Survey on Spatio-Temporal View Invariant Human Pose Recovery

Xavier Perez-Sala, Email: xavier.perez-sala@upc.edu a,c,
Sergio Escalera, Email: sergio@maia.ub.es b,c and
Cecilio Angulo, Email: cecilio.angulo@upc.edu a

a CETpD-Upc Technical Research Center for Dependency Care and Autonomous Living, Universitat Politècnica de Catalunya, Neàpolis, Rambla de l’Exposició, 59-69, 08800 Vilafranca i la Geltrú, Spain
b Dept. Mathematics, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007, Barcelona, Spain
c Computer Vision Center, Campus UAB, Edifici 0, 08193, Bellaterra, Spain

Abstract. Human Pose Recovery approaches have been studied in the field of Computer Vision for the last 40 years. Several approaches have been reported, and significant improvements have been obtained in both data representation and model design. However, the problem of Human Pose Recovery in uncontrolled environments is far from being solved. In this paper, we define a global taxonomy to group the existing methods and discuss their main advantages and drawbacks, identifying new challenging lines of research.

Keywords. Human Pose Recovery, Computer Vision.

Introduction

Human body pose estimation, or pose estimation in short, refers to the process of estimating the configuration of the underlying kinematic structure of a person from a sensor input, which is the case of 3D pose recovery, or the 2D projection of the skeletal articulation into the image evidence. Vision-based approaches are often used to provide such a solution, using cameras as sensors. Pose recovery is an important issue for many computer vision applications such as video indexing, surveillance, automotive safety and behavior analysis, as well as many other Human Computer Interaction applications.

Body pose estimation is a challenging problem because of the many degrees of freedom to be estimated. Moreover, limbs vary greatly in appearance due to changes in clothing and body shape, as well as changes in viewpoint.

In Figure 1 we summarize, in chronological order, some of the historical analyses that have been performed during the last centuries in this particular field of research. Some of them will be reviewed in the next sections. In order to update recent advances in the human pose recovery field, providing a general and standard taxonomy to group state-of-the-art approaches, we group the methods in five main modules: appearance, viewpoint, spatial relations, temporal relations, and behavior. Image evidences should be interpreted and related to appearance. Depending on the appearance detected or due to
spatio-temporal post processing, many works infer a coarse or a refined viewpoint of the body, as well as other research restrict the possible viewpoints detected in the training dataset. Since the body pose recovery task implies the location of body parts in the image, spatial relations use to be taken into account. In the same way, when a video sequence is available, the motion of body parts is also studied to refine the body pose or to analyze the behavior being carried out. Finally, the block of behavior refers to those methods that takes into account particular activities or information about scene and context to provide a feedback to previous pose recognition modules to improve final recognition. The global taxonomy used in the rest of the paper is illustrated in Figure 2.

The rest of the paper is organized as follows: Section 2 reviews the state-of-the-art methods grouped in the previous taxonomy. Then, Section 3 discuss advantages and drawbacks of state-of-the-art methods, identifying challenging lines of research.
1. State of the Art

In the next subsection we describe the state-of-the-art related to Human Pose Recovery and group these works in the taxonomy defined in Figure 2.

1.1. Appearance

In order to obtain an accurate detection and tracking of the human body parts, prior knowledge of pose and appearance is required. For a better conceptual understanding of the human appearance methodology, we split the taxonomy in global, local, pixel-based, and logical methods. The main goal of all these approaches is to serve as an initialization for posterior pose recovery approaches.

1.1.1. Global appearance

By global appearance we refer to descriptors or output of classifiers which codify full body information about people in images, either from detection or segmentation.

Silhouettes and contours Global and multiple silhouettes, as well as contours, are used to fit human body in images. An example of using a synthesized knowledge of the image to estimate the human pose is [1], where the pose is mapped directly from the silhouettes through a Mixture of Experts.

Global discriminative classifiers A common technique used for detecting people in images consists on describing image regions using standard descriptors (i.e. HOG [2]) and training a discriminative classifier (e.g. Support Vector Machines) as a global descriptor of human body [2] or as a multi-part description and learning parts [3]. This technique has been widely applied in the field of pedestrian detection in Advanced Driver Assistance Systems (ADAS) (see [4] for a detailed review). Some authors have extended this kind of approaches including spatial relations between inside object descriptors in a second level discriminative classifier, as in the case of poselets [5].

