Survey on 2D and 3D 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 Vilanova i la Geltru, 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 model based methods and discuss their main advantages and drawbacks.

Keywords. Human Pose Recovery, Model Based methods, Computer Vision.

Introduction

Human body pose recovery, or pose recovery in short, refers to the process of estimating the configuration of the underlying kinematic structure of a person, which is the case of 3D pose recovery, or the 2D projection of the skeletal articulation into the image evidences. Vision-based approaches are often used to provide such a solution, using cameras as sensor inputs. Pose recovery is an important issue for many computer vision applications, including 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 order to update recent advances in the human pose recovery field, a general and standard taxonomy to classify model based approaches of the State-of-the-Art is provided. The proposed taxonomy is composed by five main modules: appearance, viewpoint, spatial relations, temporal consistence, and behavior. Since this survey analyzes computer vision approaches for human pose recovery, image evidences

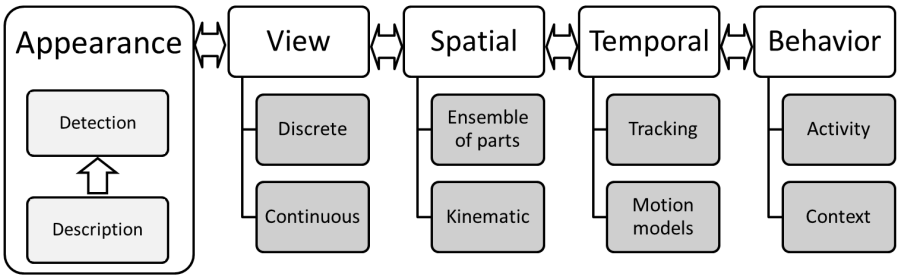

Figure 1. Taxonomy of Human Pose Recovery approaches.

should be interpreted and related to some previous knowledge of the body 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 pose estimation approaches 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 are 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, on the one hand, to those methods that take into account particular activities or the information about scene to provide a feedback to the previous modules, improving the final pose recognition. On the other hand, several works implicitly take into account the behavior by the election of datasets containing certain activities. The global taxonomy used in the rest of the paper is illustrated in Figure 1.

The rest of the paper is organized as follows: Section 1 reviews the State-of-the-Art methods grouped in the proposed taxonomy, and Section 2 discuss advantages and drawbacks of State-of-the-Art methods.

1. State of the Art

In the next subsections, the State-of-the-Art related to human pose recovery is reviewed and model based works are classified in the taxonomy defined in Figure 1.

1.1. Appearance

In order to obtain an accurate detection and tracking of the human body, prior knowledge of pose and appearance is required. This previously known information about the human body can be codified in two sequential stages: *description* of the image and *detection* of the human body (or parts), usually applying a previous learning process.. The entire procedure from image description to the detection of certain regions can be performed at three different levels: pixel, local and global. Respectively, they lead to image segmentation [30,15,16], detection of body parts [25,3,41] and full body location [9,6]. It is widely accepted that describing

the human body as an ensemble of parts improves the recognition of human body in complex poses, despite of an increasing of computational time. By contrast, global descriptors are successfully used in the human detection field, allowing the fast detection of certain poses (ex. pedestrians), as well as they serve as initialization in human pose recovery approaches. The sub-taxonomies for both *detection* and *description* stages are detailed below:

1.1.1. Detection

Detection stage refers to these specific image detections or output of classifiers which codify the human information in images. This synthesis process can be performed in four general areas summarized below.

Discriminative classifiers A common technique used for detecting people in images consists on describing image regions using standard descriptors (i.e. Histogram of Oriented Gradients (HOG) [9]) and training a discriminative classifier (e.g. Support Vector Machines) as a global descriptor of human body [9] or as a multi-part description and learning parts [10]. 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* [6].

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 by Rother, Kolmogorov and Blake [26] learns a color model from an initial evidence of a person, as well as background objects, to optimize a probabilistic functional using Graph Cuts.

Templates Example-based methods for human pose estimation have been proposed to compare the observed image with a database of samples [4]. 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 may increase the number of false positive detections when more template variations are allowed.

