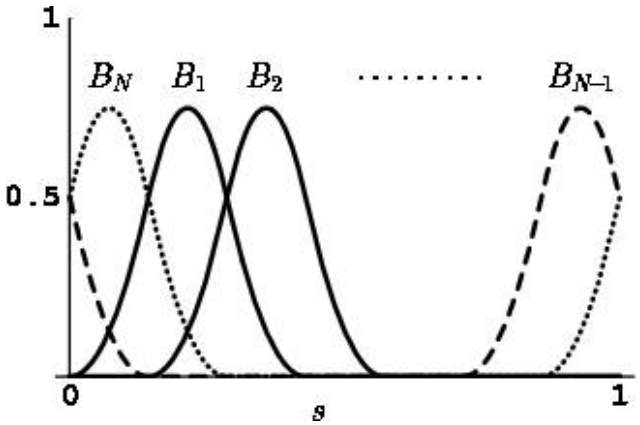
spline & control points

Explicit Shape Representations

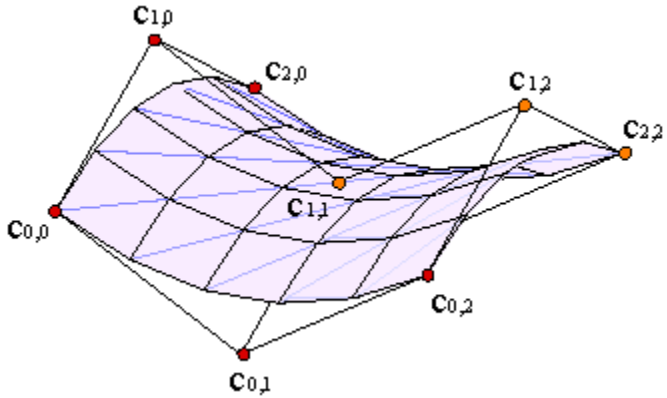
Splines can be extended from curves to surfaces or higher dimensional structures. A spline surface $S \subset \mathbb{R}^d$ can be defined as:

$$S(s,t) = \sum_{i,j} C_{i,j}\, B_i(s) B_j(t),$$

where $C_{i,j} \in \mathbb{R}^d$ denote control points and $B_1, \ldots, B_N : [0,1] \to \mathbb{R}$ denote a set of spline basis functions. Depending on whether the surface is closed or open these basis functions will have a cyclic nature (as below) or not:



basis functions



spline surface & cntrl. points

Implicit Shape Representations

One example of an implicit representation is the indicator function of the surface S, which is a function $u : V \to \{0, 1\}$ defined on the surrounding volume $V \subset \mathbb{R}^3$ that takes on the values 1 inside the surface and 0 outside the surface:

$$u(x) = \begin{cases} 1, & \text{if } x \in \text{int}(S) \\ 0, & \text{if } x \in \text{ext}(S) \end{cases}$$

Another example is the signed distance function $\phi : V \to \mathbb{R}$ which assigns all points in the surrounding volume the (signed) distance from the surface S:

$$\phi(x) = \begin{cases} +d(x, S), & \text{if } x \in \text{int}(S) \\ -d(x, S), & \text{if } x \in \text{ext}(S) \end{cases}$$

Depending on the application it may be useful to know for every voxel how far it is from the surface. Signed distance functions can be computed in polynomial time.

Explicit Versus Implicit Representations

In general, compared to explicit rerpresentations the implicit representations have the following strengths and weaknesses:

- \- Implicit representations typically require more memory in order to represent a geometric structure at a specific resolution. Rather than storing a few points along the curve or surface, one needs to store an occupancy value for each volume element.
- \- Moving or updating an implicit representation is typically slower: rather than move a few control points, one needs to update the occupancy of all volume elements.
- \+ Methods based on implicit representations do not depend on a choice of parameterization.
- \+ Implicit representations allow to represent objects of arbitrary topology (i.e. the number of holes is arbitrary).
- \+ With respect to an implicit representation many shape optimization challenges can be formulated as convex optimization problems and can then be optimized globally.

Multiview Reconstruction as Shape Optimization

How can we cast multiple view reconstruction as a shape optimization problem? To this end, we will assume that the camera poses (rotations and translations) are given.
Rather than estimate the correspondence between all pairs of pixels in either image we will simply ask:

How likely is a given voxel x on the object surface S?