Global generative classifiers As in the case of discriminative classifiers, generative approaches have been proposed to address person detection. However, in the case of generative approaches they use to deal with the problem of person segmentation. For instance, the approach of [6] learns a color model from an initial evidence of person, as well as background objects, to optimize a probabilistic functional using Graph Cuts.
1.1.2. Local appearance

It is widely accepted that describing human body as an ensemble of parts improves recognition of human body parsing approaches [3]. Next, we summarize common descriptors used for describing body parts.

**Edges**  Edges are the most widely applied features invariant to changes in the appearance of a person (i.e. the appearance because of different clothes changes meanwhile external edges of body parts are maintained). In this sense, HOG, SIFT, and wavelets features, among others, use to be considered [2].

**Motion**  Optical flow is the most common feature used to model path motion. Additionally, some other works track visual descriptors and codify the motion provided by some visual regions as an additional local cue [7].

**Color and texture**  Color information is usually codified by means of histograms or space color models (i.e. Gaussian Mixture Model) meanwhile texture use to be an additional cue for local description of body parts once parts have been detected. Texture is then described using, for example, wavelets such as Gabor filters.

**Depth**  Recently, depth cues have been included in several human pose recognition systems because of the depth maps provided by the multi-sensor Kinect\textsuperscript{TM}. The new depth representation based on infrared maps offers near 3D information from a cheap sensor synchronized with RGB data. Based on this representation, novel depth and multimodal descriptors have been proposed [8,9,10,11]. These approaches detect extrema of geodesic maps and compute histograms of normal vectors distribution.

**Templates**  Example-based methods for human pose estimation have been proposed to compare the observed image with a database of samples. One standard technique is to apply a normalized cross-correlation measure between the stored template data set and a test image. A limitation of current example-based approaches is the restriction to the poses used in training, which limits the variability of regions to be detected or increase the number of false positive detections when more template variations are allowed.

Finally, it is important to note that previous state-of-the-art local descriptors require from a first stage of detection the interest points or parts. In this sense, we refer the reader to [12] and [13] for a fair list of region detection and descriptors.

1.1.3. Pixel-based appearance

Some pixel-based approaches have recently showed robust results for the segmentation of human body. This is the case of the Random Forest approach of [14,15], where simple random off-sets of pixel-based depth features are computed and learned in a probabilistic forest of trees. In the approach of [16] another pixel-based classification based on color modeling is presented over RGB.

1.1.4. Logical

It is important to notice that new descriptors including logical relations have been recently proposed. This is the case of the Group-lets approach of [17], where local features are codified using logical operators.
1.2. Viewpoint

Viewpoint estimation is not only useful to determine the relative position and orientation among objects (or human body) and camera (i.e. camera pose), but also allows to significantly reduce the ambiguities in 3D pose [18]. Note that in camera pose literature it is named *pose* in short, however in this section it will be explicitly named camera pose to differentiate from human body posture, named *pose* in the rest of this document. Many works can be found in upper body pose estimation and in pedestrian detection literature, where only front or side views are respectively studied. Just to say an example, while the detector of [19] is in principle capable of detecting people from arbitrary views, its detection performance has only been evaluated on side views. Other works explicitly restrict the possible views, for example, to frontal and lateral viewpoints [20].

Research where 3D viewpoint is explicitly estimated is divided in discrete classification and continuous viewpoint estimation (Figure 2). The discrete approach is treated as a problem of viewpoint classification category, where the viewpoint of a query image is classified into a limited set of possible previous known or unknown views [21,22]. In these works, the 3D geometry and appearance of objects is captured by grouping local features into parts and learning about their relations. Image evidence can also be used to directly categorize the viewpoint. In the first stage of [18] a rough viewpoint is estimated for pedestrians by training eight viewpoint-specific detectors. In the next stage, this classification is used to refine the viewpoint in a continuous way, estimating the rotation angle of the person around the vertical axis. Projections of 3D exemplars of body configurations are evaluated under the previously detected 2D body parts, and the exemplar with the most probable projection is chosen as an initial 3D pose. The continuous approach to viewpoint estimation refers to estimating the real valued viewpoint angles for an example object or human in 3D.

From the point of view of registration, monocular non-rigid shape registration [23] can be seen as similar problem to body pose estimation, since points in the deformable shape could be seen as body joints. Given still images, the simultaneous camera pose and shape estimation is studied for rigid surfaces [24], as well as for deformable shapes [25].

