Thesis Proposal

Spatio-Temporal View Invariant Human Pose Recovery in Cluttered Scenes

PhD Student

Xavier Pérez Sala

Advisor

Cecilio Angulo Bahón

Co-Advisor

Sergio Escalera Guerrero

Universitat Politècnica de Catalunya

(UPC)

PhD in Artificial Intelligence

June 14, 2012

Contents

1 Introduction

Human pose recovery, or pose recovery in short, refers to the process of estimating the underlying kinematic structure of a person from a sensor input [1]. Vision-based approaches are often used to provide such a solution, using cameras as sensors [2]. Pose recovery is an important issue for many computer vision applications such as video indexing [3], surveillance [4], automotive safety [5] and behavior analysis [6], as well as many other Human Computer Interaction applications [7, 8]. However, the location of individual body parts is not required in some applications. When the whole body is tracked as a single object, it is termed human tracking or detection [9].

Body pose estimation is a challenging problem because of the many degrees of freedom to be estimated. In addition, appearance of limbs highly varies due to changes in clothing and body shape (with the extreme and usual case of self occlusions), as well as changes in viewpoint manifested in 2D non-rigid deformations. Moreover, dynamically changing backgrounds of real-world scenes make complex the data association among different frames. These difficulties have been addressed in several ways depending on the input data provided. Sometimes, 3D information is available because multiple cameras could be installed in the scene. Nowadays, a number of human pose estimation approaches from depth maps are also being published since the recent market release of low cost depth cameras [10]. In both cases, the problem is still challenging but ambiguities related to the 2D image projection are avoided since 3D data could be combined with RGB information. In many applications, however, only one camera is available. In such cases, either only RGB data is considered when still images are available, or they can be combined with temporal information when input images are provided in a video sequence.

The most of pose recovery approaches recover the human body pose in the image plane. However, recent works go a step further and the human pose is estimated in 3D [11]. Probably, the most challenging issue in 3D pose estimation is the projection ambiguity of 3D pose from 2D image evidences. This problem is particularly difficult for cluttered and realistic scenes with multiple people, partially or fully occluded during certain intervals of time.

Monocular data is the less informative input to address the 3D pose recovery problem, and there is not a general solution for cluttered scenes. There exist different approaches, depending on the activity that people in the video sequence are carrying out, as well as global solutions with limited performance. However, we found a lack of works tacking into account the activity, the task or the behavior to refine the general approach. In Figure 1 it is summarized, 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.

From our point of view, full human pose recovery integrates five modules (shown in Fig. 2): *Appearence, Viewpoint, Spatial relations, Temporal relations* and *Behavior*. State-of-the-art approaches pay more or less attention in these different aspects, however, directly or indirectly these modules are taken into account: Image evidence should be

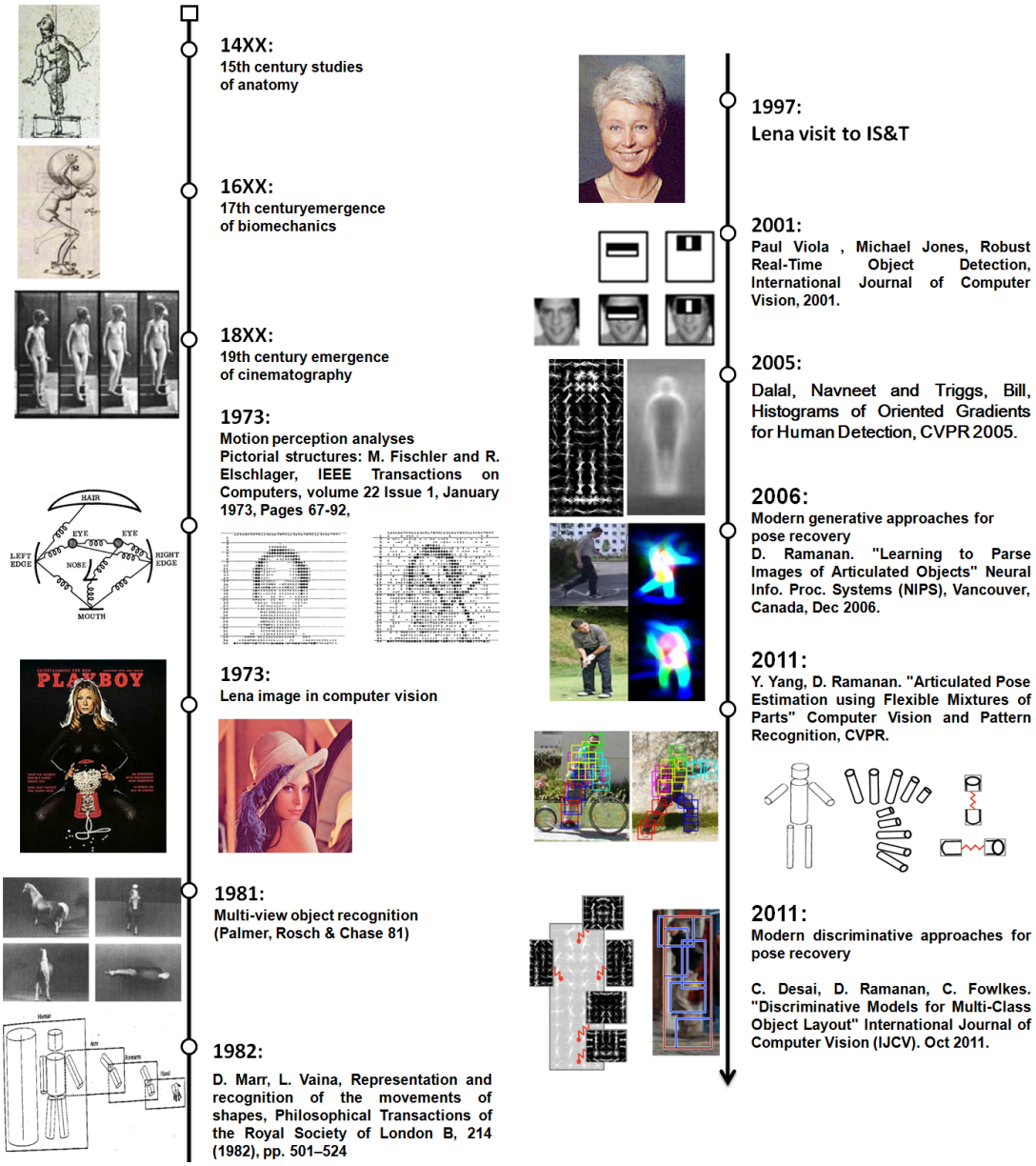

Figure 1: Chronology of Human Pose Recovery and Behavior Analysis, from their philosophical and artistic beginnings, to the most recent applications.

interpreted and related with the *a priori* knowledge of the human 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, it is common in the literature of pose recovery to consider the spatial relations among body parts in the image; In the same way, when a video sequence is available, motion of body parts is also studied to refine the body pose or to analyze the behavior being carried out; the *Behavior* block is an extension to temporal analysis, which can be directly taken into account by a jointly pose and task classification, or indirectly, through the selection of a certain database.

Activities that humans perform are directly related with particular poses. Hence, the main objective of this work is the estimation of the human pose using the feedback between the pose and the task that is being performed. Moreover, spatio-temporal relations among body parts will be also used, including feedback from appearance and viewpoint. The rest of this Thesis proposal is organized as follows: Section 2 introduces the state-of-the-art of human pose recovery, which is discussed in Section 3, together with our initial hypotheses. In Section 4 the goals for this Thesis are proposed, followed by the expected contributions itemized in Section 5.2. Section 5 details the suggested project and the working plan for this Thesis proposal, Section 6 exposes the affiliations and other resources and, finally, our current state of the research is reported in Section 7.

2 State of the Art

Human pose recovery refers to the process of estimating the configuration of the body parts of a person, which is the case of 3D pose recovery, or the 2D projection of the skeletal articulation into the image plane which correctly fits with the image evidence. This process could be preceded by detection and tracking phases, typically used in pedestrian detection applications. Though an initial detection phase usually reduce the computation time of the system, it is achieved at the expense of limiting the possible poses which can be estimated. For more information related to these topics refer to surveys on human detection and tracking [5, 12, 13].

Pose estimation surveys also exit in the literature [9, 14], as well as more general studies involving recent works on vision-based human motion analysis [15, 1]. All of them, besides many woks on this topic, offers a taxonomy. Hence, research is divided in two categories, 2D and 3D approaches, in [15], while [1] defines a taxonomy with three categories: model-free, indirect model use, and direct model use. As far as we know, work in [9] can be considered the most complete survey in the literature. They define taxonomies for model building (i.e. the likelihood function, from human body model, image descriptors, etc.) and estimation (i.e. finding the most plausible pose given a likelihood function).

In order to update recent advances in the human pose recovery field and provide a general and standard taxonomy to group state-of-the-art approaches, reviewed methods are clustered according to the five main modules proposed in [14]: *Appearence, View-*

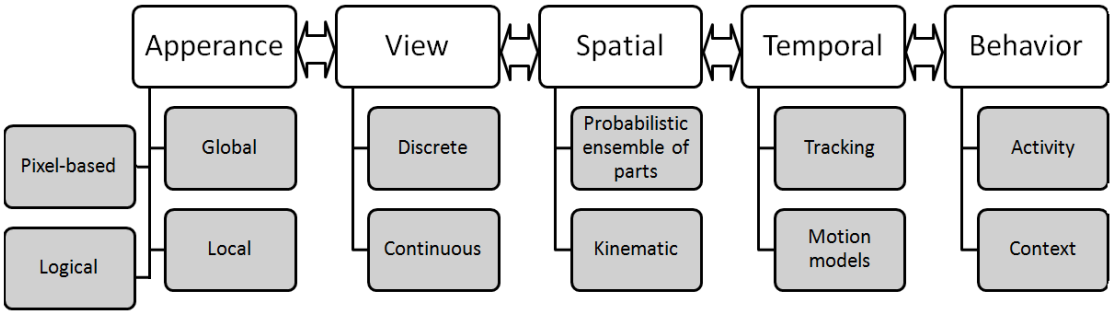

Figure 2: Taxonomy for Human Pose Recovery.

point, Spatial relations, Temporal relations and *Behavior*. Furthermore, subgroups are defined for each of the five main modules of the human pose recovery taxonomy. The whole taxonomy used in the rest of the paper is illustrated in Figure 2.

