J. Vis. Commun. Image R. 32 (2015) 10–19

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

A survey of human pose estimation: The body parts parsing based methods ☆

Zhao Liu *, Jianke Zhu, Jiajun Bu, Chun Chen

College of Computer Science, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 15 May 2014
Accepted 22 June 2015
Available online 2 July 2015

Keywords:
Human pose estimation
Articulated object detection
Survey
Body parts parsing
Motion capture
Feature Extraction
Appearance models
Structure models

ABSTRACT

Estimating human pose from videos and image sequences is not only an important computer vision problem, but also plays very critical role in many real-world applications. Main challenges for human pose estimation are variation of body poses, complicated background and depth ambiguities. To solve these problems, considerable research efforts have been devoted to the related fields. In this survey, we focus our attention on the recent advances in vision-based human pose estimation. We first present a general framework of human pose estimation, and then go through the latest technical progress on each stage. Finally, we discuss the limitations of the existing approaches and foresee the future directions to be explored.

1. Introduction

Human pose estimation (HPE) is the process of inferring the 2D or 3D human body part positions from still images or videos. Conventional HPE methods usually employ extra hardware devices to capture human poses and construct a human skeleton based on the captured body joints. These methods are either expensive or inefficient. During the past decade, considerable research efforts have been devoted to HPE problem in computer vision domain.

Although having investigated the issues of human body part configuration, human body detection and human motion [1] in the previous studies, there still lacks a survey to summarize the most recent progress on body pose estimation. In this survey, we mainly review the recent advances in vision-based human pose estimation. Human pose estimation includes nearly all the human-related problems in computer vision, ranging from the whole human body pose parsing to the detailed body parts localization. As it is hard to cover all these fields within a single survey, we mainly focus on the body part parsing methods. For better comparison of different body part parsing methods, we divide them into four parts, including 2D single person parsing in images, 2D multi-person parsing in images, 2D single person parsing in videos and 3D single person parsing in images and videos. Moreover, we discuss the limitations of the existing approaches and foresee the future trend.

Human pose estimation techniques become more and more mature in the past decades. Being the great interest of different domains, new applications constantly emerge along with the technological evolutions. Human pose estimation is not only an important computer vision problem, but also plays critical role in a variety of real-world applications in the following.

Video Surveillance. Video surveillance aims at tracking and monitoring the locations and motions of pedestrians in special circumstances. It is the earliest application area that HPE technologies have been used. The common scenes are the supermarket and airport passageway.

Human–Computer Interaction (HCI). Advanced human computer interaction systems with human pose estimation have been developed rapidly. In these systems, instructions can be analyzed accurately by capturing the human body poses. In recent years, intelligence driving emerges as a novel practical application.

Digital Entertainment. Digital entertainment, including computer games, computer animation and films, has become a huge industry and an active domain in recent years. For instance, People enjoys the pleasure the body sensor games give to them. Also, In the pre-production of the special effects for movie Avatar [2], actors wear the special equipments to animate the activities of Avatars.

Medical Imaging. Human pose estimation has been widely used in the automatic medical field. A specific instance is that HPE can

☆ This paper has been recommended for acceptance by M.T. Sun.

* Corresponding author.
E-mail addresses: liuzhao@zju.edu.cn (Z. Liu), jkzhu@zju.edu.cn (J. Zhu), bjj@zju.edu.cn (J. Bu), chenc@zju.edu.cn (C. Chen).

http://dx.doi.org/10.1016/j.jvcir.2015.06.013
1047-3203/

Fig. 1. Various applications fields for human pose estimation.

Table 1
Main commercial systems for human pose estimation.

Systems	Principle	Application areas	Institution	Related URLs
Kinect Sensor	Structure light capture and machine learning	Motion Capture Multi-View Pose estimation [9]	Microsoft	http://www.xbox.com/en-US/kinect
Leap Motion	Double sensors Infrared light Vision different	Gesture Recognition [10]	Leap Motion Inc.	https://www.leapmotion.com/
		Industrial Robot [11]		
Vicon	Reflected light based system	Animation, Military Remote sensing, Bioinformation	Oxford Metrics Limited	http://www.vicon.com/
Wii	Bluetooth communication Infrared light	Games Physical treatment [12]	Nintendo	http://www.nintendo.com/wii

be used to assist doctors to check patients' activities from the remote monitor, which greatly simplifies the therapeutic process.

Sports Scenes. In sports news and live broadcast, human pose estimation is employed to track athletes' locations and activities. Moreover, the estimated poses can be used to employed the detailed movements of their actions.

Other applications include military, children mental development, virtual reality, and so on. The related application fields of HPE are shown in Fig. 1.

In recent years, various devices and commercial systems have been released accompany with HPE technology, including Microsoft Kinect sensor [3,4], Leap Motion [5], body mounted camera [6], 3D laser scanner [7] and infrared light source [8]. These commercial systems have quite different implementation principles and application fields, as shown in Table 1.

2. Related surveys and overview

During the last decade, several surveys have been published to summarize the related work on human pose estimation. 3D HPE has attracted lots of attentions in computer vision. For instance, Hen and Paramesran [13] summarize the single camera 3D pose estimation from images and Sminchisescu [14] aims to reconstruct 3D human poses from monocular video sequences. Wearable equipments make it possible to estimate the depth in motion capture, Helten et al. [15] review the depth camera based motion capture work. Compared with methods relying on the specific hardwares, the vision-based approaches are more efficient and economic, which have been rapidly developed in these years. Poppe [1] reviews the vision-based methods for human pose estimation on the marker-less data. Moeslund et al. [16] summarize the research work on visual analysis of human, which covers various topics including pose estimation, human recognition and their applications.