1.3. Spatial Models

Spatial models encode the structure of the human body. Though mapping between image evidences in 3D pose exists (see Section 1.1), their performance is limited to specific datasets. Human body models describe kinematic properties of the body in a hard way (e.g. skeleton, bone lengths) or in a more soft manner (e.g. probabilistic assemblies of parts, grammars). Usually, accurate kinematic constraints are modeled in 3D, as well as degenerate projections of the human body in the image plane are usually modeled by probabilistic assemblies of parts, represented in a graph or a tree configuration.

1.3.1. Probabilistic assemblies of parts

Probabilistic assemblies of parts consist on detecting likely locations of the different body parts in a consistent configuration with the body structure, where such configuration is not defined by physical constraints but is described by soft restrictions.

Pictorial structures [26] are generative 2D assemblies of parts, where each part is detected with its specific discriminative detector. Pictorial structures are a general frame-
Figure 3. Examples of body models as a probabilistic assemblies of parts: a) Pictorial structures [19]; b) Human model proposed in [3]: coarse root filter (left), different part filters with higher resolution (middle), and model for spatial locations of parts (right); c) Hierarchical composition of body “pieces” [30]; d) Spatio-temporal loopy graph [31]; and e) Different trees obtained from the mixture of parts presented in [32].

work for object detection widely used for people detection and human pose estimation [27,19]. Though the original structure was a graph [26], more recent approaches represent the underlying body model as a tree [27].

Grammar models formalized in [28] provide a flexible and elegant framework for object detection [3], also used to detect humans in [3,29]. Compositional rules are used to represent objects as a combination of other objects. In this way, human body could be represented as a composition of trunk, limbs and face; as well composed by eyes, nose and mouth. Moreover, deformation rules leads to hierarchical deformations, allowing the relative movement of parts at each level, though deformation rules in [3] are treated as pictorial structures. However, which makes grammars attractive is their structural variability. Grammar models allow to choice among different subtypes for each part while deal with occlusions. Following this idea, the work of [30] based on poselets [5] represents the body as a hierarchical combination of “pieces”.

Probabilistic assembly of parts can also be performed in 3D when, for example, 3D information is available using a multi-camera system [31]. A similar model to pictorial structures is presented here, where temporal evolution is taken into account. Joints are modeled following Mixture of Gaussian distributions, however here is named “loose-limbed” model because of the loosely attachment between limbs. Instead of a tree, a loopy graph is used where nodes represent 3D position and orientation of body parts. Edges represent relative angle and position between adjacent nodes in space and time. The inference is solved with a particle filter extension for loopy graphs.

A powerful and relatively unexplored graphical representation for human 2D pose estimation are AND-OR graphs [33], which could be seen as a combination between Stochastic Context Free Grammar and multi level Markov Random Fields. Moreover, their structure allows a rapid probabilistic inference with logical constrains [34]. Much
research has been done in the graph inference area, optimizing algorithms to avoid local minima. Multi-view trees represent an alternative because since a global optimum can be found using dynamic programming, hard pose priors, or branch and bound algorithms [32].

In order to deal with high deformations of human body, as well as its changes in appearance, parameters of the body model and appearance could be learned simultaneously [32]. Active Shape Models (ASM) [35] and Active Appearance Models (AAM) [36] are labeled models which are able to deform their shape according to statistical parameters learned from the training set. These approaches provide an alternative to example-based approaches, which compare the image evidence with a database of samples. Examples of probabilistic assemblies of parts are shown in Figure 3.

\subsection{1.3.2. Kinematic models}

Due to the efficiency of trees and because of the human body does not differ too much of an acyclic graph, most of the kinematic models are represented as a tree. Contrarily of trees explained above, whose nodes represent body parts, nodes of kinematic trees usually represent joints, each one parameterized with its degrees of freedom (DOF). In the same way that probabilistic assemblies of parts are more used in 2D, accurate measures of kinematic models have more sense in a 3D representation. However, the use of 2D kinematic models is reasonable for motions parallel to the image plane (e.g. gait analysis). For example, though 3D data from multi-camera system is used in [20], only frontal and lateral 2D models are learned, limiting the performance of the system to both viewpoints. 2D pose is also estimated in [37] from a degenerate 2D model learned from image projections. In this case, not only parallel movements are allowed and different movements are considered when walking in opposite directions.

3D recovery of human pose from monocular images is the most challenging situation in human pose estimation [38]. The recovered number of Degrees of Freedom (DOF) varies greatly among different works, from 10 DOF for upper body pose estimation, to full-body with more than 50 DOF. However, the number of possible poses is huge, even for a model with few DOF and a discrete parameter space. Because of that, kinematic constraints such as joint angle limits are typically applied over kinematic models. Other solutions rely on reducing the dimensionality using unsupervised techniques as Principal Component Analysis (PCA). The continuous state space is clustered in [37], and PCA is applied over each cluster in order to deal with non-linearities of the human body performing different actions. As well as in [20], where it is used a Hierarchical PCA depending on human pose, modeling the whole body as well as body parts separately.