2.1 Appearance

In order to obtain an accurate detection and tracking of the human body parts, prior knowledge of pose and appearance is required. In this section, state-of-the-art approaches that address the description of human appearance are reviewed. The appearance of people in images varies among different lighting and clothing conditions, including differences in appearance produced by changes in the point of view. Since the main goal is the recovery of the kinematic configuration of a person, the system should generalize over these kinds of variations. This generalization can be partially handled in the image domain by extracting image descriptors. Typical image descriptors include silhouettes, motion, edges, depth or templates, among others. For a better conceptual understanding of the human appearance methodology, appearance taxonomy is divided into *global, local, pixel-based*, and *logical* methods.

2.1.1 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. Main global appearance descriptors are related to silhouettes/contours and global discriminative or generative classifiers.

Silhouettes and contours Silhouettes and their boundaries (contours) provide powerful descriptors invariant to changes of color and texture, as well as they can be extracted in a robust way when background is mainly static. An example of using a synthesized knowledge of the image to estimate the human pose is [16], where the 3D pose is mapped directly from the silhouettes obtained by background subtraction (Figure 3(a)). A Mixture of Experts is used to learn the pose from silhouette shape descriptors. Results

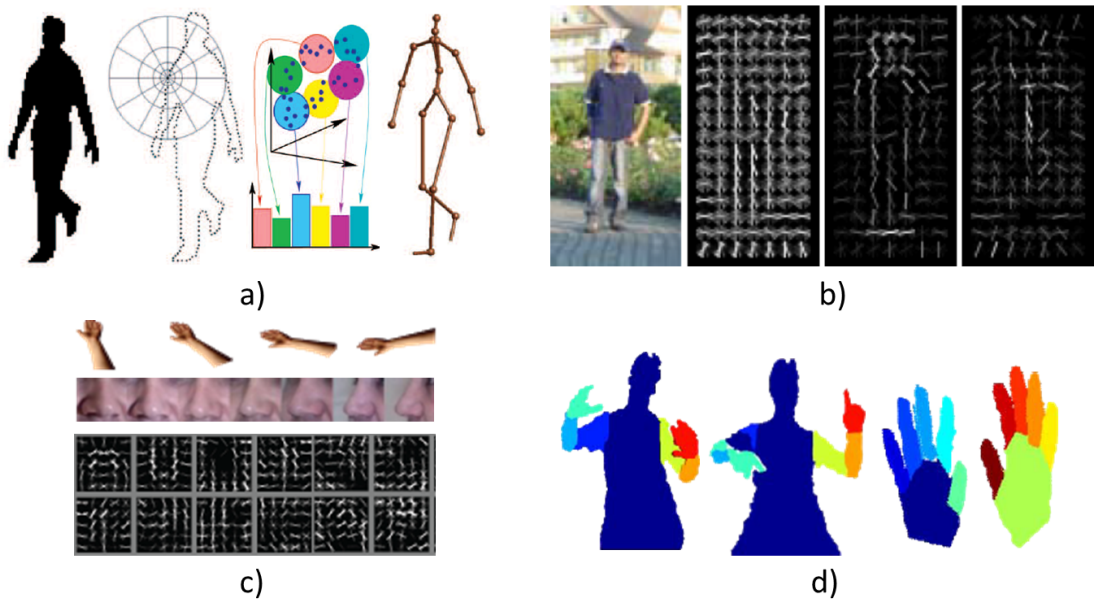

Figure 3: a) Mapping proposed in [16]: Silhouette is extracted through background subtraction, local shape contexts are computed on contour points to fill the histogram description, and three-dimensional pose is obtained from this histogram; b) HOG method introduced in [17]: Image of a person and its HOG descriptor, and this descriptor weighted by the positive and negative classification areas; c) Vocabulary of basis parts introduced in [18]; d) Graph cut approach of [19] for body and hands segmentation.

show walking motions with turns reconstructed from monocular video sequences. However, these methods suffer from bad segmentations in real-world scenes, as well as the difficulty of recovering some Degrees of Freedom (DOF) because of the lack of depth information. For these reasons, approaches based on silhouettes and contours are being revisited because of the recent market release of non-expensive depth cameras. Silhouettes from depth maps can be more robustly segmented, as much as they contain depth information.

Global discriminative classifiers A common technique for detecting people in images consists on firstly describing image regions using standard descriptors (i.e. Histogram of Oriented Gradients (HOG) [17]), then training a discriminative classifier (e.g. Support Vector Machines) as a global descriptor of human body [17] (Figure 3(b)) or as a multi-part multi-view description, and finally learning parts [20] (Figure 4(b)). This technique has been widely applied in the field of pedestrian detection in Advanced Driver Assistance Systems (ADAS) (see [5] for a detailed review). Some authors have extended this kind of approaches also including spatial relations between body parts descriptors in a

second level discriminative classifier, as in the case of the *poselets* approach introduced in [21]. These approaches are usually employed as an initialization technique of posterior pose recovery methodologies.

Global generative classifiers As in the case of discriminative classifiers, generative approaches have been proposed to address people detection. However, in the case of generative approaches they use to deal with the problem of person segmentation. For instance, the approach introduced in [22] learns a color model from an initial evidence of person (or background objects) to optimize a probabilistic functional using Graph Cuts theory.

2.1.2 Local appearance

It is widely accepted that describing human body as an ensemble of parts improves parsing approaches for human body recognition [20]. In this sense, most state-of-the-art descriptors are used to describe human limbs as parts of the image and use the evidence of the responses to perform a global spatial optimization. Next, we summarize standard descriptors used for describing body parts.

Edges Gradient-based features are the most widely applied and invariant descriptors 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 [17], SIFT [23], and wavelets features [24], among others, use to be considered. Moreover, in order to make the local description of body parts invariant to rotation, patch normalization, i.e. rotating the patch by means of gradient orientation [25], is normally applied. A rich vocabulary of parts is constructed in [18], where body parts are expressed as a linear combinations of small sets of parts basis, i.e. HOG filters (Figure 3(c)). Results show improved performance and reduced overfitting in object detection tasks [20], as well as in face detection [26] and body pose estimation [27].

Motion Optical flow [28], as well as a number of variants [29, 30], are the most typical features calculated to model path motion and they can be used to classify human activities [31]. Moreover, some other works track visual descriptors and codify the motion provided by certain visual regions as an additional local cue [32]. In this sense, following the same idea of HOG, Histogram of Optical Flow (HOF) can be constructed [31] to describe regions, as well as body parts movements.

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 regions of interest have been detected. Texture is then described using, for example, Discrete Fourier Transform (DFT) [33] or wavelets such as Gabor filters [24]. Color and texture are not often calculated as main descriptors in the topic of human pose estimation because of the huge variability of the human body in terms of clothes, skin color, and changing backgrounds, among others.

However, these descriptors are applied to perform soft identifications while tracking processes, as shown in [11, 34].

Depth We can not forget about the most recent contributions in the field of visual representation in computer vision. Recently, depth cues have been included in several human pose recognition systems because of the depth maps provided by the multi-sensor Kinect^{TM}. The new depth representation based on infrared maps offers an advantage over traditional time-of-flight systems based on multi-camera systems, providing near 3D information using a cheap sensor synchronized with RGB data. Based on this representation, new depth and multi-modal descriptors have been proposed, as well as classical methods has been revisited taking advantage of new visual cues. Examples are Gabor filters over depth maps for hand description [35] or the approach in [36], that proposes a novel keypoint detector based on saliency of depth maps which is stable to certain human poses. Interest points, based on identifying geodesic extrema on the surface mesh can be classified as, e.g., hand, foot, or head using local shape descriptors. This approach also provides a natural way of estimating a 3D orientation vector for a given interest point. In [37], depth patches are described with a combination of two novel descriptors endowing a description variant to 6 DOF. This description would help to normalize the local shape descriptors to simplify the classification problem as well as to directly estimate the orientation of the body parts in space. A similar descriptor mixed with RGB and motion information is proposed in [38], which is used in a gesture recognition framework [39]. The surveillance system proposed in [4] applies the orientation-invariant Fast Point Feature Histogram [40] based on distribution of normal vectors to identify the robbery of objects in outdoor and indoor environments.

Templates Example-based methods for human pose recovery have been proposed to compare image evidences with a database of samples. One standard technique is to apply a normalized cross-correlation measure among the stored template data set and a query image. These approaches represent a mapping between image space and human pose providing a powerful mechanism for directly estimating 3D pose [11]. However, example-based approaches suffer from several restrictions, such as the huge amount of data to exemplify the variability of the human poses from different viewpoints, and the restriction to the variability of poses or motions used in training. Moreover, these issues are difficult to be solved since a large database of poses may introduce ambiguities in pose estimation.

Finally, it is worth noting that state-of-the-art local descriptors require from a previous detection of interest points or parts. In this sense, we refer the reader to [41] and [42] for a fair list of region detectors and descriptors.

2.1.3 Pixel-based

Some pixel-based approaches have recently showed robust results for the segmentation of human body. This is the case of the Random Forest approach in [43, 19], where

simple random off-sets of pixel-based depth features are computed and learned in a probabilistic forest of trees. As a result, a global segmentation of the human body is provided (Figure 3(d)). In [22, 44] another pixel-based classification based on color modeling is presented over RGB data with successful results.