Most recently, we witness the rapid development in HPE field. For instance, the model based methods, especially the pictorial structure model, have played an important role in human body parsing [17,18].

On the other hand, since video play an more important role in recent applications, action/activity recognition have been paid much attention, although closely related with HPE, the technique used in action/activity recognition are quite different. We refer the readers to [19–22] to see more details of action/activity recognition. Moreover, deep learning based methods have attracted lots of research attentions [23,24]. As all of these technologies will continue to be the focus for a few years in future, there is a need for the discussion in hot topics.

To clearly illustrate the recent studies on human pose estimation, we categorize them into two stages: preprocessing and body parts parsing. Fig. 2 summarizes the whole process of the common human pose estimation. The preprocessing stage includes feature extraction, camera calibration, body detection and foreground

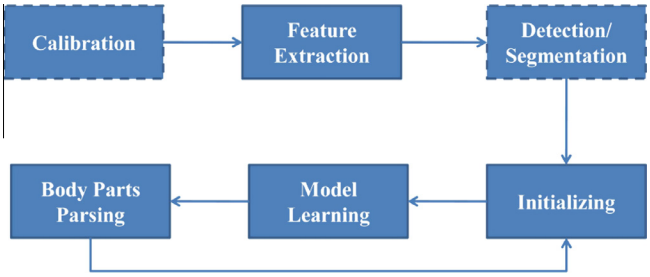

Fig. 2. The framework of the common human pose estimation system, stages in the boxes with dash lines means that they may be removed in some methods.

segmentation. According to this framework, the remainder of this survey is organized as follows. Section 3 investigates the related studies on preprocessing for HPE. Section 4 gives a comprehensive review on body parts parsing, which is the most important stage in HPE. Various datasets and evaluation methodologies for HPE are discussed in Section 5. Finally, Section 6 concludes this survey and foresee the future directions for HPE study.

3. Preprocessing work

The preprocessing stage for HPE includes camera calibration, foreground segmentation and human body detection, in this section we review the recent advance on these techniques.

3.1. Data calibration

In HPE, the human poses are always captured from different camera views and there exist deviation between the capture data, so data calibration is always take as a preprocess stage. Camera calibration is the most used methods, in [25], Stoll et al. first synchronize then calibrate the input video before establish a 2D-3D skeleton model. Also, Juergen et al. [26] use a classifier combination strategy to integrate the pose data come from multi-view cameras. In addition, active body sensors like the Kinect [9,27] are also used for calibrate the human motion. Some methods use fusion from multiple sensor terminals. For instance, Ennis et al. [28] measure the factors from different sensors in a multi-sensory conversation scenes. And Shiratori et al. [6] use the multiple body-mounted cameras to capture the motion data.

3.2. Body localization

Body localization, which contain human detection and foreground segmentation, aims at locating human beings in the complex images. In human pose estimation, various body localization methods are used as a pre-processing stage.

3.2.1. Human detection

Human detection aims at estimating the rough location and scale of people in images. Mykhaylo et al. [29] first estimate the body joint positions, and then use these positions to refine a human body detection bounding box. Eichner et al. [17] use the upper body detector to distinguish the human body from background, and then a soft labeling process is applied to further refine the result. Similar idea is taken by Sapp et al. [30], a feature based model is able to capture the joint extension in adjacent frames with the bounding box from the upper body detector. Furthermore, body detection is used in the multi-person pose estimation. Eichner and Ferrari [31] initialize the multi-person locations by an upper body detector. In this case, the detection window set is used to determine the number of persons being parsed in each image.

3.2.2. Background subtraction

Some HPE methods need the pixel-wise human segmentation as their input. Traditional segmentation methods model the foreground and background for each pixel based on different features. Kim and Grauman [32] employ the local dense regions to model the foreground, which is useful especially for capturing the human body shape. In a video sequence, the background image sometimes keeps stationary while the foreground is apt to change among the different frames. To capture the variations of foreground regions, Guillot et al. [33] try to build a background model based on a collection of registered images, which is updated with the illumination changes. Moreover, the regions of interests are selected by the keypoints unmatched to the background at each frame. Generally, dynamical cues are useful in the consequent images and video segmentation. Lee et al. [34] use the dynamic intra-frame motion cues like the surrounding difference in a space–time model to label the object-like regions. In another case, Anestis and Vittorio [35] propose a point clustering method to efficiently segment objects in the video sequences.

Table 2 gives a brief summarization of some typical methods which need the pre-processing stages we referred above. Notice that features extraction is a necessary stage for all these methods. So, we compare different feature extraction methods individually.

4. Body parts parsing

Body parts parsing aims at locating different body parts in the images, which is the most important step in human pose estimation. In this section, we review the recent technique advances in parsing human body parts.

The body parsing methods varying from 2D body parsing to 3D body parsing, and from images to videos. To make a clear illustration, we divided these methods into four subcategories, which are single person parsing in single 2D images, single person parsing in 3D images/videos, single person parsing in 2D videos and multi-person parsing in 2D images. In each subcategory, we will discuss the features, body appearance models and structure models orderly. For the appearance model, we refer to methods which parse each part of the body individually, while for the structure model, we refer to methods utilizing the relationship between different body parts.

