Computer Vision and Image Understanding 113 (2009) 29–47

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Fast nonparametric belief propagation for real-time stereo articulated body tracking

Olivier Bernier *, Pascal Cheung-Mon-Chan, Arnaud Bouguet

France Telecom R&D, Technopole Anticipa, Orange Labs ft/nsm/rd/tech/iris/via, 2 Av. Pierre Marzin, 22307 Lannion Cedex, France

ARTICLE INFO

Article history:
Received 27 November 2007
Accepted 1 July 2008
Available online 15 July 2008

Keywords:
Articulated tracking
Graphical model
Belief propagation
Body pose
Factor graph
Particle filter
Stereo
3D contours
Real-time
Pose estimation

ABSTRACT

This article proposes a statistical approach for fast articulated 3D body tracking, similar to the loose-limbed model, but using the factor graph representation and a fast estimation algorithm. A fast Nonparametric Belief Propagation on factor graphs is used to estimate the current marginal for each limb. All belief propagation messages are represented as sums of weighted samples. The resulting algorithm corresponds to a set of particle filters, one for each limb, where an extra step recomputes the weight of each sample by taking into account the links between limbs. Applied to upper body tracking with stereo and colour images, the resulting algorithm estimates the body pose in quasi real-time (10 Hz). Results on sequences illustrate the effectiveness of this approach.

1. Introduction and motivation

Vision-based articulated body pose estimation and tracking, either for monocular, stereo or multiple camera sequences, is a challenging problem, especially if a real-time algorithm is needed (see [28] for a recent general survey of vision based human motion capture). Proposed methods to address this problem can be roughly classified between deterministic methods [5,9,13,22,30] and statistical ones [10,11,15,25,39,52]. Deterministic methods are generally based on the minimisation of a function and can lose track for fast motions or occlusions because of local minima [10]. The use of multiple cameras around the subject [21,22,29] can somewhat alleviate this problem by providing occlusion free viewpoints. On the other hand statistical methods tend to be more robust but also more computationally expensive. They generally take the form of an estimation of the current probability density of the body parameters using a Bayesian approach.

The main difficulty facing both deterministic and statistical methods is the high dimension of the body parameters space, especially when the 3D pose is to be recovered, as opposed to the 2D projected pose on the image plane of a camera. For the statistical methods in particular, the random exploration of this state space is prohibitively expensive, as the number of needed samples grows exponentially with this dimension. One possibility to avoid this difficulty is to restrict the pose and movement state to learned specific cases [3,6,45,48,47,46]. For unrestricted tracking algorithms, a more intelligent state exploration, using for example covariance scaled sampling [39] or taking into account the intrinsic ambiguity of tracking with a monocular camera [40], is needed. Another approach is to use the image itself to generate the samples, for example using a learning method. Parameter sensitive hashing [34] (PSH) is a learning based method which can be used to sample the modes of the image likelihood [10]. Coupled with conditional random fields and grid filters, the PSH method can be used for real-time articulated body tracking [42]. Regression can also be applied to learn directly, from a set of scale invariant feature transforms (SIFT) features extracted from the image, the pose of the body [2].

One limitation of learning algorithms is the representativity of the learning bases. The variability of body appearance due to lighting changes, clothes and limbs deformation, complex backgrounds, as well as the variability of 3D postures, cannot be covered by any learning base, even generated synthetically. The number of needed samples is indeed astronomically high. Prior knowledge is needed to reduce the complexity of the learning task. The choice of the features to extract from the image can be viewed as an example of a prior, which can reduce the difficulty of the problem. Such features hopefully are discriminant enough to enable pose estimation and robust enough to hide a part of the variability above. Not all variability can be eliminated, however, and the resulting search space is still very large. Another example of such a prior is the known

* Corresponding author.
E-mail address: olivier.bernier@orange-ftgroup.com (O. Bernier).

1077-3142/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2008.07.001

articulated structure of the body. Decomposing the state space into the product of more simple spaces using this articulated structure is a natural way to avoid the high dimension of the parameter state. Indeed, this was the first approach for 2D articulated body tracking using deterministic algorithms (see for example [20]). This approach can also be applied to stochastic methods, by representing the articulated structure as a graphical model [32]. The estimation of the current pose is then equivalent to a Bayesian estimation on this graphical model. A number of methods can be applied for this estimation, with the added complication that as the state space is continuous, a density approximation method must be used. For example, in [15], the authors discretise the state space and use the loopy belief propagation algorithm (LBP) [50,53] for estimation of the body pose in 2D for specific views. The 3D body pose is recovered using using a standard algorithm [43]. In [36], the authors use nonparametric belief propagation, a nonparametric version of the loopy belief propagation algorithm [18] based on the Monte Carlo (sample set) integration method and representing messages as sums of Gaussians to estimate a 3D pose, using Gibbs sampling to estimate the product of such sums. A faster method based on mode propagation and kernel fitting is proposed in [16], but only for 2D pose estimation. The nonparametric belief propagation approach was recently extended to take into account occlusions for monocular tracking of articulated objects [41,37]. Belief propagation is not the only method for inference and estimation on a graphical model, and a mean field approximation coupled with a sample set representation can also be used for 2D articulated body tracking [17,52].