2.1.4 Logical

Finally, for conceptual completeness about the state-of-the-art on appearance approaches for human pose recovery, it is important to notice that new descriptors including logical relations have been recently proposed. This is the case of the Group-lets approach [45], where local features are codified using logical operators, increasing the discriminant capability of the classifiers and showing improved performance recognizing human actions in still images in comparison to classical approaches.

2.2 Viewpoint

Viewpoint estimate is useful to determine the relative position and orientation between objects (or human body) and camera (i.e. camera pose), and also allows to significantly reduce ambiguities in 3D pose [11]. Note that in camera pose literature it is named *pose* in short, however, in this section it will be explicitly named camera pose. The word *pose* will keep the same meaning as in the rest of the document: human body posture.

Usually, body viewpoint is not directly estimated in human tracking or pose recovery literature, however, it is indirectly considered because the possible viewpoints to be detected are constrained, for example, in the training dataset. Many woks can be found in upper body pose estimation and pedestrian detection literature, where only front or side views are respectively studied. Just to say an example, while in [25] a detector is presented which is able to detect people from arbitrary views, its performance has only been evaluated on walking side views. Other works explicitly restrict the possible views, for example, to frontal and lateral viewpoints [46].

Research where 3D viewpoint is explicitly estimated can be divided into discrete classification and continuous viewpoint estimation. 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 initially known [47, 48] or unknown [49] views. In these works, 3D geometry and appearance of objects is captured by clustering local features and learning their relations. Image evidence can also be used to directly categorize the viewpoint. In the first stage of [11] a rough viewpoint is estimated for pedestrians by training 8 viewpoint-specific detectors. In the following 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 examples of body configurations are evaluated under the previously detected 2D body parts, and the sample with the most suitable projection is chosen as a 3D pose proposal. The continuous approach to viewpoint estimation refers to computing the real valued viewpoint angles for an object or human in 3D. In [50], discrete and continuous viewpoint estimation are treated, as well. Discrete viewpoint is classified in a set of canonical views through a mixture-of-HOG

approach, focusing on viewpoints instead of categories [20]. Assuming orthographic projection, continuous viewpoint is measured by extending the mixture model to deal with offset viewpoint angles, with respect to the canonical orientations.

Continuous viewpoint estimation is widely studied in the field of shape registration, which refers on finding correspondences between two sets of points and recovering the transformation that maps one point set to the other. Monocular non-rigid shape registration [51] can be seen as a similar problem to body pose estimation, since points in the deformable shape can be interpreted as body joints. Indeed, at least one approach exists which provides a solution for both problems in a common framework [52]. Given still images, simultaneous camera pose and shape estimation is studied for rigid surfaces [53], as well as for deformable shapes [54]. In both works, prior knowledge of the camera is provided by modeling the possible camera poses as a Gaussian Mixture Model (GMM), which consist on uniform camera poses [55] in certain regions around the modeled surface. In [53], point correspondences and camera pose are iteratively established by hypothesizing the projections of the known 3D rigid shape onto the different camera priors. In [54], they go a step further by extending the previous work for deformable surfaces. Possible deformations of the shape are also modeled as a GMM. The simultaneous camera pose, correspondences and non-rigid shape are estimated by the joint hypothesis of both GMM. Following this work, in [56] is presented a probabilistic formulation for video sequences, inspired on Simultaneous Localization And Mapping (SLAM).

In [57] and [58], the viewpoint of the head is estimated from depth data acquired with range scanner and low cost depth cameras, respectively. Both works can deal with partial occlusions and different facial expressions, since they first detect the parts of the image belonging to the head using discriminative random regression forests. Moreover, as it is explained in previous section (Section 2.1), camera pose estimation can be directly estimated from image evidences because depth information allows to build descriptors variant to 6 DOF.

2.3 Spatial models

Spatial models encode the structure of the human body. Though there exist approaches which directly map the appearance to 3D pose (see Section 2.1), 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. pictorial structures, 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.

2.3.1 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 also is described by soft restrictions which can deal with the high variability of the body poses and viewpoints.

Pictorial structures [59] are generative 2D assemblies of parts, where each part is detected with its specific discriminative detector. Pictorial structures are a general framework for object detection widely used for people detection [20, 25] and human pose estimation [60, 25, 61]. Though the traditional structure for representation is a graph [59], more recent approaches represent the underlying body model as a tree, due to inference facilities studied in [60]. Constraints between parts are modeled following Gaussian distributions, which seems does not match, for example, with a typical walking movement between thigh and shank. However, Gaussian distribution does not correspond to a restriction in the 2D image plane, and it is applied in a parametric space where each part is represented by its position, orientation and scaling [60]. A general approach for pedestrian detection and 2D body pose estimation is presented in [25], where strong part detectors were discriminatively trained. Moreover, during the fitting phase, margin of classifiers are used as likelihood in the generative model (shown in Figure 4(a)).

Grammar models formalized in [62] provide a flexible and elegant framework for object detection [20], also used to detect humans in [20, 63]. Compositional rules are applied 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 (e.g. eyes could be displaced with respect to the face as well as its displacements are related to the whole body). Though deformation rules in [20] are treated as pictorial structures (shown in Figure 4(b)), which makes grammars attractive is their structural variability. Grammar models allow to choice among different subtypes for each part while deal with occlusions [63]. Following this compositional idea, [64] is based on *poselets* [21] to represent the body as a hierarchical combination of body "pieces" (shown in Figure 4(c)).

Probabilistic assembly of parts can also be performed in 3D when, for example, 3D information is available using a multi-camera system [65]. A similar model to pictorial structures is presented in [65], where temporal evolution is taken into account (shown in Figure 4(d)). 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 (i.e. adjacent body parts in the same frame, and the same body part in adjacent frames). The inference of 3D human pose is solved with a particle filter extension for loopy graphs. However, the presented system requires a background subtraction step. They overcome this issue in [66], where authors also deal with a monocular image sequence. The human pose problem is divided in three stages: first of all 2D body pose is estimated using their previous work, then 3D human pose is reconstructed following the mapping approach explained above [16], and finally 3D poses for walking people are refined through Bayesian inference.

A powerful and relatively unexplored graphical representation for human 2D pose estimation is AND-OR graph [67], which could be seen as a combination between Stochas-

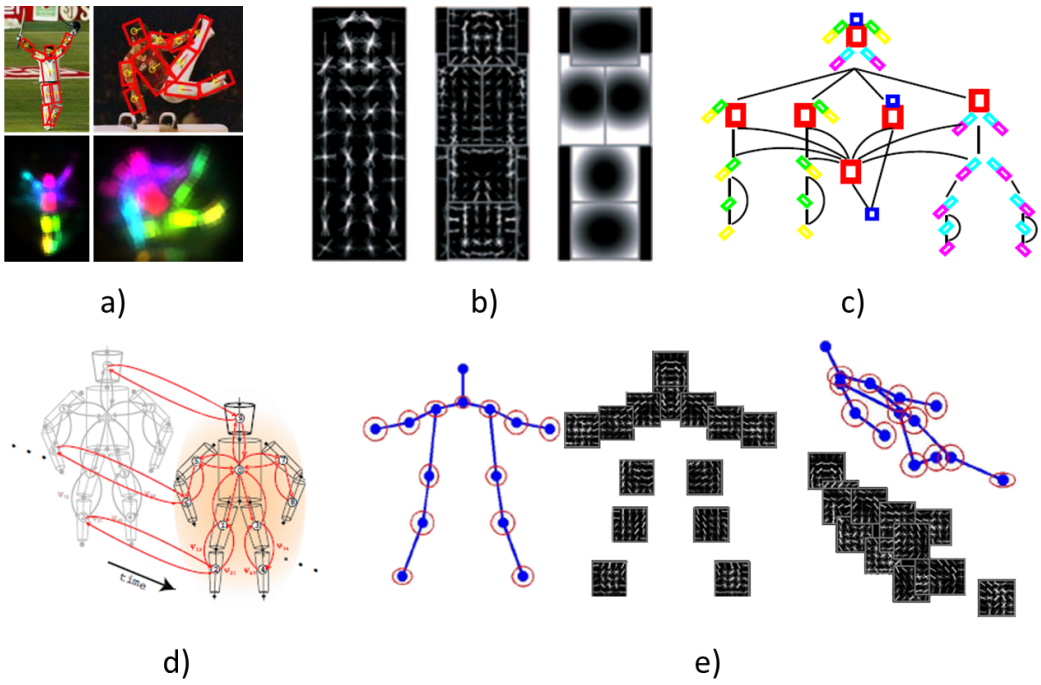

Figure 4: Examples of body models as a probabilistic assemblies of parts: a) Pictorial structures presented in [25] estimating the human 2D pose (top) for different sports from likelihoods of parts detectors (bottom); b) Human model proposed in [20]: coarse filter (left), different part filters with higher resolution (middle), and model for spatial locations of parts (right); c) Hierarchical composition of body "pieces" of [64]; d) Spatio-temporal loopy graph proposed in [65]; e) Different trees obtained from the mixture of parts presented in [27].

tic Context Free Grammar and multi level Markov Random Fields. Moreover, their structure allows a rapid probabilistic inference with logical constrains [68]. A fully connected graph to represent body models as a probabilistic assembly of parts is presented in [69], also using discriminative part detectors. However, restrictions of the method, such as absolute orientations of detected parts, only allows to be applied in upright poses. In addition, fully connected graphs are difficult to optimize. A lot of research has been developed in this area, optimizing algorithms to avoid local minima. Multi-view trees could be an alternative because a global optimum can be found using dynamic programming [27]. In this case, a mixture of parts with a tree as the underlying model encode the spatial structure of a human body. Powerful descriptors are calculated and learned jointly with relative positions between adjacent body parts. 2D pose is successfully computed for different datasets, showing high flexibility and efficiency. They exploit the fact that the parameters of tree models can be efficiently inferred, however trees also suffer the well-known double-counting phenomena, which can be solved by

considering hard pose priors [70] or branch and bound algorithms [71]. However, these solutions are more susceptible to suffer overfitting for specific datasets.