Interest points Salient points or parts in the images can also be used to compute the pose or the behavior is being carried out in a video sequence [18]. In this sense, we refer the reader to [20] for a fair list of region detectors.

1.1.2. Description

Detection stage analyses information extracted from the images in the description phase [19]. The most common methods applied for describing image cues are detailed below.

Silhouettes and contours Silhouettes, as well as edges and contours, are used to fit human body in images [2] because the most of the body pose information remains in its silhouette. However, methods using contours rely on a background subtraction stage because of the difficulty of extracting human silhouettes in complex scenarios.

Motion Optical flow [5] is the most common feature used to model path motion. Additionally, other works track visual descriptors and codify the motion provided by certain visual regions as an additional local cue [18].

Color and texture On the one hand, color and texture information by themselves are usually codified by means of space-color histograms or Gabor filters, respectively. On the other hand, gradients on image intensities are the most widely applied features for describing the appearance of a person. In this sense, HOG and SIFT, among others, use to be considered [9].

Depth Recently, depth cues have been included in several human pose recognition systems because of the depth maps provided by the multi-sensor Kinect^{TM}. This new depth representation offers near 3D information from a cheap sensor synchronized with RGB data. Novel depth and multi-modal descriptors have been proposed based on this representation [24,23,8]. These approaches compute fast and discriminative descriptions by detecting extrema of geodesic maps and compute histograms of normal vectors distribution. However, they require an specific image cue, and depth maps are not always available.

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 by Yao and Fei-Fei [43], where local features are codified using logical operators, allowing an intuitive and discriminative description of image (or region) context.

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 [1]), but also allows to significantly reduce the ambiguities in 3D body pose [4]. On the other hand, viewpoint can be implicitly taken into account by restrictions in the system or the election of a certain dataset. Many woks can be found in face, upper body pose estimation and pedestrian detection literature, where only front or side views are respectively studied. Just to say an example, while the detector proposed in [3] 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 [17].

Research where 3D viewpoint is estimated is divided in discrete classification and continuous viewpoint estimation (Figure 1). 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 [29,36] or unknown [35] views. 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 the work by Andriluka, Roth and Schiele [4], a discrete viewpoint

[1] 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.

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 by the projection of 3D exemplars onto 2D body parts detections. 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 reconstruction [27] can be seen as a similar problem to body pose estimation, since points in the deformable shape could be seen as body joints [33]. Given still images, the simultaneous continuous camera pose and shape estimation is studied for rigid surfaces [21], as well as for deformable shapes [28]. In both works, prior knowledge of the camera is provided by modeling the possible camera poses as a Gaussian Mixture Model (GMM).

1.3. Spatial Models

Spatial models encode the structure of the human body. Though mapping between image evidences and body pose exists [2], 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. ensembles of parts, grammars). On the one hand, accurate kinematic constraints are usually modeled by 3D skeletons. On the other hand, the most of the degenerate projections of the human body in the image plane are modeled by ensembles of parts.

1.3.1. Ensembles of parts

Ensembles 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 [13] are generative 2D assemblies of parts, where each part is detected with its specific detector (shown in Figure 2(a)). Pictorial structures are a general framework for object detection widely used for people detection and human pose estimation [11,3].

Grammar models formalized in [12] provide a flexible and elegant framework for detecting objects [10], also applied for human detection in [10,14]. 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. From a theoretical point of view, deformation rules leads to hierarchical deformations, allowing the relative movement of parts at each level; however, deformation rules in [10] are treated as pictorial structures (shown in Figure 2(b)). Which makes grammars attractive is their structural variability while dealing with occlusions.

Ensembles 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 in [31], where temporal evolution is also taken into account (shown in Figure 2(d)). Joints are modeled following Mixture of Gaussian distributions, however it is named "loose-limbed" model because of the loosely attachment between limbs.

