Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

A Bayesian approach to simultaneously recover camera pose and non-rigid shape from monocular images[☆]

Francesc Moreno-Noguer[*], Josep M. Porta

Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Llorens Artigas 4-6, 08028 Barcelona, Spain

ARTICLE INFO

Article history:
Received 27 January 2014
Accepted 2 May 2016
Available online 8 June 2016

Keywords:
Deformable surfaces
Pose estimation
Bayesian belief networks
SLAM

ABSTRACT

In this paper we bring the tools of the Simultaneous Localization and Map Building (SLAM) problem from a rigid to a deformable domain and use them to simultaneously recover the 3D shape of non-rigid surfaces and the sequence of poses of a moving camera. Under the assumption that the surface shape may be represented as a weighted sum of deformation modes, we show that the problem of estimating the modal weights along with the camera poses, can be probabilistically formulated as a maximum a posteriori estimate and solved using an iterative least squares optimization. In addition, the probabilistic formulation we propose is very general and allows introducing different constraints without requiring any extra complexity. As a proof of concept, we show that local inextensibility constraints that prevent the surface from stretching can be easily integrated.

An extensive evaluation on synthetic and real data, demonstrates that our method has several advantages over current non-rigid shape from motion approaches. In particular, we show that our solution is robust to large amounts of noise and outliers and that it does not need to track points over the whole sequence nor to use an initialization close from the ground truth.

1. Introduction

Recovering the 3D shape of non-rigid objects from monocular image sequences is known to be a severely ill-conditioned problem because very different shape configurations may have a similar projection [14, 35, 38]. As shown in Fig. 1 the problem becomes even further underconstrained if the camera moves while the shape deforms, and both non-rigid shape and camera motion have to be simultaneously estimated. In order to resolve the inherent ambiguity between camera motion and shape deformation and turn the problem into a tractable one, prior knowledge about the object's behavior or the camera dynamics is then necessary.

Traditional approaches seek to reduce the space of possible shapes by introducing deformation models, either physically inspired ones [9, 23, 24, 43, 46] or learned from training data [6, 7, 8, 10, 17, 22, 25, 28, 30, 38, 39, 50]. Surface deformations are then expressed as weighted combinations of modes, and estimating the shape entails at retrieving the weights of this linear combination by minimizing image based objective functions. However, since these objective functions are often complex, their convergence is only guaranteed if the shape is precisely initialized. In addition, most of these approaches either assume the pose of the camera to be known or retrieve the shape with no camera referential.

Recent approaches in non-rigid structure-from-motion (NRSFM) have shown that deformation modes can be learned along with the shape and motion parameters [3, 15, 33, 36, 42, 44, 47, 49]. Yet, while these techniques are especially interesting in situations where training data is hard to obtain, they typically require a number of points to be tracked throughout the whole sequence, which is difficult to satisfy in practice, especially when dealing with non-rigid objects that suffer from self occlusions. Furthermore, existing NRSFM approaches have shown to be effective only for relatively small deformations and they are quite sensitive to the presence of outliers and noisy observations.

In this paper, we propose a new formulation to the problem of simultaneously retrieving non-rigid 3D shape and camera motion that overcomes some of the limitations of previous approaches. We make two basic assumptions that are widely used in previous literature [29, 35, 37, 38]. First, we assume that the deformation modes are available. And second, we assume that some 2D-to-3D correspondences can be established between the input images and a reference image in which the shape is already known. Yet, in contrast to NRSFM

[☆] This paper has been recommended for acceptance by Jan-Michael Frahm, Dr.-Ing.

[*] Corresponding author.

E-mail addresses: fmoreno@iri.upc.edu (F. Moreno-Noguer), porta@iri.upc.edu (J. Porta).

URLs: http://iri.upc.edu/people/fmoreno (F. Moreno-Noguer), http://iri.upc.edu/people/porta (J. Porta).

http://dx.doi.org/10.1016/j.imavis.2016.05.012
0262-8856/© 2016 Elsevier B.V. All rights reserved.

