Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 2926–2944
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

A spatio-temporal 2D-models framework for human pose recovery in monocular sequences

Grégory Rogez*, Carlos Orrite-Uruñuela, Jesús Martínez-del-Rincón

CVLab, Aragon Institute for Engineering Research, University of Zaragoza, Spain

Received 6 July 2007; received in revised form 23 November 2007; accepted 25 February 2008

Abstract

This paper addresses the pose recovery problem of a particular articulated object: the human body. In this model-based approach, the 2D-shape is associated to the corresponding stick figure allowing the joint segmentation and pose recovery of the subject observed in the scene. The main disadvantage of 2D-models is their restriction to the viewpoint. To cope with this limitation, local spatio-temporal 2D-models corresponding to many views of the same sequences are trained, concatenated and sorted in a global framework. Temporal and spatial constraints are then considered to build the probabilistic transition matrix (PTM) that gives a frame to frame estimation of the most probable local models to use during the fitting procedure, thus limiting the feature space. This approach takes advantage of 3D information avoiding the use of a complex 3D human model. The experiments carried out on both indoor and outdoor sequences have demonstrated the ability of this approach to adequately segment pedestrians and estimate their poses independently of the direction of motion during the sequence.

Keywords: Human motion analysis; Human shape modelling; Pose inference

1. Introduction

Human motion capture and analysis has grown to become one of the most active research topics in computer vision over the past decade [1]. This is mainly motivated by the wide spectrum of promising applications in many fields such as video-surveillance, human–machine interfaces, medical diagnosis, sports performance analysis or biometrics.

The human motion analysis divides into three main interacting levels as described in Ref. [2]: human detection, human tracking and human behavior understanding. The detection stage that aims at segmenting people from the rest of the image is a significant issue since the performance of the other two processes highly depends on it. Human activity understanding relies on accurate detection and tracking, but a good prior knowledge of pose can also improve considerably both detection and tracking.

Many efficient systems are based on the use of a model which is, most of the time, a representation of the human body. In previous works, the structure and appearance of the human body have been represented as 2D or 3D stick figure [3], 2D (active) contour or shape [4–6], binary silhouette [7] or 3D volumetric model [8,9]. The selection of the appropriate model is a critical issue and the use of an explicit body model is not simple, given the high number of degrees of freedom of the human body and the self-occlusions inherent to the monocular observation.

People are able to deduce the pose of a known articulated object (e.g. a person) from a simple binary silhouette. The possible ambiguities can be solved from dynamics when the object is moving. Following this statement, the first step of this work consists in constructing a human model that encapsulates within a point distribution model (PDM) [10] both body silhouette information provided by the 2D-shape and structural information given by the 2D skeleton joints. In that way, the 2D pose could be inferred from the silhouette and vice versa. Due to the high non-linearity of the resulting feature space, mainly caused by the rotational deformations inherent to the articulated structure of the human body, the use of non-linear statistical models will be considered in this work. This approach will be compared to

* Corresponding author. Tel.: +34 635 983 614.
E-mail addresses: grogez@unizar.es (G. Rogez), corrite@unizar.es (C. Orrite-Uruñuela).

0031-3203/$30.00 © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.02.012

other two methods previously tested for solving non-linearity issue. Such non-linear statistical models have been previously proposed by Bowden [11] that demonstrated how the 3D structure of an object can be reconstructed from a single view of its outline. While Bowden only considered the upper human body and the frontal view, in this work the complete body will be modelled and viewpoint changes will be taken into account.

One of the difficulties when employing 2D-models relies on dealing with this viewpoint issue. Most of the previous related works are based on the fundamental assumption of "in-plane" motion or only present results obtained from data satisfying such condition [12]. Few consider motion-in-depth and out-of-plane rotation of the tracked people. Freeing algorithms from the view dependency appears as a critical issue for practical applications. Therefore, the goal of this work is to construct 2D dynamical models that can perform independently of the orientation of the person with respect to the camera and that can respond robustly to any change of direction during the sequence.