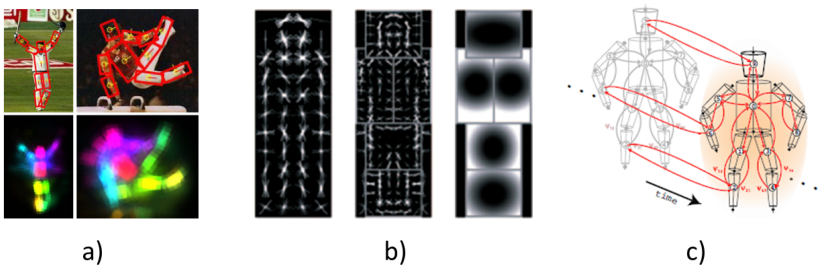

Figure 2. Examples of body models as a ensembles of parts: a) Pictorial structures [3]; b) Human model proposed in [10]: root filter (left), filters with higher resolution (middle), and model for spatial locations of parts (right); and c) Spatio-temporal loopy graph [31].

A powerful and relatively unexplored graphical representation for human 2D pose estimation are AND-OR graphs [45], 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 [7]. Much research has been done in the graph inference area, optimizing algorithms to avoid local minima. Multi-view trees represent an alternative because a global optimum can be found using dynamic programming, hard pose priors, or branch and bound algorithms [42].

1.3.2. 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 the 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 ensembles of parts are more frequently considered in 2D, accurate measures of kinematic models are more appropriate for in a 3D representation. However, the use of 2D kinematic models is reasonably useful for motions parallel to the image plane (e.g. gait analysis [17]). 2D pose is also estimated in [1] with a degenerate 2D model learned from image projections. In this case, not only parallel movements are allowed, hence, different movements are interpreted when walking in opposite directions.

3D recovery of human pose from monocular images is a the most challenging situation in human pose estimation [40]. 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. Moreover, the number of possible poses is huge, even for a model with few DOF and a discrete parameter space. Because of this reason, kinematic constraints such as joint angle limits are typically applied over kinematic models. Other solutions rely on reducing the dimensionality applying unsupervised techniques as Principal Component Analysis (PCA) over the possible 3D poses [33]. The continuous state space is clustered in [1], 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 [17], where it is used a Hierarchical PCA depending on human pose, modeling the whole body as well as body parts separately.

1.4. Temporal Models

In order to reduce the search space, temporal consistence is studied when a video sequence is available. Motion of body parts may be incorporated to refine the body pose or to analyze the behavior that is being performed.

1.4.1. Tracking

Tracking is applied 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 multiple hypothesis that can be propagated in time.

Single tracking is applied in [17], 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. Though tracking is performed in 2D, they do not loose generality at these stage since they work with movements parallel to the image plane. In contrast, 3D tracking with multiple hypotheses is computed in [4], leading to a more accurate and consistent 3D body pose estimation. In the topic of shape recovery, a probabilistic formulation is presented in [22] 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 [1].

Motion models are introduced in [39], combined with body models of walking and running sequences. A reduction of dimensionality is performed by applying Principal component analysis (PCA) over sequences of joint angles from different examples, obtaining an accurate tracking. This work is extended in [37] for golf swings from monocular images in a semi-automatic framework. Scaled Gaussian Process Latent Variable Models can also represent more different human motions [38] for concrete actions, such as walking and golf-swings, from monocular image sequences, despite of imposing hard priors on pose and motion.

1.5. Behavior

The block of behavior in our taxonomy refers to those methods that take into account particular activities or information about scene and context, to provide

a feedback to previous pose recognition modules, improving the final recognition task. Most approaches previously described do not directly include this kind of information. However, databases are usually organized by actions (e.g. walking, jogging, boxing [32]) and algorithms use to over-fit these actions (e.g. walking [4], golf swings [37]).In this sense,, 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 behavior analysis is not usual in the State-of-the-Art of pose estimation, some works exist taking into account behavior or activity to estimate an accurate body pose, learning different models depending on the action that is being performed. Different subspaces are computed for each action in [1]. Some works in the literature go a step forward and jointly recover pose and behavior. In the work by Yao and Fei-Fei [44], the authors include context information about human activity and its interaction with objects to improve final pose estimation of subjects and activity recognition. It was demonstrated that ambiguities among classes are better discriminated, and better results are obtained. The work by Singh and Nevatia [34] takes profit from such joint estimation of human pose and action being performed. A set of key poses is learned for each action and the 3D pose is accurately recovered using the specific model of such action. However, though a joint approach for pose tracking and action recognition in is presented in [34], they do not consider any feedback between both estimations.