In order to deal with high deformations of human body, as well as its changes in appearance, the parameters of the body model and appearance could be learned simultaneously [27]. Active Shape Models (ASM) [72] and Active Appearance Models (AAM) [73] are labeled models which are able to deform their shape according to statistical parameters learned from 2D or 3D [74] training set. AAM, moreover, are able to learn the appearance surrounding the anatomical landmarks, reliably labeled in the examples of the training set. Though ASM and AAM learn models for the whole body [75], the local appearance and deformations of body parts learned allow a 2D body pose estimation. These approaches provide a solution versus example-based (see Section 2.1) approaches, which compare the image evidence with a database of samples. While the body parts detection in [11] is performed by multi-view pictorial structures, 3D reconstruction is estimated by projecting 3D examples over the 2D image evidence.

2.3.2 Kinematic models

Due to the efficiency of trees and similarity between human body and acyclic graphs, most of the body kinematic models are represented as a tree. Contrarily to trees explained above, whose nodes represent body parts, nodes in kinematic trees usually represent joints, each one parametrized 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, 2D kinematic model is a reasonable tool for motions parallel to the image plane (e.g. gait analysis). For example, though 3D data from multi-camera system is used in [46], only frontal and lateral 2D models are learned, limiting the performance of the system to both viewpoints. 2D pose is also estimated in [76] from a degenerate 2D model learned from image projections. In this case, not only parallel movements are allowed, so different movements are interpreted when walking in opposite directions.

3D recovery of human pose from monocular images is the most challenging situation in human pose estimation due to projection ambiguities. Since information is lost during the projection from real world to the image plane, several 3D poses match with 2D image evidences [77]. Kinematic constraints on pose and movement are typically employed to solve the inherent ambiguity in monocular human pose reconstruction. Therefore, different works have focused on reconstruct the 3D pose given the 2D joint projections from inverse kinematics [78, 79], as well as the subsequent tracking [80, 81]. In [80], the human body is modeled as a kinematic chain, parametrized with twists and exponential maps. Tracking is performed in 2D, from a manual initialization, projecting the 3D model into the image plane under orthographic projection. This kinematic model is also used in [82], adding a refinement with the shape of garment, providing a fully automatic initialization and tracking. However this multi-camera system requires a 3D laser range model of the subject which is being tracked. In [77], 3D pose is estimated projecting a 3D model onto the image plane in the most suitable view, through perspective image projection. The computed kinematic model is based on hard constraints on angle limits

and weak priors, such as penalties proportions and self collisions, inspired in a strong human knowledge.

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. The number of possible poses is very high, even for a model with few DOF and a discrete parameter space. Hence, kinematic constraints such as joint angle limits are typically applied over kinematic models. Other solutions rely on reducing the dimensionality by choosing characteristic poses [6] or using unsupervised techniques as Principal Component Analysis (PCA). A set of distinctive pose priors is manually chosen in [6], or learned from Motion Capture (MoCap) 3D data if available. For each different action in database, keyposes are selected which result in discontinuities in pose energy. Then, 3D pose can be estimated by projecting pose priors onto the image evidence. In [46] it is used a Hierarchical PCA depending on human pose, modeling the whole body as well as body parts separately, allowing complex deformations. More sophisticated 3D models could be used, where the human shape is modeled in addition to the kinematic structure. Skin is modeled as rectangular or trapezoidal patches [83] and can be used for human segmentation [84], however, these complex models use to need a manual initialization, multi-camera systems or accurate 3D models [82].

2.4 Temporal Models

Temporal models can be seen as the temporal counterpart 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 carried out. The state-of-the-art of temporal techniques is divided into tracking and motion models.

2.4.1 Tracking

Tracking is a temporal technique to ensure the coherence among poses over the time. Tracking can be applied separately to all body parts, as well as only to a representative position of the body. Moreover, 2D tracking can be performed to the pixel positions or it can be considered that the person is moving in 3D. Tracking techniques can also be divided according to the number of hypothesis, which can be one that is maintained over the time or several hypotheses can be propagated in time. Other works achieve temporal coherence through the minimization of pose changes along a sequence in batch.

Single tracking is applied in [46], where only the central part of the body is estimated through a Hidden Markov Model (HMM), finally the 2D body pose is estimated from the refined position of the whole 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 considered in [11]. First, the whole body is tracked in 2D, then 3D poses at each frame are estimated and propagated along all the sequence, finally the results are refined with a Bayesian framework, achieving consistent tracking and 3D pose estimation for all frames. Note that joint 3D pose and tracking allow an implicit tracking of the body parts. In the topic

of shape recovery, a probabilistic formulation is presented in [56] 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/joints) and their covariances are propagated along all the sequence, optimizing the simultaneous 3D tracking for all the points.

2.4.2 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, though 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.

Statistical motion models are widely applied for human tracking and 3D pose estimation, usually learned from Motion Capture (MoCap) using multi camera systems. Since human body motions tend to be highly non-linear, common methods cluster the state space and specifically reduce the dimensionality in each region, where clusters are usually directly related to activities being performed in the database. Different approaches differ on the state representation and clustering methods, as well as in procedures for dimensionality reduction.

Consecutive 3D positions of each body joint are clustered in [81], clusters being described with their principal eigenvectors using Singular Value Decomposition (SVD). Here, the reduction of dimensionality of human gestures helps to estimate the 3D pose at each frame. In [76], the state space represented by joint angles is also clustered, PCA is applied over each cluster to reduce the dimensionality and a Gaussian auto regressive process is applied in order to deal with non-linearities of the human body performing different actions. However, the number of possible movements in the video sequence is a critical parameter, since a same action seen from different viewpoints can be interpreted as different movements. In [85], motion models are learned from MoCap sequences of walking and running. A reduction of dimensionality is performed by applying PCA over sequences of joint angles from different examples. This work is extended in [86] for modeling golf swings from monocular images. Scaled Gaussian Process Latent Variable Models (SGPLVM) can also represent more different human motions [87] such as walking and golf-swings together from monocular image sequences.

A clear weakness of using priors is overfitting on the training data because they can only generalize over a small set of specific movements. In [88], a general trajectory based on the Discrete Cosine Transform (CDT) is introduced to reconstruct different movements from, for example, faces and toys. In this case, trajectory model is combined with spatial models of the tracked objects. Applications of such motion models related to human pose can be found in [89], where it is achieved a 3D reconstruction of moving points tracked from humans and scenes; as well in [90], where articulated trajectories are reconstructed for upper body models.

2.5 Behavior

The block of behavior refers in our taxonomy to those methods that take into account particular activities or information about scene and context, to provide a feedback to precedent pose recognition modules, improving the final recognition task. Most approaches previously described do not directly include this kind of extra information. However, databases are usually organized by actions which are being performed (e.g. walking, jogging, boxing [91]) and algorithms use to over-fit these actions (e.g. walking [11], golf swings [86]). From our point of view, the election of a specific training dataset is a direct or indirect choice of the set of actions that the system will be able to detect. It is important to point out that taxonomies in the literature for behavior, activity, gesture and sub-gesture, for example, are not broadly detailed. The term *behavior* is used here as a general concept which includes actions and gestures.

Though is not usual, some works exist taking into account behavior or activity to estimate a better body pose, learning different models depending on the action that is being performed. Different subspaces are computed for each action in [76]. However, the number of actions chosen is a critical parameter, since actions seen from different viewpoints are interpreted as different movements. This phenomenon occurs because a degenerate 2D model is learned from image projections, instead of building a 3D view invariant model. Some works in the literature go a step forward and jointly recover pose and behavior. In the work of [92], the authors include extra information about human activity and its interaction with objects to improve final pose estimation of subjects and activity recognition. This technique includes "the object" as an extra parameter in a probabilistic graphical model. It was demonstrated that ambiguities among classes are better discriminated, and better results are obtained. In [93], monocular 3D body pose and viewpoint are estimated from an activity-specific manifold. However, though they achieve good generalization among different body shapes, manifolds are learned from 2D visual inputs for specific viewpoints and activities. So, the generalization to other motion domains is not clear. Finally, the work in [6] takes profit from such joint estimation of human pose and action being performed. Here, a set of pose priors is learned (or manually chosen) for each action, as well as Gaussian distributions for each joint, to deform the skeleton between pose priors. Then, after action estimation during test phase, 3D pose is accurately recovered using the specific pose priors of such action. Though a joint approach for pose tracking and action recognition in cluttered scenes is presented in this work, they do not consider any feedback between both estimations. In addition, since Gausssian distributions represent the space search of each joint, they could improve the performance of the system by projecting the covariances to the image plane, and not just the 3D skeleton to fit 2D image evidences. Other interesting improvement of this work could be the addition of motion models to pose priors, instead of Gaussian distributions, in order to achieve a more accurate tracking.

3 Initial hypothesis

Recovering the 3D pose of humans from monocular images is an ill-posed problem. Some information is lost because 2D image evidence is a projection of the real world, i.e. different human poses can have the same observations, as well as similar poses can result in very different observations. In the literature, human pose is retrieved by constraining the search space in appearance (e.g. learning strong filters for body parts [18]), limiting the feasible viewpoints [11], and confining the body parts in certain regions [25, 27], as well as restricting the possible configurations of these parts by kinematic constraints [76, 80, 77]. Temporal coherence is also taken into account [11], likewise motion of body parts is restricted to be consistent with previously known human movements [87]. As far as we know, though certain works jointly estimate pose and activity [92, 6], as well as viewpoint [93], our Thesis proposal is the first approach for body estimation which studies the possibilities of mixing all this multi-modal knowledge.