1.1. Related work

There are basically two main schools of thought on human pose recovery: model-based *top-down* approaches and model-free *bottom-up* strategies. Model-based approaches presuppose the use of an explicit model of a person's kinematics [9,13]. The number of degrees-of-freedom and the high dimensionality of the state space make the tracking problem computationally difficult. Recent research has investigated the use of learnt models of human motion to constraint the search in state space by providing strong priors on motion [12,14,15]. In *bottom-up* strategies, the individual body parts can be detected and then probabilistically assembled to estimate the 2D pose as in Ref. [16] or an example-based method can be followed. This last method consists in comparing the observed image with a database of samples as in Refs. [17–19] to cite a few. In some cases, a mapping from 2D image space to 3D pose space is learnt for directly estimating the 3D pose [20–22]. Instead of storing and searching for similar examples, Agarwal and Triggs [20] use non-linear regression of joint angles against shape descriptor vectors to distill a large training database into a compact model. Grauman et al. [22] inferred the 3D structure from multi-view contour using a probabilistic "shape+structure" model. As mentioned before, this idea was first introduced by Bowden [11].

Shape-models have appeared as powerful tools for human motion analysis. Baumberg and Hogg [4] used active shape models to track pedestrians from a fixed camera. The same active shape tracker was considered by Siebel and Maybank [6] that extended it by a head detector and a region tracker, all integrated in the visual surveillance system ADVISOR. Fan et al. presented in Ref. [23] a compound structural and textural image model for pedestrian registration. In Ref. [24], the authors exploit the shape deformations of a person's silhouette as a discriminative feature for gait recognition, indicating that methods based on shape perform better than methods based on kinematics alone. Giebel et al. [25] proposed a Bayesian framework for tracking pedestrians from a moving vehicle: a method for learning spatio-temporal shape representations from examples was outlined, involving a set of distinct linear subspace models. Recently, Zhang et al. [26] introduced a statistical shape representation of non-rigid and articulated body contours. To accommodate large viewpoint changes, a mixture of a finite number of view-dependent models is employed.

1.2. Overview of the work

This paper presents a novel probabilistic spatio-temporal 2D-models framework (STMF) for human motion analysis. In this approach, the 2D-shape of the entire body is associated to the corresponding stick figure allowing the joint segmentation and pose recovery of the subject observed in the scene. The first step of this work, described in Section 2, thus relies on the construction of the "shape–skeleton" training data set: contour parameters are associated to the corresponding 2D joints extracted from many different training views of the same walking sequences (varying azimuth angle of the camera).

The framework construction is then detailed in Section 3. First, a novel technique is presented for shape clustering that establishes dynamics correspondences between the different training views. Basically, a structure-based clustering of the training shapes is achieved by partitioning the 3D pose parameters subspace, thus dividing the gait cycle into a series of basic steps. The resulting labelling is then used to construct the non-linear models in each training view where a mixture of PCA models is learned using the expectation maximization (EM) algorithm [10,27,28], the clusters being used at initialization. The method is compared to other two approaches previously developed to deal with non-linearity: nearest neighbor (NN) classifier and independent component analysis (ICA).

Using the motion-based partitioning and the spatial clustering directly provided by the training views, a *spatio-temporal clustering* is obtained in the global shape–skeleton eigenspace: the different clusters correspond in terms of dynamic (temporal clusters) or viewpoint (spatial clusters). A local 2D-model is then built for each spatio-temporal cluster, generalizing well for a particular training viewpoint and state of the considered action. All those models are concatenated and sorted, what leads directly to the construction of the global STMF presented in Fig. 1.

Given this huge amount of data, an efficient search method is needed. In that way, *temporal* and *spatial constraints* are considered to build a probabilistic transition matrix (PTM). This matrix limits the search in the feature space by giving a frame to frame estimation of the most probable local models to be considered during the fitting procedure. This constraint-based search is described in Section 4.

Once the model has been generated (off-line), it can be applied (on-line) to real sequences. Given an input human blob provided by a background subtraction, the model is fitted to jointly segment the body silhouette and infer the posture. This model fitting is explained in Section 5.

Experiments are presented in Section 6 where both segmentation and 2D pose estimation are tested. The main goal of