Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 47 (2018) 68–84

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Unsupervised construction of human body models

Action editor: Alessandra Sciutti

Thomas Walther, Rolf P. Würtz *

Department of Electrical Engineering and Information Technology and Institute for Neural Computation, Ruhr-University Bochum, Germany

Received 13 March 2017; received in revised form 27 June 2017; accepted 8 August 2017
Available online 24 August 2017

Abstract

Unsupervised learning of a generalizable model of the visual appearance of humans from video data is of major importance for computing systems interacting naturally with their users and others. We propose a step towards automatic behavior understanding by making the posture estimation cycle more autonomous. The system extracts coherent motion from moving upper bodies and autonomously decides about limbs and their possible spatial relationships. The models from many videos are integrated into a meta-model, which shows good generalization with respect to different individuals, backgrounds, and attire. This model allows robust interpretation of single video frames without temporal continuity and posture mimicking by an android robot.

Keywords: Structure learning; Learning a visual representation; Upper body pose estimation

1. Introduction

Humans show unmatched expertise in visually analyzing and interpreting the movements of other humans. This skill of social perception is one of the foundations of effective and smooth interaction of humans inhabiting a complex environment. The benefits of machines capable of interpreting human motion would be enormous: applications in health care, surveillance, industry and sports (Gavrila, 1999; Moeslund, Hilton, & Krüger, 2006) promise a broad market. Despite significant effort (Poppe, 2007) to transfer human abilities in motion estimation and behavioral interpretation to synthetic systems, automatically *looking at people* (Gavrila, 1999) remains among the 'most difficult recognition problem[s] in computer vision' (Mori, Ren, Efros, & Malik, 2004) there is still no technical solution matching human competency in vision-based motion capturing (VBMC). Furthermore, humans can understand body poses even in still images.

Artificial vision systems must be enhanced by learning lessons from human perception. Here, we present a system that is able to acquire conceptual models of the upper human body in a completely autonomous manner: the learning procedures are based on only a few general principles, namely the gestalt rule of "common fate", which states that coherently image parts with coherent motion belong to a single object, and the rule that object properties persistent over time are important for recognizing the object, while malleable ones should be ignored. This strategy significantly reduces human workload and allows self-optimization of the generated models. While autonomous model learning and knowledge agglomeration take place in simple scenarios, the conceptual nature of the retrieved body representations allows for generalization to more complex scenarios and holds opportunities for model *adaptation and enhancement loops*, which might perform continuous, non-trivial learning as found in the human brain. A much simpler example of such a system has been presented

* Corresponding author.
E-mail addresses: thomas.walther@rub.de (T. Walther), rolf.wuertz@ini.rub.de (R.P. Würtz).

http://dx.doi.org/10.1016/j.cogsys.2017.08.001
1389-0417/© 2017 Elsevier B.V. All rights reserved.

by Prodöhl, Würtz, and von der Malsburg (2003), where a neural network learns the gestalt rule of collinearity from common fate.

Fig. 1 provides a schematic overview of the system and is referred to throughout the paper for all components. In Section 2 we give an overview of VBMC approaches that have been considered or used in this work and discuss their strengths and weaknesses. Section 3 describes the details of learning a body model from a single video of human motion. This consists of the following subsystems:

- A central requirement for autonomous model learning is the exclusion of irrelevant features. This is achieved by motion-based background elimination (Section 3.1, Fig. 1(c)).
- The "common fate" rule is implemented by measuring and clustering point trajectories to select coherently moving parts, called limb patterns, and constraints on relative motion (Section 3.1, Fig. 1(d)).
- For a matchable description of limbs we extract skeletons from those limb patterns (Section 3.2, Fig. 1(e)).
- The next step is the generation of limb templates to be filled with color (Fig. 1j), shape (Fig. 1i), and texture (Fig. 1k) (Section 3.3).
- Single limbs are combined into a complete body model, which describes the encountered relative movements and their constraints as well as the appearance of each limb template to a *pictorial structure* (Section 3.4, Fig. 1(e)).