State-of-the-art presented in the previous section is divided in the five focus of interest of this work, nevertheless, it is also important to take into account the relations among them:

- The selection of reliable body part detectors is a key step in the most of approaches that estimate human pose from monocular RGB cues [25, 27, 11, 20]. On the other hand, promising works which do no pay enough attention to image evidences could be seen their results improved by using better detectors [60]. **Strong part detectors reduce the dimensionality of the problem, relaxing the difficulties resulting from cluttered scenes, while maintaining the useful information to estimate the human pose**.

- Unlike motion priors, pose priors are not often used in body pose estimation. Though priors on body configuration are applied in the literature, they use to learn pose deformations in 2D [76, 46]. Learning the 2D projection of a 3D object deals to unrealistic kinematic constraints [76] or allows to estimate only natural movements when such motions are parallel to the image plane [46]. By contrast, 2D models are learned from 3D objects in [74], as well as 3D pose priors are used in [6] in an action recognition framework. However, 3D poses are chosen very different among them in order to cover the action-pose estate, instead of dealing with an accurate kinematic reconstruction. **Pose priors on 3D kinematic skeletons should decrease non-linearities of human body, usually magnified by non-linearities due to camera projection, improving flexibility and performance of kinematic models**.

- Human pose can have many different observations, because of the variations between people in shape and appearance, as well as different environments and camera viewpoints. Furthermore, different poses can result in the same observation. In order to solve these problems, it makes sense the approach introduced in [54], where the search space is restricted by combining possible camera poses and shape model (i.e. body model). Moreover, camera poses and shape deformations (i.e.

body poses) are learned separately. Hence, the potentially huge amount of data resulting of all possible views combined with of all possible body poses is avoided. Going a step further to [77], where a 3D skeleton is projected to the image plane, **the joint 3D body pose and camera viewpoint should decrease the non-linearities of body pose and camera projection, achieving more accurate results at the expense of increasing the computational time**. Then, 3D uncertainty of space search can be projected onto the 2D image plane [54], in order to match with body parts detectors and accelerate the process.

- 3D body pose from monocular image sequences is a hard problem with well known ambiguities [77]. Moreover, while body poses and camera poses are going to be learned in 3D, certain restrictions could be needed to reduce the search space. Temporal coherence in 3D [11] is a powerful tool to reduce the uncertainty, while maintaining 3D information. However, in order to reduce more the state space, 3D motion models can be learned for certain actions (e.g. walking). Hence, **combined spatio-temporal models should provide a robust framework to estimate 3D body pose from monocular image sequences in a small number of frames**.

- Both motion models and spatial models are attempts to explain spatio-temporal deformations in different dimensions. Nevertheless, they are supposed to be independent and both learnings are performed separately. Therefore, important information for human motion only noticed in a joint domain could be ignored because of a poor relevance in temporal or in spatial domains separately. In addition, independent learning and application of both models leads to a bottleneck of computational time and memory storage. To provide a solution to these issues, **jointly learned spatio-temporal models should decrease the stored data and the estimation time, while providing similar or even better results**.

- Motion priors [87, 88] and spatial priors [6] reduce the search space in spite of limiting the human poses and movements which can be detected by the system. In order to relax these restrictions on the possible detected behaviors, a variety of activities can be added into the training and test databases [6]. However, an important drawback is the increasing of reconstruction ambiguities, because similar poses or motions can be produced by different actions. Unsupervised learning over pose and motion spaces could help to solve these problems. **Choosing the specific model, depending on the pose or motion that is being performed, should relax the hard prior constraints while maintaining the reduced search space in multi-activity databases**.

- Since different motion and pose models could improve the performance and flexibility of the system, a reliable mechanism for models selection shall be an important focus of research. In this way, **each model should have a direct relation with the activity or behavior that is being detected, which leads to jointly estimate behavior and human pose**. Therefore, global solution could

be refined by specific spatio-temporal models of activities in the training set, in a simultaneous activity and human pose estimation framework.

4 Goals

The main goal of this Thesis proposal is to develop methods for estimating 3D human pose in cluttered scenes, providing feedback from activity recognition. To this end, self-contained goals are proposed in a staggered manner:

- Strong part detectors reduce the dimensionality of the human pose problem, relaxing the difficulties resulting from cluttered scenes, while maintaining the useful information to estimate the human pose. Our work is not supposed to do research on novel descriptors, therefore an **intensive research on state-of-the-art body part descriptors** is required, complementing the research which have been done yet.

- Since pose priors on 3D kinematic skeletons should decrease non-linearities of human body, improving the flexibility and performance of kinematic models, one important objective of this work is the **construction of 3D pose priors able to deform the skeleton as humans do when performing certain tasks**.

- Research proposed in this work requires a large amount of data to train and test the final system. Consequently, it must be done an **intensive research on available databases, as well as complementing the state-of-the-art with new databases**.

- The joint 3D body pose and camera viewpoint should decrease the non-linearities of body pose and camera projection, achieving more accurate results at the expense of increasing the computational time. Another important goal of this Thesis proposal is the **joint body pose and camera estimation from monocular image sequences, given 3D priors of the skeleton and the camera poses**.

- Combined spatio-temporal models should provide a robust framework to estimate 3D body pose from monocular image sequences in a small number of frames. Therefore, it should be done an intensive **study of the applicability of general motion models [88], as well as action specific motion priors [87], to be jointly combined with 3D kinematic body models**.

- Jointly learned spatio-temporal models should decrease the stored data and the estimation time, while providing similar or even better results than both models learned separately. In consequence, we will **study the benefits of the joint learning of motion priors and body pose priors for certain human actions**.

- Unsupervised learning over pose and motion spaces could help to solve reconstruction ambiguities when using multi-activity databases. Choosing specific models,

depending on the pose or motion that is being performed, should relax the hard prior constraints while maintaining the reduced search space in such databases. Therefore, it is required a strong **study on clustering pose priors, as well as motion priors, and the recursive optimization algorithm for finding the pose which maximizes the cluster, while finding the best cluster to reduce the search space**.

- Since learned pose and motion models should have a direct relation with activity or behavior being detected, the final goal proposed is the **joint estimation of behavior and human pose**.

5 Project

The Thesis proposed here is referred to do research in the field of human pose estimation. In previous sections, the state-of-the-art (Section 2) is presented, it is discussed and initial hypotheses are presented (Section 3), and goals are proposed in Section 4. In the incoming sections, an overview of the suggested framework will be exposed in Section 5.1, followed by the expected contributions of this Thesis, itemized in Section 5.2, then, the working plan of this project is detailed in Section 5.3, resources are listed in Section 6 and, finally, the current state of research is reported in Section 7.

5.1 Overview

This Thesis is devoted to explore the current limits of the state-of-the-art in the field of human pose estimation and provide novel solutions to unsolved problems. Human pose recovery is an important issue for many computer vision applications such as video indexing, surveillance, automotive safety and behavior analysis, for example. However, body pose estimation is a difficult problem because many degrees of freedom have to be estimated, appearance of limbs highly varies due to changes in clothing and body shape, as well as changes in 3D viewpoint are manifested as 2D non-rigid deformations. In order to outperform these difficulties, a framework is proposed to jointly estimate the 3D human body pose and camera viewpoint, restricting the search-space with strong body part detectors and motion priors, as well as receiving feedback from behavior analysis (see Figure 5.) The proposed Thesis can be summarized in the following goals:

- Intensive research on state-of-the-art body part descriptors.
- Intensive research on available databases, as well as complementing the state-of-the-art with new databases.
- Construction of 3D pose priors able to deform the skeleton as humans do when performing certain tasks.
- Joint body pose and camera estimation from monocular image sequences given 3D priors of the skeleton and the camera poses.

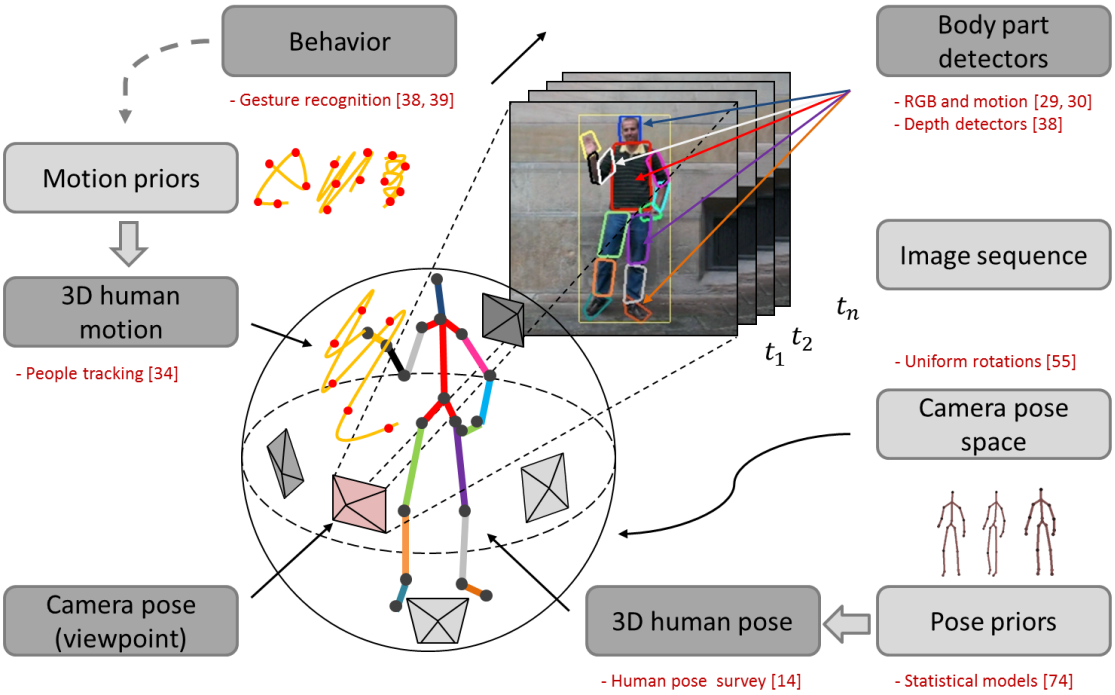

Figure 5: Overview of the proposed system and published contributions in red.