4.1. 2D single person parsing in images

Single person parsing in 2D images aims at locating each part of a single person in the input image set, which is the fundamental body parts parsing problem and most of the proposed work are related to it.

4.1.1. Feature extraction

The features used in person parsing can be divided into the handcrafted features and the learned features according to their representation. The most used handcrafted features are the HOG [38] and shape context [29]. To obtain more robust feature representation, researchers usually combine the different local features together. For instance, Sapp and Taskar combine the shape, contour, geometric and appearance features together in their cascaded

Table 2
Different pre-processing steps in HPE system.

Pre-stage	Goal	Features used	Main methods and related work
Camera calibration	Unification for multiple views	color silhouettes	Sums of Gaussians [25] Classifier combination [26]
Foreground segmentation	Human shape estimation	boundary, hog sift, shape	Region modeling [32,33] Dual decomposition [36]
Human Detection	Human location and size estimation	Shape context color, geometry edges	Shape Estimation [37] Regression [17] Appearance transfer [31] Adaboost classification [29]

pictorial structures model [39]. In recent years, the deep learning methods are widely used in HPE for feature extraction. For instance, Toshev and Szegedy [23] use a seven-layered convolutional DNN in their body joint regressor to represent the joint context and predict the body location. Also, Chen and Yuille [40] train Deep CNNs on the image patches around the body joints to learn the probabilities for the absence and spatial relationship of different body parts. In contrast to the local features, the global features represent the human poses holistically. Ouyang et al. [24] aim at extracting high-level global features from the fusion of traditional HPE models, they first extract high-level representation from different sources and combine them in a deep model. Similarly, Tompson et al. [41] integrate the deep convolutional network with the Markov random field to get a novel part based body detector, this new architecture takes advantage of the use of multi-scale features.

4.1.2. Appearance models

Poselet [42] is a popular model for monocular HPE. Specifically, a poselet is a set of linear support vector machines, which bridges the gap between the body part appearance and configuration. Wang and Li [43] propose a part-based tree model by constructing poselet examples through a visual category distinctive tree nodes which are used to represent the human body parts. This tree model is approximate to the pictorial structure model and can be employed in the animal body parts parsing. The original poselet is limited in representing only single part of the human body. To solve this problem, Wang et al. [44,45] extend poselet to a multi-level form, which denotes each level of the extended model as a different mixture of poselet elements varying from a single body part to the whole body. Similarly, Srinivasan and Shi [46] utilize the hierarchical segmentation model for parsing different level of human body parts. In [47], Dantone et al. train a two-layer regression forest as body part regressor, where the first layer models each body part independently and the second layer models the relationship between these body parts. Both the human body and the object have a hierarchical structure and can be modeled in a whole formulation. Based on this hypothesis. Yao and Fei-Fei [48] tackle the human-object interaction through a coarse-to-fine random field model, in which the higher level represents the image information and the lower level stores the appearance of the body-object parts interaction.

4.1.3. Structure models

The body parts in an articulated structure are inherently dependent of each other. Therefore, imposing the articulated constraints on the tree model is helpful in body parts parsing. In the original tree model, each body part is represented by a tree node, and all the body parts are connected to its neighboring body parts. Ramakrishna et al. [49] build an inference machine framework to model the interactions between body joints. Tran and Forsyth [50] present a full-relational model, which includes the additional left–right symmetry body joints relationship. Instead, Karlinsky and Ullman [51] propose a new model with the linking features that replace the kinematic constraints in the original PS model. With this modification, the connectivity of the part can be achieved by retrieving using the image features. Pictorial structure model (PSM) is a special case of the tree model which was first introduced forty years ago [52] and now has become the most popular generative model in HPE. The original pictorial structures model is shown in Fig. 3. It differs from other tree-form models in that each of its nodes is modeled individually in a deformable form, and spring like connections are used to connect different parts. This special structure enables the PSM to have rich appearance variations. The PSM had not been applied to human pose estimation until investigated by Felzenszwalb and Huttenlocher [53]. Ferrari et al. [17,54] construct a spatio-temporal model that facilitates a time window to tie up the nodes for a certain body part between different frames. Yang and Ramanan [55,18] propose a mixture model to describe the body joints and their relationship, each body joint is represented as a non-oriented mixture part, and these parts are learned in a structured SVM. Sapp and Taskar [56] employ quadratic deformation cost as the geometry features in MODEC model in order to refine the binary term between neighbor parts. Recently, Kiefel and Gehler [57] modified the pictorial structure model to a more flexible form by replacing each body part with a binary random variable. Pishchulin et al. [58,59] propose such a method, where both the unary and binary terms of pictorial structure model are replaced with poselet hypotheses and the whole structure is kept fixed.