If the voxel $x \in V$ of the given volume $V \subset \mathbb{R}^3$ was on the surface then (up to visibility issues) the projection of that voxel into each image should give rise to the same color (or local texture). Thus we can assign to each voxel $x \in V$ a so-called photoconsistency function

$$\rho : V \to [0, 1],$$

which takes on low values (near 0) if the projected voxels give rise to the same color (or local texture) and high values (near 1) otherwise.

A Weighted Minimal Surface Approach

The reconstruction from multiple views can now be formulated as finding the maximally photoconsistent surface, i.e. a surface S_{opt} with an overall minimal photoconsistency score:

$$S_{opt} = \arg\min_S \int_S \rho(s) ds. \tag{1}$$

This seminal formulation was proposed among others by Faugeras & Keriven (1998). Many good reconstructions were computed by starting from an initial guess of S and locally minimizing this energy using gradient descent. But can we compute the global minimum?

The above energy has a central drawback:

The global minimizer of (1) is the empty set.

It has zero cost while all surfaces have a non-negative energy. This short-coming of minimal surface formulations is often called the shrinking bias. How can we prevent the empty set?

Imposing Silhouette Consistency

Assume that we additionally have the silhouette S_i of the observed $3D$ object outlined in every image $i = 1, \ldots, n$. Then we can formulate the reconstruction problem as a constrained optimization problem (Cremers, Kolev, PAMI 2011):

$$\min_S \int_S \rho(s) ds, \quad \text{such that } \pi_i(S) = S_i \; \forall i = 1, \ldots, n.$$

Written in the indicator function $u : V \to \{0, 1\}$ of the surface S this reads:

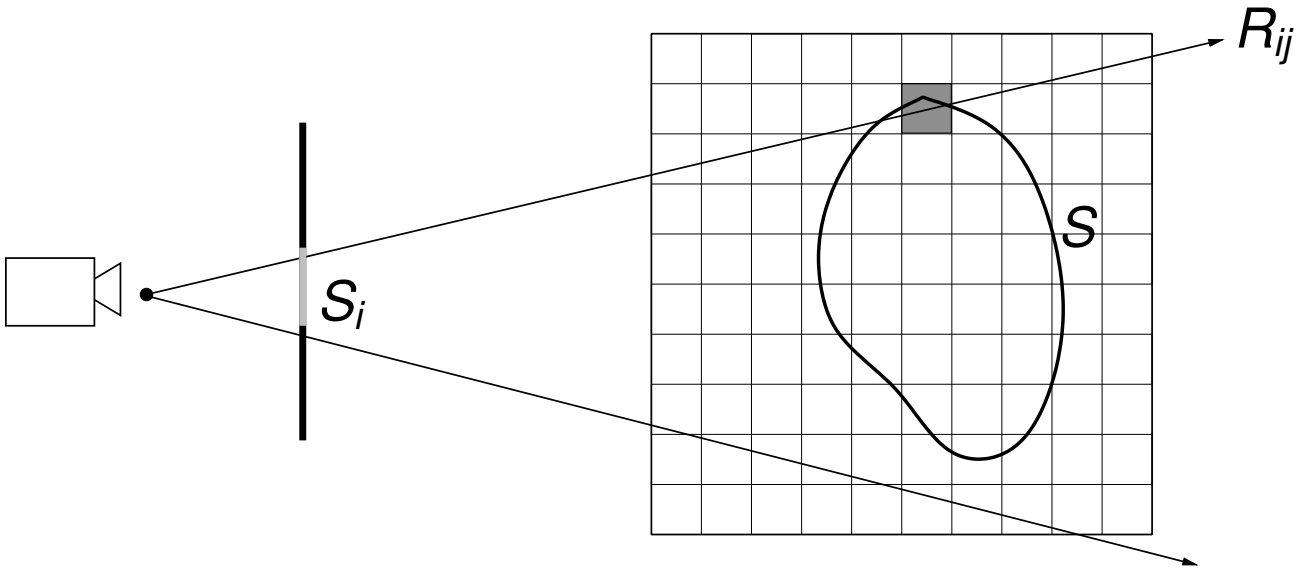
$$\min_{u: V \to \{0,1\}} \int_V \rho(x) |\nabla u(x)| \, dx$$

$$\text{s. t.} \quad \int_{R_{ij}} u(x) \, dR_{ij} \geq 1, \text{ if } j \in S_i \qquad (*)$$

$$\int_{R_{ij}} u(x) \, dR_{ij} = 0, \text{ if } j \notin S_i,$$

where R_{ij} denotes the visual ray through pixel j of image i.