2. Discussion and conclusion

In this survey, past and current trends in the field of human pose recovery are reviewed. Moreover, a new taxonomy is defined and State-of-the-Art methods are classified in appearance, viewpoint, spatial relations, temporal consistence, and behavior modules. We reviewed the State-of-the-Art descriptors and detectors for full body, body parts, and pixel-level codification of human information. It is widely accepted that describing the human body as an ensemble of parts improves recognition of human body parsing approaches, as well as the descriptors with the best performance in the State-of-the-Art are based on HOG filters. We showed that main methods for viewpoint analyses can be split in discrete and continuous domains. Spatial models were reviewed and divided into ensembles of parts and kinematic models depending on their flexibility. Ensembles of parts approaches 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 huge amount of body movements, combined with viewpoint and projection. Kinematic approaches 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. When the action performed in a video sequence is known, strong motion priors help in the pose estimation problem, specially in the challenging case of monocular video sequences,

reducing the search space despite of limiting the possible movements that can be detected. 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 pose estimation.

3. Acknowledgments

This work is partly supported by the Spanish Ministry of Science and Innovation (projects TIN2009-14404-C02 and TIN2011-28854-C03-01) and the Comissionat per a Universitats i Recerca del Departament dInnovaci, Universitats i Empresa de la Generalitat de Catalunya.

References

[1] Agarwal, A., Triggs, B.: Tracking articulated motion with piecewise learned dynamical models. In: ECCV, vol. 3, pp. 54–65 (2004)
[2] Agarwal, A., Triggs, B.: Recovering 3d human pose from monocular images. PAMI **28**(1), 44–58 (2006)
[3] Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR, pp. 1014–1021. IEEE (2009)
[4] Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: CVPR, pp. 623–630. IEEE (2010)
[5] Barron, J., Fleet, D., Beauchemin, S.: Performance of optical flow techniques. IJCV **12**(1), 43–77 (1994)
[6] Bourdev, L.D., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: ICCV, pp. 1365–1372 (2009)
[7] Chen, Y., Zhu, L., Lin, C., Yuille, A., Zhang, H.: Rapid inference on a novel and/or graph for object detection, segmentation and parsing. ANIRS **20**, 289–296 (2007)
[8] Clapes, A., Reyes, M., Escalera, S.: User identification and object recognition in clutter scenes based on rgb-depth analysis. In: AMDO (2012)
[9] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, vol. 1, pp. 886–893 (2005)
[10] Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI **32**(9), 1627–1645 (2010)
[11] Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. IJCV **61**(1), 55–79 (2005)
[12] Felzenszwalb, P., McAllester, D.: Object detection grammars. Tech. rep., University of Chicago (2010). TR
[13] Fischler, M., Elschlager, R.: The representation and matching of pictorial structures. Computers, IEEE Transactions on **100**(1), 67–92 (1973)
[14] Girshick, R., Felzenszwalb, P., McAllester, D.: Object detection with grammar models. PAMI **33**(12) (2011)
[15] Hernández, A., Reyes, M., Escalera, S., Radeva, P.: Spatio-temporal grabcut human segmentation for face and pose recovery. In: CVPRW, pp. 33–40. IEEE (2010)
[16] Hernández-Vela, A., Zlateva, N., Marinov, A., Reyes, M., Radeva, P., Dimov, D., Escalera, S.: Graph cuts optimization for multi-limb human segmentation in depth maps. In: CVPR. IEEE (2012)
[17] Karaulova, I., Hall, P., Marshall, A.: A hierarchical model of dynamics for tracking people with a single video camera. In: British Machine Vision Conference, vol. 1, pp. 352–361 (2000)