- Study of the applicability of general motion models, as well as action specific motion priors, to be jointly combined with 3D kinematic body models.
- Study the benefits of the joint learning of motion priors and body pose priors for certain human actions.
- Study on clustering pose priors, as well as motion priors, and the recursive optimization algorithm for finding the pose which maximizes the cluster, while finding the best cluster to reduce the search space.
- Joint estimation of behavior and human pose.

5.2 Expected contributions

Fulfill the objectives presented in Section 4 would result in the following contributions:

1. Contribute to the development of techniques dealing with 3D human pose estimation from strong part detectors, searching for the most suitable configuration of 3D body priors which maps into the image likelihoods, making all computations in 3D until the 2D image fitting.

2. Extend the methods available in the literature for the simultaneous reconstruction of deformable meshes and camera pose estimation to the field of human pose recovery. In particular, we plan to propose a method for jointly estimate the body pose and the camera viewpoint from monocular image sequences.

3. Define methods to reduce the search-space without limiting the flexibility of the system by refining 3D body pose and tracking by specific models. In particular, specific motion priors, to be jointly combined with 3D kinematic body models, also learned for certain tasks.

4. Study the joint estimation of behavior and human pose, through a recursive framework which exploits the information of both estimations.

5.3 Working Plan

This section describes the expected tasks in the development of the proposed research, moreover the foreseen planning is presented in Figure 6 as a in a Gantt chart that spans over four years.

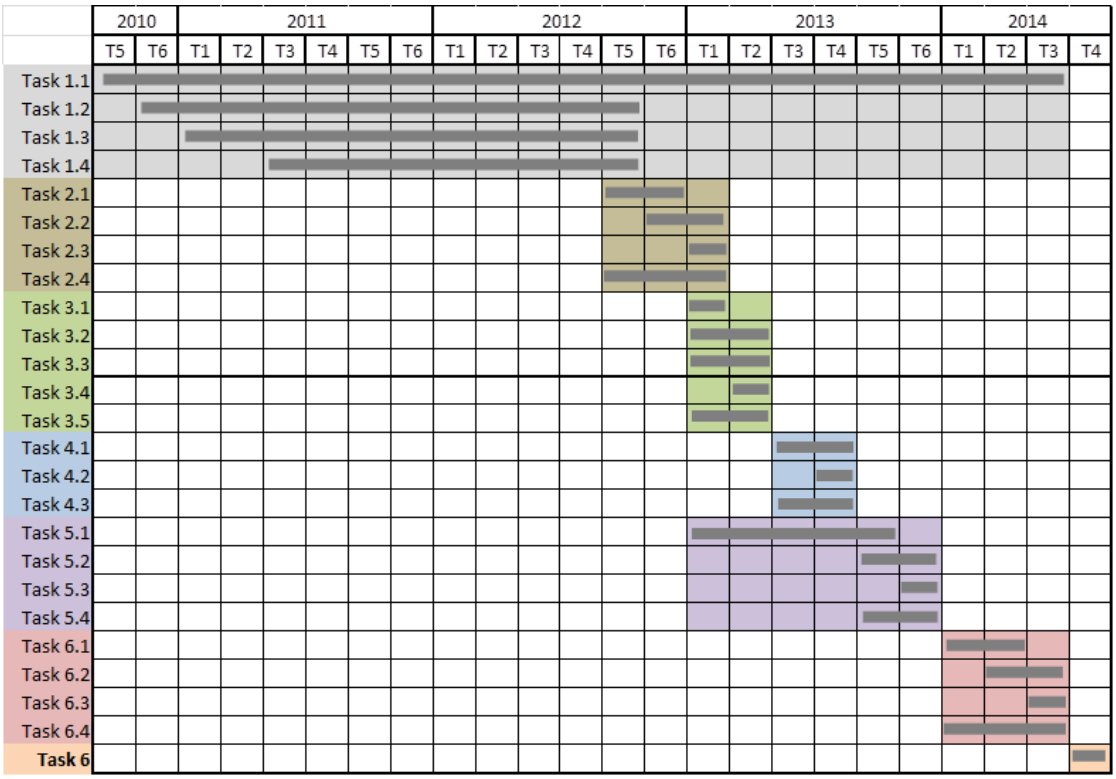

Figure 6: Work planning for the proposed Thesis. Each column represents a two months period.

5.3.1 Task 1: State of the art, Library of body part descriptors and Databases

The first step of the proposed Thesis is to explore the state-of-the-art, paying a special attention to body parts detectors and human datasets. Our work is not supposed to do research on novel descriptors, therefore an intensive research on state-of-the-art body part descriptors is required. In this way, obtain or build databases for learn body part descriptors is also a must, as well as databases with 3D information to test the global framework of 3D human pose estimation.

Task 1.1: State-of-the-art of human pose recovery, viewpoint estimation and behavior analysis.

Task 1.2: Construction of a library including the state-of-the-art detectors to find body parts in cluttered scenes, incorporating current trends and novel detectors proposed.

Task 1.3: Extensive search for public databases must also be done. Databases with labeled body parts are needed to learn and test body part detectors, as well as datasets including 3D information of humans are required to test the 3D human pose estimation.

Task 1.4: Validation & publication: Validation of the proposed methods on to the state-of-the-art databases, in comparison with current trends and similar methodologies. Final results will be summarized and published in international scientific forums.

5.3.2 Task 2: 3D body pose from 2D image evidences

The second step of this Thesis proposal searches to infer the human 3D pose from video sequences. In this first approach, it will be done for only one subject and hard kinematic constraints (e.g. bone length).

Task 2.1: Build likelihood maps for body parts: head, trunk and limbs. This task includes the election of the most suitable part detectors and the study of the most reliably detected parts to sort the subsequent image search.

Task 2.2: Estimate 3D pose during walking action given body pose priors with fixed bone lengths and fixed camera viewpoint, searching for the most suitable pose which matches with likelihood maps. Though it is the first approximation to 3D human pose in this work, which makes this task challenging is that all computations are performed in 3D, instead of the image search, which is a projection over 2D image likelihood.

Task 2.3: Estimate 3D pose for several actions, maintaining previous restrictions. This task includes the learning of different pose priors for more complex actions than walking, such gestures or jumping, and be able to estimate the 3D pose for all of them.

Task 2.4: Validation & publication: Validation of the proposed methods on to the state-of-the-art databases, in comparison with current trends and similar methodologies. Final results will be summarized and published in international scientific forums.

5.3.3 Task 3: 3D body tracking

This step refers to the refinement of the 3D human pose through the addition of motion information in conjunction with 3D kinematic body models. By adding motion priors, restrictions of previous tasks shall be relaxed after this analysis. However, previous constraints of fixed bone length and fixed camera viewpoint should be maintained during this task.

Task 3.1: Motion priors in 3D will be learned for walking action. All the process will be performed in 3D until the image search, projecting the 3D information over 2D image likelihoods of body parts.

Task 3.2: General vs. specific motion priors study about the applicability of both models. General priors and action specific motion models will be tested, as well as the combination of both approaches in a hierarchical way: first proposing a rough trajectory through general priors and then refining the solution by specific models.

Task 3.3: Different motion priors will be learned for more complex actions than walking (such gestures or jumping), if the results of **Task 3.2** indicate that specific motion priors outperform the general ones.

Task 3.4: Joint learning of motion and pose priors for certain human actions. Study the benefits of a common learning of both priors, as well as study the viability of a common learning with appearance parameters.

Task 3.5: Validation & publication: Validation of the proposed methods on to the state-of-the-art databases, in comparison with current trends and similar methodologies. Final results will be summarized and published in international scientific forums.

5.3.4 Task 4: Joint 3D body pose and viewpoint estimation

The forth step of the proposed Thesis is to explore the possibilities of the joint 3D human pose and viewpoint estimation, given body pose models, priors of the camera pose and motion priors for walking action.

Task 4.1: Relax viewpoint constraints jointly learning 3D body pose and viewpoint, maintaining restrictions over fixed bone length (i.e. body pose only could be estimated for one subject). All computation is performed in 3D and fitted into 2D image evidences.

Task 4.2: Kinematic of various subjects will be learned, relaxing constraints of bones length but increasing memory storage and the search space.

Task 4.3: Validation & publication: Validation of the proposed methods on to the state-of-the-art databases, in comparison with current trends and similar methodologies. Final results will be summarized and published in international scientific forums.

5.3.5 Task 5: Feedback with activity estimation

The following task proposed in this Thesis refers to improve previous results on joint 3D pose and viewpoint estimation by a feedback with activity classification.

Task 5.1: Cluster study on pose priors, as well as motion priors, and the recursive optimization algorithm for finding the pose which maximizes the cluster, while finding the best cluster to reduce the search space.

Task 5.2: Study influences of behavior/activity on the clusters of poses and motion priors. This knowledge will be used on the selection of specific priors depending on the action is being performed.

Task 5.3: Joint 3D body pose, viewpoint and activity estimation in cluttered scenes in a full framework.

Task 5.4: Validation & publication: Validation of the proposed methods on to the state-of-the-art databases, in comparison with current trends and similar methodologies. Final results will be summarized and published in international scientific forums.

5.3.6 Task 6: Real scenarios and applications

This step consists on applying the previous research on to the field of social robotics. Applications are divided in two topics: content retrieval and Human Robot Interaction (HRI):

Task 6.1: Content retrieval of monocular video sequences given smart queries. These queries could include human poses, gestures and complex human behaviors. Video indexing can be applied in several domains, processing sequences batch. This task is oriented to video sequences where appear long term patients, with the final objective to provide statistics to clinicians about the evolution motor diseases.