\subsection{1.4. Temporal Models}

Temporal models can be seen as the temporal homologous of spatial models. When a video sequence is available, the motion of body parts may be incorporated to refine the body pose or to analyze the behavior that is being performed.

\subsubsection{1.4.1. Tracking}

Tracking is used to ensure the coherence among poses over the time. Tracking can be applied separately to all body parts or only a representative position for the whole body.
can be taken in account. Moreover, 2D tracking can be performed to the pixel positions or it could be considered that the person is moving in 3D. Other subdivision of tracking is the number of hypothesis, which can be one that is maintained over sequence or several hypothesis can be propagated in time. Other works achieve temporal coherence through a minimization of pose changes along a sequence in batch.

Single tracking is applied in [20], where only the central part of the body is estimated through a Hidden Markov Model (HMM), finally the 2D body pose is recovered from the refined position of the body. Tracking is performed in 2D, however they do not loose generality at these point since they work with movements parallel to the image plane. In contrast, 3D tracking with multiple hypothesis is used in [18]. In the topic of shape recovery, a probabilistic formulation is presented in [39] which simultaneously solves the camera pose and the non-rigid shape of a mesh (i.e. body pose in this topic) in batch. Possible positions of landmarks (i.e. body parts) and their covariances are propagated along all the sequence, optimizing the simultaneous 3D tracking for all the points.

1.4.2. Motion models

The human body can perform a huge diversity of movements, however specific actions could be defined by smaller sets of movements (e.g. in cyclic actions as walking). In this way, a set of motion priors can describe the whole body movements when a single action is performed. However, hard restrictions on the possible motions recovered are as well established. A potential issue of motion priors is that the variety of movements that can be described highly depends on the amount and diversity of the training data [37].

Motion models are introduced in [40], learned from motion capture sequences of walking and running. A reduction of dimensionality is performed by applying Principal component analysis (PCA) over sequences of joint angles from different examples. This work is extended in [41] for golf swings from monocular images. Scaled Gaussian Process Latent Variable Models can also represent more different human motions [42] for concrete actions, such as walking or golf-swings, from monocular image sequences.

1.5. Behavior

The block of behavior refers to those methods that takes into account particular activities or information about scene and context to provide a feedback to previous pose recognition modules to improve final recognition. Most approaches described in previous sections do not include this kind of extra information. In the work of [43], the authors include extra information about human activity and its interaction with objects or other subjects to improve final pose recovery of subjects and activity recognition. This is done by including an extra parameter to a probabilistic graphical model. This approach showed that ambiguities among classes are better discriminated, and better results are obtained.

2. Discussion and conclusion

In this survey, we reviewed past and current trends in the field of human pose estimation. We defined a new taxonomy and grouped state-of-the-art methods in appearance, viewpoint, spatial, temporal, and behavior modules. We reviewed the state-of-the-art descriptors for full body, body parts, and pixel-level segmentation. Most of these approaches
are used as initialization of posterior methods for pose recovery. We showed that main methods for viewpoint analyses can be split in discrete and continuous domains. Spatial models were reviewed and divided into probabilistic assemblies of parts and kinematic models depending on their flexibility. Approaches from the first group result very useful to fit with 2D image evidences since they occur in a 2D degenerate space and kinematic restrictions are too hard to deal with the hough amount of body movements combined with viewpoint and projection. The second group can deal with 3D pose more accurately, reducing the search space through physical constraints. We also reviewed temporal models and split them into tracking and motion models. 3D information in tracking approaches improves 2D methods since nonlinearities due to viewpoint projection are reduced, however it implies computing 3D pose and includes an extra computational cost. Using multiple hypotheses the cumulated error of the tracking procedure is reduced. We showed that strong motion priors help in the pose estimation problem, specially in the challenging case of monocular video sequences. However, usually these approaches are limited to specific actions. Finally, we described the benefits of including extra information related to human activities and context. Scene understanding has recently demonstrated to be a powerful field of research which provides a useful feedback to the object recognition problem, and thus, to the problem of human pose recovery. This kind of inference is not frequently considered in the human pose recovery approaches, but it could be incorporated in a higher layer of knowledge (i.e. “ambient intelligence” layer), where context and scene information can provide feedback to any module of the approach to improve final recognition.

References