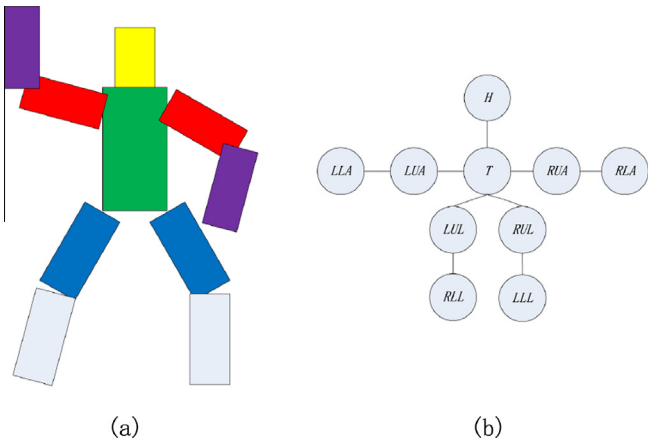

Fig. 3. The pictorial structures model [52]: (a) the tree nodes representing each body parts and (b) the connection between different body nodes.

4.2. 2D Multi-person parsing in images

Most of body part parsing methods aim at estimating the body parts of single person from images. However, the recent trend is to estimate the poses of multiple people from images.

4.2.1. Feature extraction

Local features contain the image properties for a certain patch, which are helpful especially to distinguish the target foreground from the background. Among various local features, SIFT and HOG [38] are the most frequently used in multi-person parsing.

4.2.2. Appearance models

Compared with the single person parsing, the interaction between people should be taken into consider in this cases. Sun and Savarese [60] present an articulated part-based model, in which the different body parts of multiple people are represented individually as a shape template, and the occlusion between different body part are modeled in a coarse-to-fine representation.

4.2.3. Structure models

Yang and Ramanan [61] extend their non-oriented mixture part model, they add "touch codes" to model the interactions of people, however, this new model can only be used in the two-person parsing case. Eichner and Ferrari [31] propose a more practical multi-person parsing model, they employ a front-to-back structure PSM with an extra occlusion probability prior along with the inter-person exclusion penalty to prevent two body parts from sharing a same image region. In the traditional PS model, the dependency of body parts is only related to kinematic constraints.

4.3. 2D single person parsing in videos

The problem of parsing human parts in videos is also important. Different from parsing human body parts from images, the parsing of human body parts in videos always includes temporal cues to reflect the relationship of the same body parts in adjacent frames.

4.3.1. Feature extraction

In video analysis, temporal features are widely used. Weinland et al. [62] propose the motion history volume (MHV) feature, they use the Fourier transforms in cylindrical coordinates to perform the collection process. Sapp et al. [30] use optical flow in their extended cascaded pictorial structures model to track the between-frame body joints, this work is extended in Zuffi et al. [63], they employ the dense optical flow to track the body parts in adjacent three frames.

4.3.2. Appearance models

Aim at estimating both arms in clutter scenes, Gkioxari et al. [64] extend the poselet by adding temporal information, they train a set of linear SVMs for the image patches extracted from the key points. Compared with single frame method, the tracking-based method has an extra tracking stage. Weiss and Taskar [65] use efficient value-function approximation to learn the feature extracted from individual inputs adaptively. In a recent work, Tokola el al. [66] also use a hidden Markov model to track the object hypotheses generated from each frame.

4.3.3. Structure models

Temporal information is important in video-based HPE. A typical video-based model is proposed by Sapp et al. [30], in this model, a six parts structure is built across the adjacent frames. Although being able to capture the relationship of the body joints across adjacent frames, it requires heavy manual annotations and lots of computational time to extract features. The Hidden Markov Model (HMM) is an efficient structure model in recognition, which has also been used in HPE. For example, Fathi and Mori [67] employ HMM to estimate the motion correlation in video sequence. Similarly, Navaratnam et al. [68] model hypotheses from multiple frames in a HMM, where the Viterbi algorithm is used to track the poses in different frames.

From the above all, we summarize the typical methods for 2D body parts parsing into Table 3.

4.4. 3D single person parsing in images and videos

In recent years, as 3D data have been widely used in HPE field and the hardware for capturing 3D human joints developed. 3D human pose estimation becomes a promising technique area. To get the 3D information is the key in 3D human parsing, there are two ways for capturing the 3D information: from the multi-view images/video or from the depth information in monocular image. The techniques used are quite different.

4.4.1. Feature extraction

In 3D HPE, researchers combine features different from those used in 2D HPE to reflect dimension variation. In [4], Shotton et al. employ the depth pixel difference features, which is able to effectively capture the body part information from single image. A similar work is proposed by Bo and Sminchisescu [70], where 3D poses can be estimated from 2D images through twin Gaussian process regression using the HMAX and HOG features. Sedai et al. [71] concatenate histogram of shape context and histogram of local appearance context features into a new fusion form feature. Also, combining the 2D features and 3D features is always helpful. Straka et al. [72] extract the 3D voxels from the 2D images and combine these voxels with the human silhouette images. Chen et al. [73] use the body joints and coordinates to align the symmetric body parts. Since deep features provide more information, they are recently integrated in the 3D person parsing system [74].

4.4.2. Monocular depth methods

Salzmann and Urtasun [75] perform a Gaussian process on a set of random variables to learn a model linking the 2D images and 3D joints set. For real-time parsing, Plagemann et al. [76] propose an novel method which is able to label body parts according to the interest points learned by a boosted classifier. Since the spatial ambiguities cannot be directly reduced in 3D HPE, extra constraints are always imposed. For instance, Wei and Chai [77] employ the stationary body structure constraints along with 2D–3D bone projection constraints. Additionally, Buys et al. [78] combine color and depth information to initialize the body labels for training. Several methods try to estimate the human body parts from the depth images by learning a regression forest, as in [79,80]. Temporal information is added to associate the body poses in videos. Recently, Kaliamoorthi and Ramakrishna [81] employ a kinematical model to represent the body joints relationship and use a constant position model to acquire the motion priors. Self-occlusion is a hard problem in articulated human pose estimation. To tackle this issue, Cho et al. [82] use a Markov random field to represent the occlusion relationship of different body parts. Moreover, they design an adaptive inference algorithm to estimate the 3D body poses in occlusion states.