[18] Laptev, I.: On space-time interest points. In: IJCV, vol. 64, pp. 107–123 (2005)
[19] Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. PAMI **27**(10), 1615–1630 (2005)
[20] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. IJCV **65**(1-2), 43–72 (2005)
[21] Moreno-Noguer, F., Lepetit, V., Fua, P.: Pose priors for simultaneously solving alignment and correspondence. In: ECCV, pp. 405–418. Springer-Verlag (2008)
[22] Moreno-Noguer, F., Porta, J.: Probabilistic simultaneous pose and non-rigid shape recovery. In: CVPR, pp. 1289–1296. IEEE (2011)
[23] Plagemann, C., Ganapathi, V., Koller, D., Thrun, S.: Real-time identification and localization of body parts from depth images. In: ICCV, pp. 3108–3113 (2011)
[24] Pugeault, N., Bowden, R.: Spelling it out: Real-time asl fingerspelling recognition. In: ICCV, pp. 1114–1119 (2011)
[25] Ramanan, D.: Learning to parse images of articulated bodies. p. 1129 (2007)
[26] Rother, C., Kolmogorov, V., Blake, A.: "grabcut": interactive foreground extraction using iterated graph cuts. In: ACM SIGGRAPH, pp. 309–314 (2004)
[27] Salzmann, M., Moreno-Noguer, F., Lepetit, V., Fua, P.: Closed-form solution to non-rigid 3d surface registration. In: ECCV, pp. 581–594 (2008)
[28] Sánchez-Riera, J., Ostlund, J., Fua, P., Moreno-Noguer, F.: Simultaneous pose, correspondence and non-rigid shape. In: CVPR, pp. 1189–1196. IEEE (2010)
[29] Savarese, S., Fei-Fei, L.: 3d generic object categorization, localization and pose estimation. In: ICCV, pp. 1–8. IEEE (2007)
[30] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M.: Real-time human pose recognition in parts from single depth images. In: CVPR, vol. 2, p. 7 (2011)
[31] Sigal, L., Bhatia, S., Roth, S., Black, M., Isard, M.: Tracking loose-limbed people. In: CVPR, vol. 1, pp. I–421. IEEE (2004)
[32] Sigal, L., Black, M.: Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Tech. rep., Brown Univertsity (2006). TR
[33] Simo-Serra, E., Ramisa, A., Alenya, G., Torras, C., Moreno-Noguer, F.: Single image 3d human pose estimation from noisy observations. In: CVPR. IEEE (2012)
[34] Singh, V., Nevatia, R.: Action recognition in cluttered dynamic scenes using pose-specific part models. In: ICCV, pp. 113–120. IEEE (2011)
[35] Su, H., Sun, M., Fei-Fei, L., Savarese, S.: Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In: ICCV, pp. 213–220. IEEE (2009)
[36] Sun, M., Su, H., Savarese, S., Fei-Fei, L.: A multi-view probabilistic model for 3d object classes. In: CVPR, pp. 1247–1254. IEEE (2009)
[37] Urtasun, R., Fleet, D., Fua, P.: Monocular 3d tracking of the golf swing. In: CVPR, vol. 2, pp. 932–938. IEEE (2005)
[38] Urtasun, R., Fleet, D., Hertzmann, A., Fua, P.: Priors for people tracking from small training sets. In: ICCV, vol. 1, pp. 403–410. IEEE (2005)
[39] Urtasun, R., Fua, P.: 3d human body tracking using deterministic temporal motion models. ECCV pp. 92–106 (2004)
[40] Valmadre, J., Lucey, S.: Deterministic 3d human pose estimation using rigid structure. ECCV pp. 467–480 (2010)
[41] Wang, Y., Tran, D., Liao, Z.: Learning hierarchical poselets for human parsing. In: CVPR, pp. 1705–1712. IEEE (2011)
[42] Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR, pp. 1385–1392. IEEE (2011)
[43] Yao, B., Fei-fei, L.: Grouplet: A structured image representation for recognizing human and object interactions. In: CVPR, pp. 9–16 (2010)
[44] Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR, pp. 17–24. IEEE (2010)
[45] Zhu, L., Chen, Y., Lu, Y., Lin, C., Yuille, A.: Max margin and/or graph learning for parsing the human body. In: CVPR, pp. 1–8. IEEE (2008)