Task 6.2: Monitorize patients suffering motor diseases in indoor environments. Monocular marker-less Motion Capture (MoCap) is a powerful approach to provide information to clinicians in near real-time. This task includes software optimization to process video sequences in the time required by this application.

Task 6.3: Human Robot Interaction (HRI): Optimization and adaptation of developed algorithms to be executed on a robotic platform. Human pose estimation and behavior analysis provide powerful tools to be applied in HRI.

Task 6.4: Validation & publication: Validation of the proposed methods on to the state-of-the-art databases, in comparison with current trends and similar methodologies. Final results will be summarized and published in international scientific forums.

5.3.7 Task 7: Compilation of results

The last task proposed for this Thesis refers to the elaboration of the dissertation and the preparation of the public defense.

6 Resources

The proposed research work will be developed mainly at the Technical Research Centre for Dependency Care and Autonomous Living (CETpD) from Universitat Politècnica de Catalunya, working in the framework of social robotics. Part of this research is also being done in collaboration with the BCN Perceptual Computing Lab (BCNLab) at the Universitat de Barcelona (UB) and Centre de Visió per Compuitador (CVC) at the Universitat Autònoma de Barcelona. The author is being financed by the Comissionat per a Universitats i Recerca del Departament d'Innovació, Universitats i Empresa de la Generalitat de Catalunya, through a TEM grant associated to Fundació Hospital Comarcal Sant Antoni Abat.

7 State of research

According to the planing presented in Figure 6, the current state of research can be summarized by the data collection of two datasets of 3D human motion and labeled body parts (see Figure 7), as well as the following list of publications:

1. *Biologically inspired path execution using SURF flow in robot navigation* [29]: In the field of social robotics, has been completed research on detectors and descriptors based on RGB cues, as well as using motion.

2. *Biologically Inspired Turn Control for Autonomous Mobile Robots* [30]: Complementing the previous work [29], egomotion is studied in this paper in order to compute motion information from monocular image sequences.

3. *Identificación y seguimiento de personas usando kinect por parte de un robot seguidor* [34]: Tracking of people by applying color histograms was proposed in this work to follow people with a social robot.

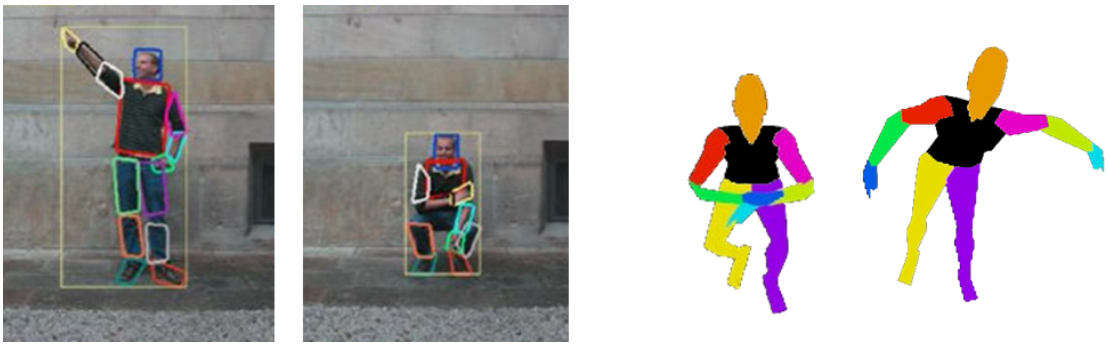

Figure 7: HuPBA datasets. Sample frames of (left) video sequences of different activities, with manually labeled body parts; and (right) 3D avatar animated with real motions captured with our framework working on Microsoft Kinect^{TM}.

4. *Uniform Sampling of Rotations for Discrete and Continuous Learning of 2D Shape Models* [55]: Different approaches to uniform samplings of rotations are reviewed in this book chapter with the final goal of building statistical models.

5. *BoVDW: Bag-of-Visual-and-Depth-Words for Gesture Recognition* [38]: A framework is proposed to classify human gestures from RGB features, as well as using motion combined with depth information.

6. *Probability-based Dynamic Time Warping for Gesture Recognition* [39]: A probabilistic approach to Dynamic Time Warping is proposed in this work in a framework of gesture classification.

7. *Survey on Spatio-Temporal View Invariant Human Pose Recovery* [14]: This paper is a survey of human pose estimation, where classical methods are reviewed and compared with current trends in this topic.

8. *Continuous Alternative to Generalized Procrustes Analysis* [74]: In this work, 2D shape models are learned from 3D objects, in order to avoid typical problems of learning certain 2D projections of 3D data.

References

[1] T. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, CVIU 104 (2) (2006) 90–126.

[2] D. Marr, L. Vaina, Representation and recognition of the movements of shapes, Proceedings of the Royal Society of London. Series B. Biological Sciences 214 (1197) (1982) 501–524.

[3] M. Eichner, M. Marin-Jimenez, A. Zisserman, V. Ferrari, Articulated human pose estimation and search in (almost) unconstrained still images, Technical Report No 272 (272).

[4] A. Clapes, M. Reyes, S. Escalera, User identification and object recognition in clutter scenes based on rgb-depth analysis, in: Articulated Motion of Deformable Objects, 2012.

[5] P. Dollár, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: An evaluation of the state of the art, PAMI (99) (2011) 1–1.

[6] V. Singh, R. Nevatia, Action recognition in cluttered dynamic scenes using pose-specific part models, in: ICCV, IEEE, 2011, pp. 113–120.

[7] E. Seemann, K. Nickel, R. Stiefelhagen, Head pose estimation using stereo vision for human-robot interaction, in: Automatic Face and Gesture Recognition, IEEE, 2004, pp. 626–631.

[8] K. Nickel, R. Stiefelhagen, Visual recognition of pointing gestures for human-robot interaction, Image and Vision Computing 25 (12) (2007) 1875–1884.

[9] R. Poppe, Vision-based human motion analysis: An overview, CVIU 108 (1-2) (2007) 4–18.

[10] S. Escalera, Human behavior analysis from depth maps, in: AMDO, 2012.

[11] M. Andriluka, S. Roth, B. Schiele, Monocular 3d pose estimation and tracking by detection, in: CVPR, IEEE, 2010, pp. 623–630.

[12] M. Enzweiler, D. Gavrila, Monocular pedestrian detection: Survey and experiments, PAMI 31 (12) (2009) 2179–2195.

[13] D. Gerónimo, A. López, A. Sappa, Computer vision approaches to pedestrian detection: visible spectrum survey, PAMI (2007) 547–554.

[14] X. Perez-Sala, L. Igual, S. Escalera, C. Angulo, Survey on spatio-temporal view invariant human pose recovery, in: CCIA, (under review).

[15] D. Gavrila, The visual analysis of human movement: A survey, CVIU 73 (1) (1999) 82–98.

[16] A. Agarwal, B. Triggs, Recovering 3d human pose from monocular images, PAMI 28 (1) (2006) 44–58.

[17] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: CVPR, Vol. 1, 2005, pp. 886–893.

[18] H. Pirsiavash, D. Ramanan, Steerable part models, in: CVPR, IEEE, 2012.

[19] A. Hernández-Vela, N. Zlateva, A. Marinov, M. Reyes, P. Radeva, D. Dimov, S. Escalera, Graph cuts optimization for multi-limb human segmentation in depth maps, in: CVPR, IEEE, 2012.

[20] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, PAMI 32 (9) (2010) 1627–1645.

[21] L. D. Bourdev, J. Malik, Poselets: Body part detectors trained using 3d human pose annotations, in: ICCV, 2009, pp. 1365–1372.

[22] C. Rother, V. Kolmogorov, A. Blake, "grabcut": interactive foreground extraction using iterated graph cuts, in: ACM SIGGRAPH 2004 Papers.

[23] D. Lowe, Distinctive image features from scale-invariant keypoints, IJCV 60 (2) (2004) 91–110.

[24] J. Daugman, et al., Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters, Optical Society of America, Journal, A: Optics and Image Science 2 (1985) 1160–1169.

[25] M. Andriluka, S. Roth, B. Schiele, Pictorial structures revisited: People detection and articulated pose estimation, in: CVPR, IEEE, 2009, pp. 1014–1021.

[26] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in: CVPR, IEEE, 2012.

[27] Y. Yang, D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, in: CVPR, IEEE, 2011, pp. 1385–1392.

[28] J. Barron, D. Fleet, S. Beauchemin, Performance of optical flow techniques, IJCV 12 (1) (1994) 43–77.

[29] X. Perez-Sala, C. Angulo, S. Escalera, Biologically inspired path execution using surf flow in robot navigation, in: Advances in Computational Intelligence, Springer, 2011, pp. 581–588.

[30] X. Perez-Sala, C. Angulo Bahón, S. Escalera, Biologically inspired turn control for autonomous mobile robots, in: CCIA, IOS Press, 2011.

[31] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: CVPR, IEEE, 2008, pp. 1–8.

[32] I. Laptev, On space-time interest points, in: IJCV, Vol. 64, 2005, pp. 107–123.

[33] R. Navarathna, S. Sridharan, S. Lucey, Fourier active appearance models, in: ICCV, IEEE, 2011, pp. 1919–1926.

[34] O. Franco Genís, X. Perez-Sala, C. Angulo Bahón, Identificación y seguimiento de personas usando kinect por parte de un robot seguidor, in: JARCA, Universidad de Sevilla, 2011.

[35] N. Pugeault, R. Bowden, Spelling it out: Real-time asl fingerspelling recognition, in: ICCV, 2011.