4.4.3. Multi-view methods

In an early work, Mori and Malik [83] recover 3D poses from a sequence of 2D images, which is based on the matching of shape context and deformable models from multi-view images. Although the idea is general, the shape context feature limits its application domain. Later, Ramakrishna et al. [84] employ anthropometric regularity as the constraints in 3D human skeleton recovering, which is effective for the multi-view camera scenes. Juergen et al. [26] propose a multi-view 3D pose estimation framework

Table 3
Different methods for 2D human body parts parsing.

Methods	Features	Multi/single person	Video/ images	Main technique
Yang and Ramanan [55,18]	HOG	Single	Images	Non-maximum suppression (NMS) SVM
Mykhaylo et al. [29]	Shape Context	Single	Images	Bootstrapping Max-product Part detector
Wang and Li [43]	HOG	Single	Images	Poselet Tree structure
Eichner and Ferrari [31]	Edgelet HOG	Multi	Images	Approximate inference TRW-S Occlusion probability
Sun and Savarese [60]	HOG	Multi	Images	Hierarchical structure Structured SVM Dynamical program
Ferrari et al. [17,54]	HOG, Shapes Edges	Single	Images	Temporal association Spatio-temporal parsing PS model
Sapp et al. [30,69]	Geometry, Color Optical flow Texture	Single	Video	Coarse to fine cascade Rich temporal features Sub-gradient descent
Fathi and Mori [67]	Edges HOG	Single	Video	Hidden Markov Model

Table 4
Different methods for 3D human body parts parsing.

Methods	Images/videos	Viewing angle	Feature	Main techniques
Straka et al. [72]	Images	Monocular	Silhouette images	Voxel scooping
Rui et al. [89]	Images	Monocular	Silhouette	Bayesian formulation
Yu et al. [85]	Images	Monocular	Pose Silhouette	Two GPLVMs [86]
Salzmann and Urtasun [75]	Images	Monocular	PHOG SIFT	Gaussian process
Ramakrishna et al. [84]	Images	Monocular	Shape	Matching pursuit algorithm
Mori and Malik [83]	Images	Multi-view	Shape contex	Sequence image matching
Juergen et al. [26]	Images	Multi-view	Edge Silhouette	Manifolds optimization
Kaliamoorthi and Ramakrishna [81]	Videos	Multi-view	Silhouette	Kinematical model
Cho et al. [82]	Videos	Multi-view	Position Orientation	MRF Adaptive inference

based on a set of action-specific manifolds, in which a 2D action recognition system is served as the prior distribution for optimization on the manifolds of the human poses. Yu et al. [85] propose the method for estimating the 3D poses from the two independent SGPLVMs [86], which is able to represent the shape variables of the body poses. Most recently, Burenius et al. [87] extend the pictorial structures model to multi-view form, the authors discretize the joint angles and skeleton views in search space. Later, they propose an extension work [88] by using random forest to initialize the body joints.

In Table 4, we summarize different 3D human body parts parsing methods in recent years. It can be seen that the silhouette feature is commonly used in 3D body parts parsing. Also, the multi-view methods provide good supplementary to the traditional monocular methods.

5. Benchmark

5.1. Datasets

Due to the large variations in different scenes, it is difficult to build a universal dataset to evaluate the human pose estimation. Alternatively, researchers have created lots of datasets to evaluate their proposed techniques for the specific task, which makes the fair comparison on the different algorithms even harder.

We summarize the current publicly available datasets into Table 5. HumanEva [90] dataset is made of a number of images capturing the synchronized people performing the interactive and walking actions. These images were acquired by using a multi-view motion capture system with the high resolution video sensors. The Buffy [17] dataset is extracted from TV clips containing 5 sections, 748 annotated images in total, where six different body parts are annotated for each of the upper body skeleton. In a recent work, Andriluka et al. [91] propose the representative MPII Human Pose benchmark, which covering different human activities both for upper-body and full-body pose estimation.

Besides the above three most used benchmark, other datasets are collected for the different purposes. For instance, a portion of images containing people are selected from PASCAL VOC [92] to evaluate the human pose estimation under the relatively uncontrolled environment. Parse people is an early but still popular dataset proposed by Ramanan [93]. Human actions in sport scenes are challenging and variable that are always used as the benchmark for HPE, i.e., Leeds sports dataset [94] and UIUC stickmen dataset [50]. In contrast to these mainstream full-body dataset, Sapp et al. [56]

Table 5
Publicly available human body pose databases.