For computer vision applications such as human–computer interfaces (HCI), real-time tracking is a necessity. All these previous statistical methods, in particular for estimating the 3D pose, are computationally very expensive and cannot be used for such real-time applications. Learning based methods can achieve real-time [42], but with the inherent limitations explained above. In this paper, we propose a new fast nonparametric belief propagation approach to track the articulated 3D body pose in real-time. We choose a model similar to the one used in [36], but with a new efficient recursive estimation method based on belief propagation on factor graphs [23]. Our method is similar in its approach to the one proposed by [52], but performs in quasi real-time for 3D upper body tracking. Compared to the nonparametric belief propagation approach of [41,37], our method does not take into account the possible occultations between limbs, but is far faster. The key point of this method is the observation that the most computationally intensive step of nonparametric belief propagation applied to articulated body tracking is the evaluation of the likelihood of a sample of a limb on the image. Our method aims to reduce the computational complexity by fixing the samples used to represent each limb for one frame. All belief propagation messages are consequently represented as a weighed sum of these samples. Each sample is evaluated on the image only once for each frame instead of multiple times as in other belief propagation approaches, thus considerably reducing the computational cost. A fully recursive estimation, equivalent to particle filters interacting through belief propagation, is obtained [4].

The proposed algorithm is applied to tracking the upper body pose using a stereo camera, in real-time. Instead of 3D points or voxels we advocate the use of 3D contours as robust features. These 3D contours are extracted using a disparity estimation in the vicinity of contour points. The disparity is evaluated independently on each side of a contour point to obtain a valid estimation even for occluding contours. The resulting features are fast to compute as the disparity is evaluated only on a small subset of all image points. These 3D contours are augmented with colour histograms for hands and head tracking, as head and hands colour is similar and generally distinctive from the rest of the image. Our method is focused on tracking knowing the initial pose. Detection of the initial pose is automatic, obtained by detecting the face and supposing a natural starting pose of the body (arms roughly along the torso). The paper is organized as follows. In Section 2 we will present the general graphical model and theoretically derive our estimation algorithm, based on belief propagation. Section 3 will present the specific model used for tracking the upper body in real-time using a stereo camera, including the images features and corresponding image compatibility functions. Results will be presented in Section 4, before a general conclusion.

2. Recursive Bayesian tracking for articulated objects

A statistical method for articulated objects tracking comprises two parts: the statistical model used to represent the articulated structure, and the estimation algorithm for this model. Classically, we use a graphical model to represent the articulated body structure, and the classical Belief Propagation algorithm as the basis for our fast estimation algorithm.

2.1. Graphical model

The most simple possibility to represent the articulated structure as a graphical model is to use a Bayesian tree structure for the graph, where nodes represent body parts (limbs) and edges represent links between parts (joints) [32] (see Fig. 1a for a simple case with three parts). The likelihood of such a structure can be efficiently estimated for a restricted class of links [12]. However our goal is also to take into account the time coherence of the body pose between consecutive discrete sampling times. A solution is to add edges representing the dependence between the state of the same part between consecutive times. The resulting graph is a Bayesian network with loops, and the tree representation cannot be used (see Fig. 1b). An equivalent approach is to use a Markov random field (or MRF) with pairwise interactions [15–17,36,37] (see Fig. 1c). In the general case, links can create cliques of higher order in the MRF, for example three-node cliques (see Fig. 1d). This can imply a more complex and computationally intensive estimation algorithm as three-node (or more) cliques are equivalent to factors of three-node states (or more) in the joint probability in the general case. To avoid this problem, and to specify the use of only pairwise factors in the global state probability, we represent the model by a factor graph [23]. A factor graph directly represents the decomposition of the global state probability (joint probability of all part states) as a product of positive factors (of individual part states). For each MRF, an equivalent factor graph can be constructed. Fig. 1e shows the factor graph equivalent to the MRF in Fig. 1d. In this representation, rectangles represent factor nodes and circles represent variable nodes (part states), and the factors of three-node states are clearly visible. Instead of this general model, we decompose all higher order factors in products of factors of only two states, to obtain a factor graph with only pairwise factors (see Fig. 1f). Note that this model is more specific than a Markov random field: the constraint of using only pairwise factors cannot be represented with a Markov random field if higher order cliques are present.

In our case the pairwise factors taken into account are: the pairwise factors between the states of two parts at the same time representing links between parts (which we call the *link factors*), the pairwise factors between the states of each part at two consecutive times (which we call the *time coherence factors*), and the pairwise factors between all parts and their corresponding observations (which we call the *image compatibility factors*). For vision based tracking, the observation linked to one part correspond to features extracted from a region of one or more images taken at the same