Image and Vision Computing 32 (2014) 728–738

Contents lists available at ScienceDirect

Image and Vision Computing

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

Coupled person orientation estimation and appearance modeling using spherical harmonics ☆

Martijn C. Liem [a], Dariu M. Gavrila [a,b]

[a] Intelligent Autonomous Systems Group, Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
[b] Environment Perception Department, Daimler Research and Development, Wilhelm Runge St. 11, 89081 Ulm, Germany

ARTICLE INFO

Article history:
Received 1 June 2013
Received in revised form 29 December 2013
Accepted 2 April 2014
Available online 18 April 2014

Keywords:
Person appearance modeling
Orientation estimation
Spherical harmonics

ABSTRACT

We present a novel approach for the estimation of a person's overall body orientation, 3D shape and texture, from overlapping cameras. A distinguishing aspect of our approach is the use of spherical harmonics for 3D shape- and texture-representation; it offers a compact, low-dimensional representation, which elegantly copes with rotation estimation. The estimation process alternates between the estimation of texture, orientation and shape. Texture is estimated by sampling image intensities with the predicted 3D shape (i.e. torso and head) and the predicted orientation, from the last time step. Orientation (i.e. rotation around torso major axis) is estimated by minimizing the difference between a learned texture model in a canonical orientation and the current texture estimate. The newly estimated orientation allows to update the 3D shape estimate, taking into account the new 3D shape measurement obtained by volume carving.
We investigate various components of our approach in experiments on synthetic and real-world data. We show that our proposed method has lower orientation estimation error than other methods that use fixed 3D shape models, for data involving persons.

1. Introduction

A person's overall body orientation (i.e. rotation around 3D torso major axis, facing direction) conveys important information about the person's current activity and focus of attention. In this paper, we focus on the estimation of overall person body orientation from few, overlapping cameras. To obtain a more accurate orientation estimate, we jointly estimate it with a 3D shape and texture representation of a person's torso-head, under a rigidity assumption. By projection onto a basis of Spherical Harmonics (*SH*), a low dimensional appearance model is created that can cope with object deformations while retaining the spatial information captured in a textured 3D representation. By comparing the texture model of the person at consecutive points in time, the relative body orientation can be estimated elegantly by using the properties of the *SH*. This in turn allows to update the 3D shape and texture model of a person.

Apart from facilitating an accurate orientation estimate, the proposed 3D shape and texture model furthermore has the potential to offer improved track disambiguation in a multiple person scenario, compared to less descriptive 2D view-based models. The 3D representation could also provide an initialization for applications aiming at full articulated 3D pose recovery.

The remainder of this paper is organized as follows. In Section 2, we discuss the related work. Section 3 starts with an overview of the proposed approach and the basic properties of spherical harmonics. Thereafter, it describes the estimation of texture, orientation, and shape in alternating fashion. In Section 4, we present experimental results on both artificial and real-world datasets. We conclude and suggest directions for future work in Section 5.

2. Related work

Extensive research has been performed in the areas of person appearance modeling [1–5], 3D body shape modeling [6–9] and pose estimation [8–16]. This section focuses on the work that we consider to be most closely related to our paper.

Modeling person appearance is most commonly done based on single view information. Color histograms representing the full extent of a person's appearance are often used [1,17], while more sophisticated methods split a single person's appearance up into several layers, incorporating some spatial information in the descriptor [18,19]. More recent methods make use of ensembles of different features to be more robust and cover different aspects of persons' appearance like color and texture information [2,3].

Viewpoint invariance is addressed by Gray et al. [4], where AdaBoost is used to learn a good set of spatial and color features for representing persons. Some approaches combine histograms created at multiple viewpoints (e.g. Liem and Gavrila [18]). Others try estimating the full

☆ This paper has been recommended for acceptance by Lijun Yin.

E-mail addresses: M.C.Liem@uva.nl (M.C. Liem), D.M.Gavrila@uva.nl (D.M. Gavrila).

http://dx.doi.org/10.1016/j.imavis.2014.04.007
0262-8856/

body texture of a person by projecting the person's appearance onto a 3D structure like a cylinder (e.g. Gandhi and Trivedi [5]). These multi-view types of methods provide robustness to perception from different angles, while the use of full body textures also maintains the spatial properties of the appearance per view.

Since the human body shape is largely non-rigid, modeling body texture is not straightforward. As we will see in the experiments, mapping the texture of a non-rigid shape onto a rigid object as done by Gandhi and Trivedi [5] may result in instable textures over time, introducing artifacts into learned texture models and making it hard to accurately compare body textures over time. Using a more accurate estimate of the 3D object shape could help in creating more discriminative appearance models.

Some research has specifically aimed to generate a more accurate 3D reconstruction of the visual hull of objects, computed using shape-from-silhouette methods. Kutulakos and Seitz [20] compute an improved voxel model of the visual hull by coloring all voxels using the camera views and checking color consistency among cameras for each voxel, eliminating inconsistent voxels. Cheung et al. [21] present a method for aligning rigid objects in multiple time steps and reconstructing an object using information from multiple time instances. Translation and rotation are estimated by matching and aligning colored surface points over time. By transforming the camera viewpoints according to the estimated transformation, new virtual viewpoints are created extending the number of viewpoints usable for shape-from-silhouette methods.