Dataset	Size	Color/gray	Type	Dim	Link
Buffy	472 frames training 276 frames testing	720 × 405 color	Upper body	2D	http://www.robots.ox.ac.uk/vgg/data/stickmen
Parse	100 images training 205 images testing	Different sizes color	Full body	2D	http://www.ics.uci.edu/dramanan/papers/parse
LSP	1000 images training 1000 testing	Different sizes color	Full body	2D	http://vision.grasp.upenn.edu/cgi-bin/index.php
FLIC	3987 images training 1016 images testing	720 × 480 color	Full body	2D	http://vision.grasp.upenn.edu/video/FLIC.zip
PASCAL VOC	47,186 images 110,008 objects	Different sizes color	Upper body	2D	http://pascallin.ecs.soton.ac.uk/challenges/VOC
MPII Human pose	410 activities 25 K images	Different sizes color	Full body	2D	http://human-pose.mpi-inf.mpg.de/
Poses in the wild	30 sequences 900 frames	Different sizes color	Upper body	2D	https://lear.inrialpes.fr/research/posesinthewild/dataset
HumanEva	50,600 frames training 26,400 frames testing	659 × 494 color 644 × 448 gray	Full body	3D	http://vision.cs.brown.edu/humaneva
PDT	40 sequences 6 performers	Different size color	Full body	3D	http://gvvperfcapeva.mpi-inf.mpg.de/public/PersonalizedDepthTracker/index.php
SMMC-10	28 sequences	176 × 144 color	Full body	3D	http://ai.stanford.edu/varung/
EVAL	3 subjects 24 sequences	Different size color	Full body	3D	http://ai.stanford.edu/varung/eccv12/

recently propose the upper body FLIC dataset, and Cherian et al. [95] propose the "Poses in the Wild" dataset which focusing on various poses in wild scenes.

By using the updated motion capture methods 3D datasets with more variations can be created, for instance, Helten et al. [96] create the PDT dataset by using a Kinect and a marked-based mocap system. Additionally, Ganapathi et al. [97,98] create the SMMC-10 and EVAL datasets, which focusing on utilizing the depth information contained in the 3D human pose data. Variation of clothes imposes considerable difficulties on HPE, to show this influence, Dantone et al. [47] collected the FashionPose dataset.

5.2. Evaluation methodology

5.2.1. 2D person parsing in images

For the task of 2D person parsing in images, including the single person parsing and multi-person parsing tasks, the most popular evaluation metric is Percentage of Correctly estimated body Parts (PCP) [17,99,31,55,50]. Specifically, PCP is proposed by Ferrari et al. [100], which evaluates the accuracy of stick predictions; Considering a predicted body part, if it overlaps the ground truth more than 0.5 of its annotated part length, this part is correctly predicted. In recent years, PCK (Percentage Correct Keypoints) is widely used to evaluate the body joints accuracy [56,18] in 2D person parsing. According to its definition, a body joint is correctly predicted if it falls within a circle with a radius of X pixels around the ground-truth body joint [30]. Besides, researchers usually employ the error evaluation to compare their approaches against the previous methods in order to show the efficiency of their approach and the improvements over previous methods [101–103].

5.2.2. 2D person parsing in videos

Most of the evaluation metrics used in images, like the PCP and PCK, can also be used in 2D person parsing in videos [54,64,66]. In addition, researchers use various methods to increase the temporal coherence. For instance, Navaratnam et al. [68] use Viterbi algorithm to perform tracking between adjacent frames, and [65] use Meta-features to reduce the bias.

5.2.3. 3D person parsing in images/videos

Many methods in 3D person paring use reconstruction error as the evaluation metric [75,77,80]. The 3D human parsing can be treated as a fine-grained object detection problem, in which the conventional precision-recall for each body part can be used as the evaluation metric [79,85]. Additionally, by mapping the predicted 3D body parts back to 2D, PCP can also be used [87].

5.3. Empirical evaluation on existing methods

We compare several human pose estimation methods in Tables 6–8. Results for 2D person parsing in images are performed on LSP and FLIC datasets, we use PCP and PCK metrics to evaluate the parsing accuracy. The images of Johnson and Everingham's are based on the Person-Centric (PC) annotations while others use the Observer-Centric (OC) annotations. Part of the results can be referred to Andriluka et al.'s work [91], and we extended the results by adding the results from Wang and Li [43], Yang and Ramanan [18], Johnson and Everingham [94]. From the results we can see that Chen and Yuille.'s method [40] performs best both on the LSP and FLIC datasets. Researchers use various datasets and evaluation metrics in 3D human pose estimation, and we conclude some representative methods in Table 8.

Finally, we make further experiments on two deep learning based human body parsing methods: Chen and Yuille [40] and Tompson et al. [41]. Both can be used to parse the upper body pose and full body pose. In our experiments, we use FLIC dataset for evaluating upper body parsing and LSP dataset for full body parsing. Figs. 4 and 5 show the visualization results on these two methods. As shown in Figs. 4 and 5, both methods perform well on most of the images, even in the cases of backward poses or poses with occlusion. Additionally, estimation errors occur when the person is hard to be identify from the background, or in upside-down pose, which is the future work to be explored.

Table 6
Comparison of different 2D person HPE methods on LSP dataset.

Methods	Tasks	Datasets	Evaluation metrics	Performance
Yang and Ramanan PAMI'13 [18]	2D single person/ images	LSP	PCP	55.1
Pishchulin et al. ICCV'13 [59]	2D single person/ images	LSP	PCP	69.2
Wang and Li CVPR'13 [43]	2D single person/ images	LSP	PCP	62.8
Ramakrishna et al. ECCV'14 [49]	2D single person/ images	LSP	PCP	67.8
Ouyang et al. CVPR'14 [24]	2D single person/ images	LSP	PCP	68.7
Chen and Yuille NIPS'14 [40]	2D single person/ images	LSP	PCP	75.0

Table 7
Comparison of different 2D HPE approaches on FLIC dataset.