[36] C. Plagemann, V. Ganapathi, D. Koller, S. Thrun, Real-time identification and localization of body parts from depth images, in: ICCV, 2011, pp. 3108–3113.

[37] A. Aldoma, M. Vincze, N. Blodow, D. Gossow, S. Gedikli, R. Rusu, G. Bradski, Cad-model recognition and 6dof pose estimation using 3d cues, in: ICCV Workshops, IEEE, 2011, pp. 585–592.

[38] A. Hernandez-Vela, M. Bautista, X. Perez-Sala, V. Ponce, X. Baró, O. Pujol, C. Angulo, S. Escalera, Bovdw: Bag-of-visual-and-depth-words for gesture recognition, in: ICPR, (under review).

[39] M. Bautista, A. Hernandez-Vela, V. Ponce, X. Perez-Sala, X. Baró, O. Pujol, C. Angulo, S. Escalera, Probability-based dynamic time warping for gesture recognition, in: ICPR, (under review).

[40] Point cloud library (pcl).
URL http://pointclouds.org/

[41] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. V. Gool, A comparison of affine region detectors, Int. J. Comput. Vision 65 (1-2) (2005) 43–72.

[42] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, IEEE Trans. Pattern Anal. Mach. Intell. 27 (10) (2005) 1615–1630.

[43] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, Real-time human pose recognition in parts from single depth images, 2011.

[44] A. Hernández, M. Reyes, S. Escalera, P. Radeva, Spatio-temporal grabcut human segmentation for face and pose recovery, in: CCVPR Workshops, IEEE, 2010, pp. 33–40.

[45] B. Yao, L. Fei-fei, L.: Grouplet: A structured image representation for recognizing human and object interactions, 2010.

[46] I. Karaulova, P. Hall, A. Marshall, A hierarchical model of dynamics for tracking people with a single video camera, in: British Machine Vision Conference, Vol. 1, 2000, pp. 352–361.

[47] S. Savarese, L. Fei-Fei, 3d generic object categorization, localization and pose estimation, in: ICCV, IEEE, 2007, pp. 1–8.

[48] M. Sun, H. Su, S. Savarese, L. Fei-Fei, A multi-view probabilistic model for 3d object classes, in: CVPR, IEEE, 2009, pp. 1247–1254.

[49] H. Su, M. Sun, L. Fei-Fei, S. Savarese, Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories, in: ICCV, IEEE, 2009, pp. 213–220.

[50] C. Gu, X. Ren, Discriminative mixture-of-templates for viewpoint classification, Computer Vision–ECCV 2010 (2010) 408–421.

[51] M. Salzmann, F. Moreno-Noguer, V. Lepetit, P. Fua, Closed-form solution to non-rigid 3d surface registration, in: ECCV, 2008, pp. 581–594.

[52] M. Salzmann, R. Urtasun, Combining discriminative and generative methods for 3d deformable surface and articulated pose reconstruction, in: CVPR, IEEE, 2010, pp. 647–654.

[53] F. Moreno-Noguer, V. Lepetit, P. Fua, Pose priors for simultaneously solving alignment and correspondence, in: ECCV, Springer-Verlag, 2008, pp. 405–418.

[54] J. Sánchez-Riera, J. Ostlund, P. Fua, F. Moreno-Noguer, Simultaneous pose, correspondence and non-rigid shape, in: CVPR, IEEE, 2010, pp. 1189–1196.

[55] X. Perez-Sala, L. Igual, S. Escalera, C. Angulo, Uniform Sampling of Rotations for Discrete and Continuous Learning of 2D Shape Models, IGI Global, Spain, 2012.

[56] F. Moreno-Noguer, J. Porta, Probabilistic simultaneous pose and non-rigid shape recovery, in: CVPR, IEEE, 2011, pp. 1289–1296.

[57] G. Fanelli, J. Gall, L. V. Gool, Real time head pose estimation with random regression forests, in: CVPR, 2011, pp. 617–624.

[58] G. Fanelli, T. Weise, J. Gall, L. V. Gool, Real time head pose estimation from consumer depth cameras, in: DAGM, 2011.

[59] M. Fischler, R. Elschlager, The representation and matching of pictorial structures, Computers, Transactions on 100 (1) (1973) 67–92.

[60] P. Felzenszwalb, D. Huttenlocher, Pictorial structures for object recognition, IJCV 61 (1) (2005) 55–79.

[61] M. Andriluka, S. Roth, B. Schiele, Discriminative appearance models for pictorial structures, IJCV (2011) 1–22.

[62] P. Felzenszwalb, D. McAllester, Object detection grammars, Tech. rep., University of Chicago, computer Science TR (2010).

[63] R. Girshick, P. Felzenszwalb, D. McAllester, Object detection with grammar models, PAMI 33 (12).

[64] Y. Wang, D. Tran, Z. Liao, Learning hierarchical poselets for human parsing, in: CVPR, IEEE, 2011, pp. 1705–1712.

[65] L. Sigal, S. Bhatia, S. Roth, M. Black, M. Isard, Tracking loose-limbed people, in: CVPR, Vol. 1, IEEE, 2004, pp. I–421.

[66] L. Sigal, M. Black, Predicting 3d people from 2d pictures, Articulated Motion and Deformable Objects (2006) 185–195.

[67] L. Zhu, Y. Chen, Y. Lu, C. Lin, A. Yuille, Max margin and/or graph learning for parsing the human body, in: CVPR, IEEE, 2008, pp. 1–8.

[68] Y. Chen, L. Zhu, C. Lin, A. Yuille, H. Zhang, Rapid inference on a novel and/or graph for object detection, segmentation and parsing, Advances in Neural Information Processing Systems 20 (2007) 289–296.

[69] M. Bergtholdt, J. Kappes, S. Schmidt, C. Schnörr, A study of parts-based object class detection using complete graphs, IJCV 87 (1) (2010) 93–117.

[70] X. Lan, D. Huttenlocher, Beyond trees: Common-factor models for 2d human pose recovery, in: ICCV, Vol. 1, IEEE, 2005, pp. 470–477.

[71] V. Singh, R. Nevatia, C. Huang, Efficient inference with multiple heterogeneous part detectors for human pose estimation, ECCV (2010) 314–327.

[72] T. Cootes, C. Taylor, D. Cooper, J. Graham, et al., Active shape models-their training and application, CVIU 61 (1) (1995) 38–59.

[73] T. Cootes, G. Edwards, C. Taylor, Active appearance models, PAMI 23 (6) (2001) 681–685.

[74] L. Igual, X. Perez-Sala, S. Escalera, C. Angulo, F. Dela Torre, Continuous alternative to generalized procrustes analysis, PAA (under review).

[75] D. Kim, J. Paik, Gait recognition using active shape model and motion prediction, Computer Vision, IET 4 (1) (2010) 25–36.

[76] A. Agarwal, B. Triggs, Tracking articulated motion with piecewise learned dynamical models, in: ECCV, Vol. 3, 2004, pp. 54–65.

[77] C. Sminchisescu, B. Triggs, Kinematic jump processes for monocular 3d human tracking, in: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, Vol. 1, IEEE, 2003, pp. I–69.

[78] X. Wei, J. Chai, Modeling 3d human poses from uncalibrated monocular images, in: ICCV, IEEE, 2009, pp. 1873–1880.

[79] J. Valmadre, S. Lucey, Deterministic 3d human pose estimation using rigid structure, Computer Vision–ECCV 2010 (2010) 467–480.

[80] C. Bregler, J. Malik, K. Pullen, Twist based acquisition and tracking of animal and human kinematics, International Journal of Computer Vision 56 (3) (2004) 179–194.

[81] N. Howe, M. Leventon, W. Freeman, Bayesian reconstruction of 3d human motion from single-camera video, in: Neural Information Processing Systems, Vol. 1999, Cambridge, MA, 1999, p. 1.

[82] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, H. Seidel, Motion capture using joint skeleton tracking and surface estimation, in: CVPR, Ieee, 2009, pp. 1746–1753.

[83] B. Allen, B. Curless, Z. Popović, The space of human body shapes: reconstruction and parameterization from range scans, in: Transactions on Graphics (TOG), Vol. 22, ACM, 2003, pp. 587–594.

[84] P. Guan, A. Weiss, A. Balan, M. Black, Estimating human shape and pose from a single image, in: ICCV, IEEE, 2009, pp. 1381–1388.

[85] R. Urtasun, P. Fua, 3d human body tracking using deterministic temporal motion models, ECCV (2004) 92–106.

[86] R. Urtasun, D. Fleet, P. Fua, Monocular 3d tracking of the golf swing, in: CVPR, Vol. 2, IEEE, 2005, pp. 932–938.

[87] R. Urtasun, D. Fleet, A. Hertzmann, P. Fua, Priors for people tracking from small training sets, in: ICCV, Vol. 1, IEEE, 2005, pp. 403–410.

[88] I. Akhter, Y. Sheikh, S. Khan, T. Kanade, et al., Nonrigid structure from motion in trajectory space, in: Neural Information Processing Systems, 2008, pp. 41–48.

[89] H. Park, T. Shiratori, I. Matthews, Y. Sheikh, 3d reconstruction of a moving point from a series of 2d projections, Computer Vision–ECCV 2010 (2010) 158–171.

[90] H. Park, Y. Sheikh, 3d reconstruction of a smooth articulated trajectory from a monocular image sequence, in: ICCV, IEEE, 2011, pp. 201–208.

[91] L. Sigal, M. Black, Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion, Tech. rep., Brown Univertsity, brown Univertsity TR (2006).

[92] B. Yao, L. Fei-Fei, Modeling mutual context of object and human pose in human-object interaction activities, in: CVPR, IEEE, 2010, pp. 17–24.

[93] A. Elgammal, C. Lee, Inferring 3d body pose from silhouettes using activity manifold learning, in: CVPR, Vol. 2, IEEE, 2004, pp. II–681.