Mitzel and Leibe [7] learn a 3D shape model of a person over time based on stereo depth data. Using the Iterative-Closest-Point (ICP) algorithm, the model is aligned with the stereo data at each time step enabling person tracking and updating the model.

An alternative approach is to estimate fully articulated 3D body pose, either by lifting 2D part-based body models [10,16] or by using 3D models directly [11,8,9]. This approach can in principle provide accurate 3D body shape and texture models, but is computationally expensive and is hard to apply in uncontrolled, complex environments (e.g. dynamic background, multiple persons).

Some work has been done on estimating the body facing direction of persons without estimating the full articulated pose. Several 2D single frame approaches combine orientation-specific person detectors [12–14], while work by Chen and Odobez [15] estimates body and head pose in batch mode, coupling the output of underlying classifiers. These methods offer an absolute orientation estimate with respect to some global reference frame. In Gandhi and Trivedi [5], texture sampled from a cylinder surrounding a person is shifted along the rotational axis in a generate-and-test fashion to find the best matching orientation.

SH have been used in order to perform face recognition under varying lighting conditions (e.g. Yue et al. [22]). Representing 3D objects by projecting them onto an *SH* basis has been researched mainly with respect to exemplar based 3D object retrieval. A collection of rotationally invariant 3D object descriptors has been compared by Bustos et al. [23]. Recently, an *SH* decomposition of a 3D extension of the HOG descriptor was presented by Liu et al. [24]. Makadia et al. [25] present a method for the registration of 3D point clouds that uses an *SH* representation of Extended Gaussian Images (EGI) to estimate the difference in orientation between point clouds. Several rotation invariant 3D surface shape descriptors have been compared by Huang et al. [6], for the purpose of finding matching poses in sequences with high quality 3D data.

3. Appearance modeling and orientation estimation

3.1. Overview

Fig. 1 gives a schematic overview of the proposed procedure for person appearance modeling and orientation estimation. Appearance is modeled as the combination of a 3D shape model and a texture model. We take an intermediate approach between modeling a person using a fixed body shape and doing full articulated pose estimation. We estimate the largest rigid partition (core) of the body over time by regarding all non-rigid elements of the human body (arms, legs) as 'noise' around the rigid body core (torso-head).

The preprocessing step uses multi-camera data to compute a volumetric 3D reconstruction of the scene. A person detection and tracking system localizes the person of interest, and the volume corresponding to this person is segmented from the volume space. In the next step, an *SH* based shape measurement is created by using the segmented volume (Section 3.3.2). Furthermore, an *SH* based texture measurement is computed by using the learned shape model and Kalman filtered orientation estimate from the previous time step *t-1* (Section 3.3.1). Using an EM-like optimization scheme, an optimized orientation measurement is computed by iteratively comparing the texture measurement to the learned texture model from *t-1*, and adjusting the shape model's orientation to compute an improved texture measurement (Section 3.3.3). When the orientation measurement has converged, it is used to correct the Kalman filtered orientation estimate from *t-1*, and the *SH* shape and texture models from *t-1* are updated making use of the shape and texture measurements and the estimated orientation.

Our main contribution is a method for estimating a person's relative body orientation while simultaneously generating a basic model of the person's shape and texture. The estimate is made on a per-frame basis using an on-line learned appearance model consisting of low dimensional *SH* shape and texture representations. By using *SH* as a basis, orientation estimation can be performed elegantly, without the need of explicitly testing for different orientations and without a constraint on

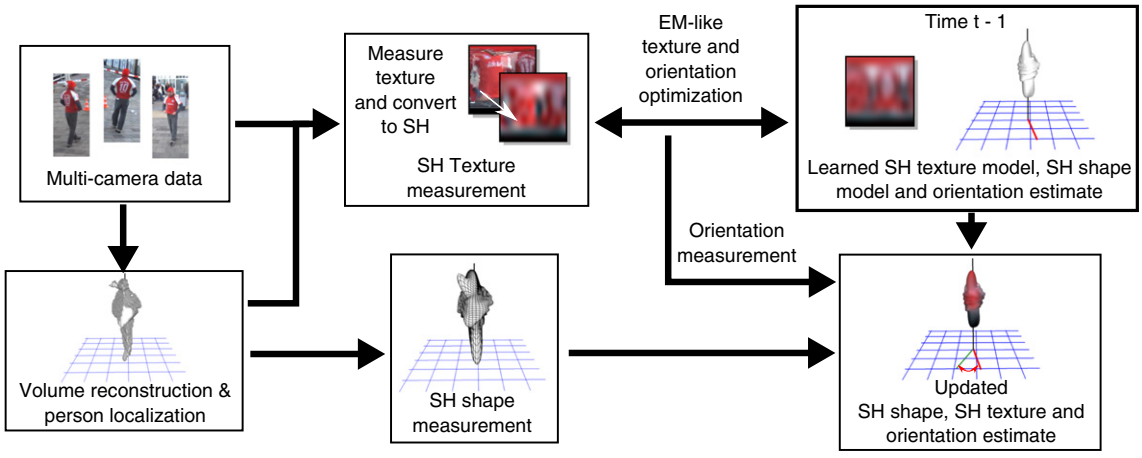

Fig. 1. Overview of the appearance modeling and orientation estimation process.