Methods	Tasks	Datasets	Evaluation metrics	Performance
Yang and Ramanan PAMI'13 [18]	2D single person/ images	FLIC	PCP	65.2
Sapp and Taskar CVPR'13 [56]	2D single person/ images	FLIC	PCP	68.3
Pishchulin et al. NIPS'13 [59]	2D single person/ images	FLIC	PCP	87.3
Chen and Yuille NIPS'14 [40]	2D single person/ images	FLIC	PCP	91.9

Table 8
Comparison of different 3D HPE approaches on image/videos.

Methods	Tasks	Datasets	Evaluation metrics	Performance
Jonathan et al. CVPR'12 [80]	3D single person/ images	MSRC-5000	reconstruction error	0.038
Yu et al. ECCV'10 [85]	3D single person/ images	CAESAR database	Precision recall	0.78 0.89
Shotton et al. CVPR'11 [79]	3D single person/ images	Real depth data	mean average precision	0.914
Burenius et al. CVPR'11 [87]	3D single person/ videos	football sequence	PCP	0.63

Fig. 4. The upper body part parsing results on Chen and Yuille [40] (the first and the third rows) and Tompson et al. [41] (the second and fourth rows) with the FLIC dataset. The first and second rows show the successful results, while the third and the fourth rows show the fair cases.

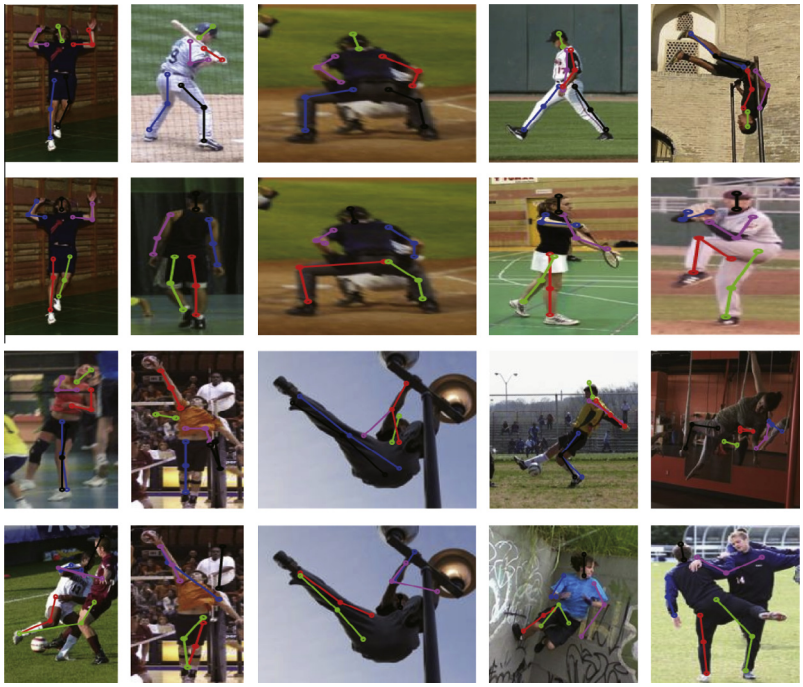

Fig. 5. The full body part parsing evaluation on Chen and Yuille [40] (the first and the third rows) and Tompson et al. [41] (the second and fourth rows) on LSP dataset.

6. Future work and conclusion

Due to challenges ranging from most of the important topics in computer vision domain, estimating human poses from images and videos is always hard. This survey summarize the recent research efforts on this problem.

However, these technologies are limited especially for the irregular poses. A future trend is to explore the unsupervised or semi-supervised learning in body parts parsing. Over-segmentation is useful to keep the contour information, which is a promising preprocessing technique.

Multi-view HPE will continue to be the focus of research in future. Transfer learning proves to be useful in cosegmentation. We are looking for the pose estimation methods based on it, especially in the application in multi-view HPE field.

The traditional human-object touching analysis methods usually parse the people and objects separately. The recent HPE method tends to study the person-object interaction [104]. The

nonrigid image descriptors can be used to determine the interactive relationship among people and objects.

Although several datasets have been built for fair evaluations on HPE, it is still desirable to collect better datasets with proper evaluation protocols. In the future, more body sensors should be employed to capture the raw data with various postures. Semantic information can be added as extra feature to derive more applications. Also, it is important to develop new techniques to efficiently annotate the human body parts with the proper representation.

Up to now, enormous research efforts have been spent on the task of estimating human poses from images or videos, yet there is still a large gap between the theoretical research and real-world applications. In this survey, we focus mainly on the recent progress for parsing human body parts by using vision-based methods. For the future work, we will investigate the effective methods to deal with the large variations on body poses and shapes.

Acknowledgments

The authors appreciate the reviewers for their extensive and informative comments for the improvement of this manuscript. This work was supported in part by National Natural Science Foundation of China under the Grant (61103105), National High Technology Research and Development Program of China (2013AA040601).

