Computer Vision and Image Understanding 136 (2015) 32–44

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

The informed sampler: A discriminative approach to Bayesian inference in generative computer vision models

Varun Jampani [a,*], Sebastian Nowozin [b], Matthew Loper [a], Peter V. Gehler [a]

[a] Max Planck Institute for Intelligent Systems, Spemannstraße 41, Tübingen, Germany
[b] Microsoft Research Cambridge, 21 Station Road, Cambridge, United Kingdom

ARTICLE INFO

Article history:
Received 15 April 2014
Accepted 4 March 2015

Keywords:
Probabilistic models
MCMC inference
Inverse Graphics
Generative models

ABSTRACT

Computer vision is hard because of a large variability in lighting, shape, and texture; in addition the image signal is non-additive due to occlusion. Generative models promised to account for this variability by accurately modelling the image formation process as a function of latent variables with prior beliefs. Bayesian posterior inference could then, in principle, explain the observation. While intuitively appealing, generative models for computer vision have largely failed to deliver on that promise due to the difficulty of posterior inference. As a result the community has favoured efficient discriminative approaches. We still believe in the usefulness of generative models in computer vision, but argue that we need to leverage existing discriminative or even heuristic computer vision methods. We implement this idea in a principled way with an *informed sampler* and in careful experiments demonstrate it on challenging generative models which contain renderer programs as their components. We concentrate on the problem of inverting an existing graphics rendering engine, an approach that can be understood as "Inverse Graphics". The informed sampler, using simple discriminative proposals based on existing computer vision technology, achieves significant improvements of inference.

1. Introduction

A conceptually elegant view on computer vision is to consider a generative model of the physical image formation process. The observed image becomes a function of unobserved variables of interest (for example presence and positions of objects) and nuisance variables (for example light sources, shadows). When building such a generative model, we can think of a scene description θ that produces an image $I = G(\theta)$ using a deterministic rendering engine G, or more generally, results in a distribution over images, $p(I|\theta)$. Given an image observation $\widehat{I}$ and a prior over scenes $p(\theta)$ we can then perform Bayesian inference to obtain updated beliefs $p(\theta|\widehat{I})$. This view was advocated since the late 1970s [24,22,45,33,31,44].

Now, 30 years later, we would argue that the generative approach has largely failed to deliver on its promise. The few successes of the idea have been in limited settings. In the successful examples, either the generative model was restricted to few high-level latent variables, e.g. [36], or restricted to a set of image transformations in a fixed reference frame, e.g. [6], or it modelled only a limited aspect such as object shape masks [16], or, in the worst case, the generative model was merely used to generate training data for a discriminative model [39]. With all its intuitive appeal, its beauty and simplicity, it is fair to say that the track record of generative models in computer vision is poor. As a result, the field of computer vision is now dominated by efficient but data-hungry discriminative models, the use of empirical risk minimization for learning, and energy minimization on heuristic objective functions for inference.

Why did generative models not succeed? There are two key problems that need to be addressed, the design of an accurate generative model, and the inference therein. Modern computer graphic systems that leverage dedicated hardware setups produce a stunning level of realism with high frame rates. We believe that these systems will find its way in the design of generative models and will open up exciting modelling opportunities. This observation motivates the research question of this paper, the design of a general inference technique for efficient posterior inference in accurate computer graphics systems. As such it can be understood as an instance of *Inverse Graphics* [5], illustrated in Fig. 1 with one of our applications.

The key problem in the generative world view is the difficulty of posterior inference at test-time. This difficulty stems from a number of reasons: *first*, the parameter θ is typically high-dimensional and so is the posterior. *Second*, given θ, the image formation process

* Corresponding author.
E-mail address: varun.jampani@tue.mpg.de (V. Jampani).

http://dx.doi.org/10.1016/j.cviu.2015.03.002
1077-3142/

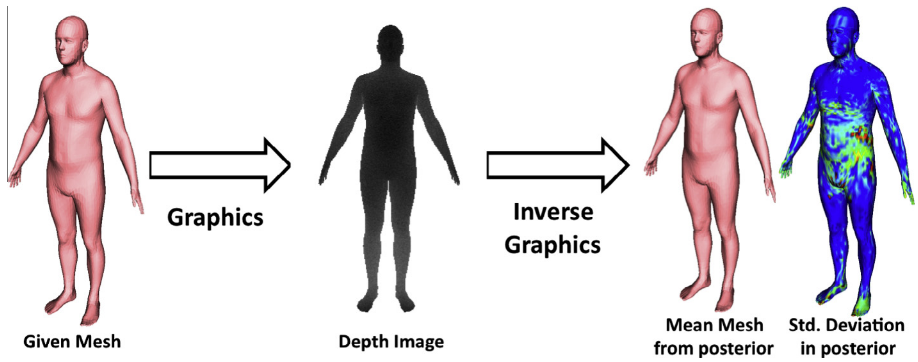

Fig. 1. An example "inverse graphics" problem. A graphics engine renders a 3D body mesh and a depth image using an artificial camera. By Inverse Graphics we refer to the process of estimating the posterior probability over possible bodies given the depth image.

realizes complex and *dynamic* dependency structures, for example when objects occlude or self-occlude each other. These intrinsic ambiguities result in multi-modal posterior distributions. *Third*, while most renderers are real-time, each simulation of the forward process is expensive and prevents exhaustive enumeration.