Imposing Silhoutte Consistency



Top view of the geometry and respective visual rays.

Any ray passing through the silhoutte must intersect the object in at least one voxel.

Any ray passing outside the silhouette may not intersect the object in any pixel.

Cremers, Kolev, PAMI 2011

Convex Relaxation and Thresholding

By relaxing the binarity constraint on u and allowing intermediate values between 0 and 1 for the function u, the optimization problem $(*)$ becomes convex.

Proposition

The set

$$\mathcal{D} := \left\{ u : V \to [0,1] \;\middle|\; \begin{array}{ll} \int_{R_{ij}} u(x)\, dR_{ij} \geq 1 & \text{if } j \in S_i \;\forall i,j \\ \int_{R_{ij}} u(x)\, dR_{ij} = 0 & \text{if } j \notin S_i \;\forall i,j \end{array} \right\}$$

of silhouette consistent functions is convex.

Proof.

For a proof we refer to Kolev, Cremers, ECCV 2008. □

Thus we can compute solutions to the silhouette constrained reconstruction problem by solving the relaxed convex problem and subsequently thresholding the computed solution.

Reconstructing Complex Geometry



3 out of 33 input images of resolution 1024×768
Data courtesy of Y. Furukawa.

Reconstructing Complex Geometry

Estimated multiview reconstruction

Cremers, Kolev, PAMI 2011

Reconstruction from a Handheld Camera

2/28 images

Estimated multiview reconstruction

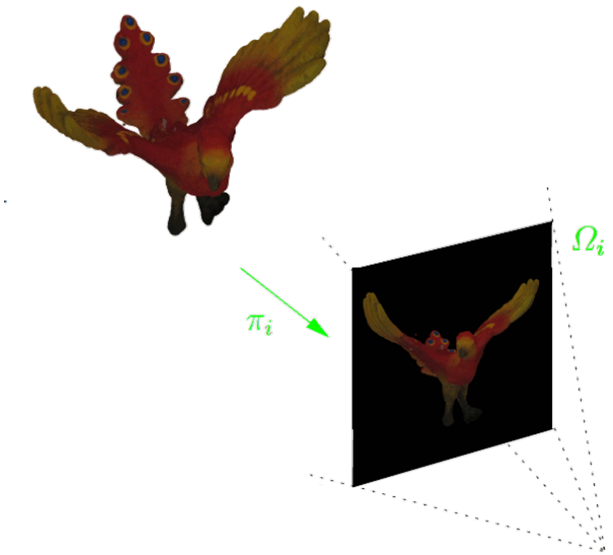
Cremers, Kolev, PAMI 2011

Multi-view Texture Reconstruction

In addition to the dense geometry S, we can also recover the texture $T : S \to \mathbb{R}^3$ of the object from the images $\mathcal{I}_i : \Omega_i \to \mathbb{R}^3$. Rather than simply back-projecting respective images onto the surface, Goldlücke & Cremers ICCV 2009 suggest to solve a variational super-resolution approach of the form:

$$\min_{T:S\to\mathbb{R}^3} \sum_{i=1}^{n} \int_{\Omega_i} \left(B\left(T \circ \pi_i^{-1}\right) - \mathcal{I}_i \right)^2 dx \; + \; \lambda \int_S \|\nabla_S \, T\| ds,$$

where B is a linear operator representing blurring and downsampling and π_i denotes the projection onto image Ω_i:



Multi-view Texture Reconstruction

The super-resolution texture estimation is a convex optimization problem which can be solved efficiently. It generates a textured model of the object which cannot be distinguished from the original:

One of 36 input images

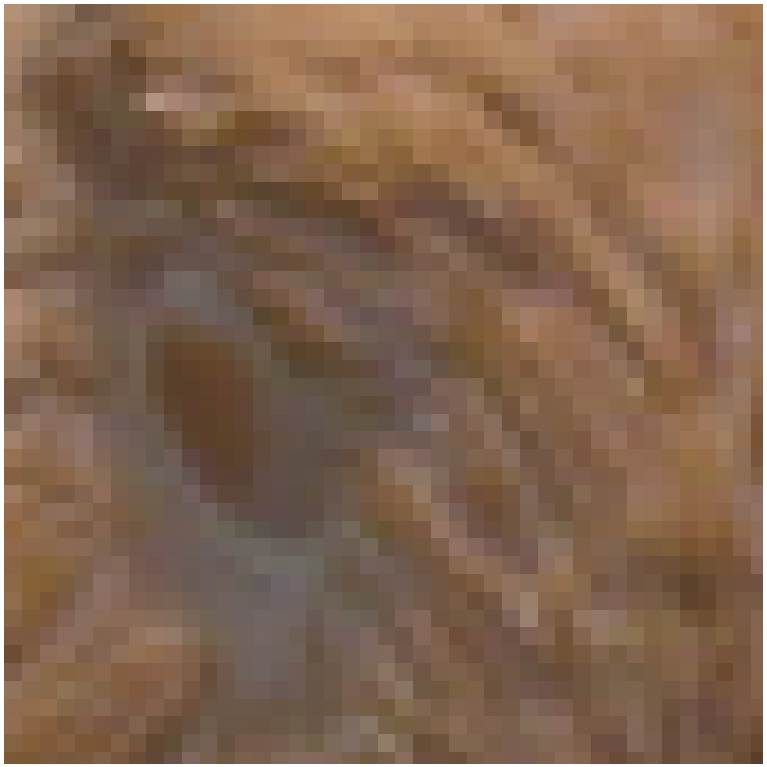
textured 3D model

Goldlücke, Cremers, ICCV 2009, DAGM 2009*

* Best Paper Award

Multi-view Texture Reconstruction

The super-resolution approach exploits the fact that every surface patch is observed in multiple images. It allows to invert the blurring and downsampling, providing a high-resolution texturing which is sharper than the individual input images:



input image close-up

super-resolution texture

Goldlücke, Cremers, ICCV 2009, DAGM 2009*

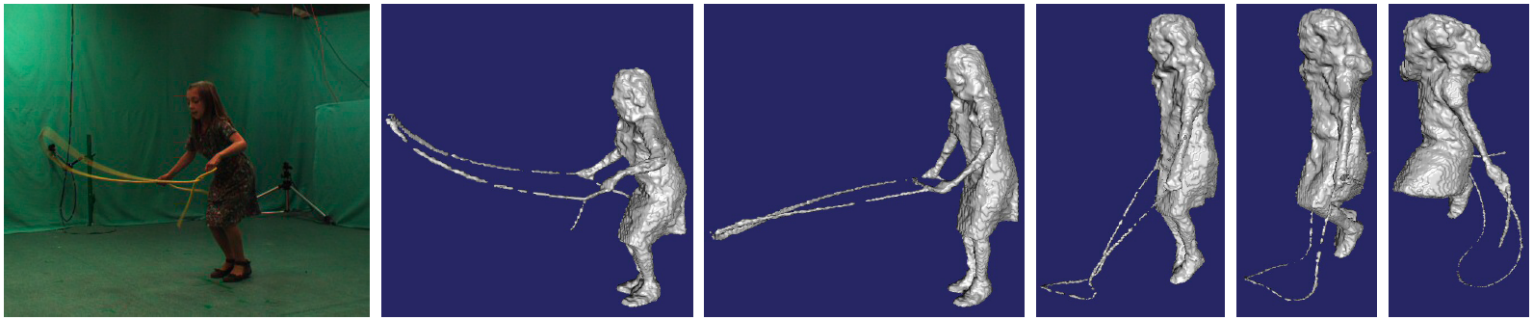
* Best Paper Award

Space-Time Reconstruction from Multi-view Video

Although laser-based reconstruction is often more accurate and more reliable (in the absence of texture), image-based reconstruction has two advantages:

- One can extract geometry and color of the objects.
- On can reconstruct actions over time filmed with multiple synchronized cameras.

Oswald & Cremers 4DMOD 2013 and Oswald, Stühmer, Cremers, ECCV 2014, propose convex variational approaches for dense space-time reconstruction from multi-view video.



1/16 input videos Dense reconstructions over time

Oswald, Stühmer, Cremers, ECCV 2014

Toward Free-Viewpoint Television

Space-time action reconstructions as done in Oswald & Cremers 2013 entail many fascinating applications, including:

- For video conferencing one can transmit a full 3D model of a speaker which gives stronger sense of presence and immersion.
- For sports analysis one can analyze the precise motion of a gymnast.
- For free viewpoint television, the spectator at home can interactively chose from which viewpoint to follow an action.

Textured action reconstruction for free-viewpoint television

Oswald, Cremers, 4DMOD 2013