Each of these subsystems is constructed by using relevant techniques from the literature described in Section 2, and we describe all modifications that were necessary for autonomous learning.

A general model must include more than a single video in order to capture possible variations in appearance and movements. Therefore, in Section 4 many such models are combined into a meta-model, which captures the invariant cues of the single models. In Section 5 we test the learned meta-model on still images with different backgrounds, individuals, attire, etc. This is a much harder task than evaluating more videos of a single person, and the failures point to ways to improve the system by adding more training. We provide test results on single images varying considerably in person, attire, and background. Then we show how the learned representations can be used to mimick observed postures on a humanoid robot. The paper ends with a brief discussion.

2. Previous work in vision-based human motion capturing

Following Poppe (2007), VBMC methods can be classified into *model-based*, *generative* approaches and *model-free*, *discriminative* methods (cf. also (Navaratnam, Fitzgibbon, & Cipolla, 2006)). Model-based schemes incorporate *top-down* and *bottom-up* techniques, while the model-free domain employs *learning-based* and *exemplar-based* pose estimation.

To stay in scope, we leave an in-depth discussion of top-down and discriminative techniques to Poppe (2007) or Walther (2011). Bottom-up solutions form an important mainstay of our own approach and are thus investigated more closely. Nevertheless, our focus is on autonomous, fully unsupervised VBMC strategies.

2.1. Bottom-up posture estimation

A generic bottom-up (or *combinatorial* (Roberts, McKenna, & Ricketts, 2007)) posture estimation system follows the principle formulated by Sigal and Black (2006a): 'measure locally, reason globally.' Local measurement treats the human body as an ensemble of 'quasi-independent' (Sigal, Isard, Sigelman, & Black, 2003) limbs, which much alleviates the complex model coupling inherent in top-down approaches. Imposing independence, 'image measurements' (Sigal et al., 2003) of single limbs can be performed separately by a dedicated *limb detector* (LD) (Ramanan, Forsyth, & Zisserman, 2007; Sigal & Black, 2006b), which moves the burden of matching a given body part model to some well-chosen *image descriptors* (Kanaujia, Sminchisescu, & Metaxas, 2007; Poppe, 2007). The selection of appropriate images descriptors as well as construction and application of LDs require domain knowledge of and concept building by human supervisors. For many object categories, histograms of oriented gradient (HOG) features seem to be a good choice, allowing object classification by linear discriminant analysis (Hariharan, Malik, & Ramanan, 2012).

To organize the data from local measurements, pending inter-limb dependencies come into play during global reasoning. 'Assemblies' (Moeslund et al., 2006) of detector responses are retrieved that comply well with kinematically meaningful human body configurations. The majority of bottom-up systems employ *graphical models* (Sigal & Black, 2006a) (GMs) to encode human body assemblies: each node in the model's graph structure correlates to a dedicated body part, whereas the graph's edges encode (mostly) 'spring-like' (Lan & Huttenlocher, 2005; Sigal et al., 2003) kinematic relationships between single limbs.

Using GMs for global inference, a configuration becomes more 'human-like' (Felzenszwalb & Huttenlocher, 2000) if all LDs return low matching cost and the 'springs' between the body parts are close to their resting positions. This can conveniently be formulated by means of an *energy functional*, whose global minimum represents the most probable posture of the captured subject. However, minimization for arbitrary graphs and energy functions is NP-hard (Felzenszwalb & Huttenlocher, 2005). Thus, Felzenszwalb and Huttenlocher (2000) propose to restrict the graphs to be *tree-like* and further restrictions on the energy function to allow for computationally feasible posture inference using *dynamic programming* (Felzenszwalb & Huttenlocher, 2005). We follow this approach by boosting the *pictorial structure* (Fischler