We believe in the usefulness of generative models for computer vision tasks, but argue that in order to overcome the substantial inference challenges we have to devise techniques that are general and allow reuse in several different models and novel scenarios. On the other hand we want to maintain correctness in terms of the probabilistic estimates that they produce. One way to improve on inference efficiency is to leverage existing computer vision features and discriminative models in order to aid inference in the generative model. In this paper, we propose the *informed sampler*, a Markov Chain Monte Carlo (MCMC) method with discriminative proposal distributions. It can be understood as an instance of a data-driven MCMC method [46], and our aim is to design a method that is general enough such that it can be applied across different problems and is not tailored to a particular application.

During sampling, the informed sampler leverages computer vision features and algorithms to make informed proposals for the state of latent variables and these proposals are accepted or rejected based on the generative model. The informed sampler is simple and easy to implement, but it enables inference in generative models that were out of reach for current *uninformed* samplers. We demonstrate this claim on challenging models that incorporate rendering engines, object occlusion, ill-posedness, and multi-modality. We carefully assess convergence statistics for the samplers to investigate their truthfulness about the probabilistic estimates. In our experiments we use existing computer vision technology: our informed sampler uses standard histogram-of-gradients features (HoG) [12], and the OpenCV library, [7], to produce informed proposals. Likewise one of our models is an existing computer vision model, the *BlendSCAPE* model, a parametric model of human bodies [23].

In Section 2, we discuss related work and explain our informed sampler approach in Section 3. Section 4 presents baseline methods and experimental setup. Then we present experimental analysis of informed sampler with three diverse problems of estimating camera extrinsics (Section 5), occlusion reasoning (Section 6) and estimating body shape (Section 7). We conclude with a discussion of future work in Section 8.

2. Related work

This work stands at the intersection of computer vision, computer graphics, and machine learning; it builds on previous approaches we will discuss below.

There is a vast literature on approaches to solve computer vision applications by means of generative models. We mention some works that also use an accurate graphics process as generative model. This includes applications such as indoor scene understanding [15], human pose estimation [29], and hand pose estimation [14]. Most of these works are however interested in inferring MAP solutions, rather than the full posterior distribution.

Our method is similar in spirit to *Data Driven Markov Chain Monte Carlo* (DDMCMC) methods that use a bottom-up approach to help convergence of MCMC sampling. DDMCMC methods have been used in image segmentation [43], object recognition [46], and human pose estimation [29]. The idea of making Markov samplers data dependent is very general, but in the works mentioned above, lead to highly problem specific implementations, mostly using approximate likelihood functions. It is due to specialization on a problem domain, that the proposed samplers are not easily transferable to new problems. This is what we focus on in our work: to provide a simple, yet efficient and general inference technique for problems where an accurate forward process exists. Because our method is general we believe that it is easy to adapt to a variety of new models and tasks.

The idea to invert graphics [5] in order to understand scenes also has roots in the computer graphics community under the term "inverse rendering". The goal of inverse rendering however is to derive a direct mathematical model for the forward light transport process and then to analytically invert it. The work of [37] falls in this category. The authors formulate the light reflection problem as a convolution, to then understand the inverse light transport problem as a deconvolution. While this is a very elegant way to pose the problem, it does require a specification of the inverse process, a requirement generative modelling approaches try to circumvent.

Our approach can also be viewed as an instance of a probabilistic programming approach. In the recent work of [31], the authors combine graphics modules in a probabilistic programming language to formulate an approximate Bayesian computation. Inference is then implemented using Metropolis–Hastings (MH) sampling. This approach is appealing in its generality and elegance, however we show that for our graphics problems, a plain MH sampling approach is not sufficient to achieve reliable inference and that our proposed informed sampler can achieve robust convergence in these challenging models. Another piece of work from [41] is similar to our proposed inference method in that knowledge about the forward process is learned as "stochastic inverses", then applied for MCMC sampling in a Bayesian network. In the present work, we devise an MCMC sampler that we show works in both a multi-modal problem as well as for inverting an existing piece of image rendering code. In summary, our method can be understood in a similar context as the above-mentioned papers, including [31].