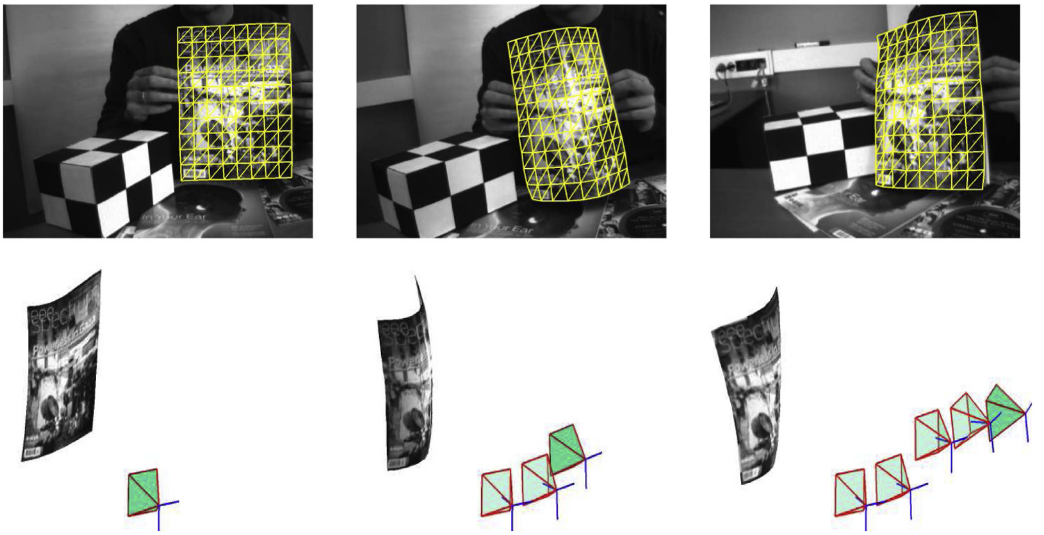

Fig. 1. Simultaneous estimation of non-rigid shape and camera pose from input images. Top: Three different frames of an input sequence with the reconstructed 3D mesh overlaid. Bottom: Re-textured side view of the retrieved surface and sample camera poses up to the current frame. Note that estimating the camera pose from only the observation of the deforming shape is very difficult even for the human eye. It would be much easier from the observation of the rigid objects, such as the calibration box, although we do not contemplate this case in the current paper.

methods, we do not require tracking the points along the whole sequence, that is, each input image may have its independent set of matches. And most importantly, our method tolerates significant amounts of outliers and noise.

Our approach draws inspiration from a recent work on Simultaneous Localization and Map Building (SLAM) [13] used to estimate camera pose while mapping a rigid and static environment. We show that by appropriately parameterizing the shape and pose, the SLAM formulation can be extended to non-rigid domains. More specifically, we formulate the problem of estimating the modal weights describing non-rigid shapes and the pose parameters of the camera as a Maximum a Posteriori (MAP) estimate which can be iteratively solved using linearization and an efficient QR factorization for sparse linear systems [12]. As we will demonstrate through testing on both synthetic and real data, besides the robustness to outliers and noise, this formulation does not require a precise initialization, which is another remarkable step-forward when compared to the previous methodologies.

This work is an extended version of our earlier paper [27] where we already proposed the probabilistic framework to integrate parameters describing both the camera motion and the surface deformation. Here, we exploit the generality of this methodology and show that it allows introducing additional constraints. As a proof of concept we will show that enforcing local inextensibility naturally fits within our formulation, yielding better accuracies on the reconstructed shapes.

2. Related work

3D surface reconstruction from monocular images has been an active research topic in Computer Vision for many years. Existing solutions may be roughly classified into those based on pre-defined or pre-learned deformation modes and those that learn the modes from input images and simultaneously retrieve shape and pose parameters.

The earliest works introduced physically-inspired deformation modes such as superquadrics [24], thin-plate splines [23] or balloons [9], used in combination with modal analysis [34] to reduce the degrees of freedom of the problem. Yet, all these approaches are only effective to capture relatively small deformations. More realistic deformations were described by complex non-linear models [4, 46], although their applicability is limited to very specific materials.

This limitation has been addressed by methods that learn the deformation modes from training data, such as the Active Appearance and Shape Models [10, 22] or the 3D Morphable Models [6]. These approaches represent surface deformations as linear combinations of rigid modes, and retrieving shape entails minimizing an image-based objective function. However, since this function is typically highly non-convex, it requires good pose and shape initializations to converge, which makes these methods appropriate for tracking shapes with a small inter-frame deformation, such as faces [30, 50]. In [17] a similar approach is used to detect human shape and pose from just a single image, although it requires manual pose initialization.

Several recent methods have been proposed to recover non-rigid shape from single images, by using deformation modes in conjunction with local rigidity constraints to reconstruct inextensible surfaces [14, 28, 35, 37, 38], and in conjunction with shading constraints to reconstruct stretchable surfaces [29]. However, none of these approaches retrieves the camera pose, and either assume that the deformation modes are aligned with the camera coordinate system or provide a solution shape for which the pose is unknown. One interesting exception is [39] which simultaneously retrieves point correspondences, pose and shape from one single image. Yet, in order to do so, it assumes hard prior constraints on the pose which may be difficult to hold in practice, and can only handle a reduced amount of points of interest.

Constraining the surface motion by linear models is also at the core of non-rigid structure-from-motion methods. Although the seminal work of Bregler et al. [8] used known deformation modes, current approaches [3, 15, 33, 36, 42, 44, 47, 49] do not require to know them and, given a video sequence, they simultaneously compute modes, pose and shape. This generality, though, comes at the price of having to impose several constraints that are difficult to hold in practice, such as requiring a sufficient number of points to be tracked throughout the whole sequence. In addition, most of these methods have only been effectively used to retrieve relatively small deformations, and tend to be sensitive to noisy correspondences, missing data, and outliers. Recently, in [1, 2], the strengths of both the NRSFM and the physic-based approaches based on Finite Elements are merged, yielding a system able to track the motion of