References

[1] R. Poppe, Vision-based human motion analysis: an overview, Comput. Vis. Image Underst. 108 (2007) 4–18.
[2] J. Cameron, Avatar. <http://www.avatarmovie.com/index.html>.
[3] J. Han, L. Shao, D. Xu, J. Shotton, Enhanced computer vision with microsoft kinect sensor: a review, IEEE Trans. Cybern. 43 (5) (2013) 1318–1334.
[4] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, A. Blake, Efficient human pose estimation from single depth images, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2013) 2821–2840.
[5] F. Weichert, D. Bachmann, B. Rudak, D. Fisseler, Analysis of the accuracy and robustness of the leap motion controller, Sensors 13 (5) (2013) 6380–6393.
[6] T. Shiratori, H.S. Park, L. Sigal, Y. Sheikh, J.K. Hodgins, Motion capture from body-mounted cameras, in: ACM SIGGRAPH Conference, 2011, pp. 31:1–31:10.
[7] N. Werghi, Segmentation and modeling of full human body shape from 3-d scan data: a survey, IEEE Trans. Syst. Man Cybern. Part C 37 (6) (2007) 1122–1136.
[8] A. Boyali, M. Kavakli, J. Twamley, Real time six degree of freedom pose estimation using infrared light sources and wiimote ir camera with 3d tv demonstration, in: Mobile and Ubiquitous Systems: Computing, Networking, and Services, 2010, pp. 137–148.
[9] J. Tong, J. Zhou, L. Liu, Z. Pan, H. Yan, Scanning 3d full human bodies using kinects, IEEE Trans. Visual Comput. Graphics 18 (4) (2012) 643–650.
[10] J. Palacios, C. Sagrüés, E. Montijano, S. Llorente, Human–computer interaction based on hand gestures using rgb-d sensors, Sensors 13 (9) (2013) 11842–11860.
[11] F. Weichert, D. Bachmann, B. Rudak, D. Fisseler, Analysis of the accuracy and robustness of the leap motion controller, Sensors 13 (5) (2013) 6380–6393.
[12] F. Anderson, M. Annett, W.F. Bischof, Lean on Wii: Physical rehabilitation with virtual reality and Wii peripherals, Annu. Rev. CyberTherapy Telemedicine 8 (2010) 181–184.
[13] H.Y. Wooi, P. Raveendran, Single camera 3d human pose estimation: a review of current techniques, in: International Conference for Technical Postgraduates, 2009, pp. 1–8.
[14] C. Sminchisescu, 3d human motion analysis in monocular video techniques and challenges, in: Proceedings of the IEEE International Conference on Video and Signal Based Surveillance, 2006, pp. 76–100.
[15] T. Helten, A. Baak, M. Müller, C. Theobalt, Full-body human motion capture from monocular depth images, in: Time-of-Flight and Depth Imaging, Lecture Notes in Computer Science, vol. 8200, 2013, pp. 188–206.
[16] T.B. Moeslund, A. Hilton, V. Krüger, L. Sigal (Eds.), Visual Analysis of Humans – Looking at People, Springer, 2011.
[17] M. Eichner, M. Marin-Jimenez, A. Zisserman, V. Ferrari, 2d articulated human pose estimation and retrieval in (almost) unconstrained still images, Int. J. Comput. Vision 99 (2) (2012) 190–214.
[18] Y. Yang, D. Ramanan, Articulated human detection with flexible mixtures of parts, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2013) 2878–2890.
[19] R. Klette, G. Tee, Understanding human motion: a historic review, in: Human Motion, vol. 36, 2008, pp. 1–22.
[20] J. Aggarwal, M. Ryoo, Human activity analysis: a review 43(3) (2011) 16:1–16:43.
[21] R. Poppe, A survey on vision-based human action recognition, Image Vis. Comput. 28 (6) (2010) 976–990.
[22] M. Ye, Q. Zhang, L.W. 0002, J. Zhu, R. Yang, J. Gall, A survey on human motion analysis from depth data, in: Time-of-Flight and Depth Imaging, vol. 8200, 2013, pp. 149–187.
[23] A. Toshev, C. Szegedy, Deeppose: human pose estimation via deep neural networks, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1653–1660.
[24] W. Ouyang, X. Chu, X. Wang, Multi-source deep learning for human pose estimation, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2337–2344.
[25] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, C. Theobalt, Fast articulated motion tracking using a sums of gaussians body model, in: IEEE International Conference on Computer Vision, 2011, pp. 951–958.
[26] G. Juergen, Y. Angela, L.J.V. Gool, 2D Action recognition serves 3D human pose estimation, in: European Conference on Computer Vision, 2010, pp. 425–438.
[27] J. Shotton, A.W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images., in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1297–1304.
[28] C. Ennis, R. McDonnell, C. O'Sullivan, Seeing is believing: body motion dominates in multisensory conversations, in: ACM SIGGRAPH Conference, 2010, pp. 91:1–91:9.
[29] A. Mykhaylo, R. Stefan, B. Schiele, Pictorial structures revisited: People detection and articulated pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1014–1021.
[30] B. Sapp, D. Weiss, B. Taskar, Parsing human motion with stretchable models, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1281–1288.
[31] M. Eichner, V. Ferrari, We are family: joint pose estimation of multiple persons, in: European Conference on Computer Vision, 2010, pp. 228–242.
[32] J. Kim, K. Grauman, Boundary preserving dense local regions, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1553–1560.
[33] C. Guillot, M. Taron, P. Sayd, Q.-C. Pham, C. Tilmant, J.-M. Lavest, Background subtraction adapted to ptz cameras by keypoint density estimation, in: The British Machine Vision Conference, 2010, pp. 1–10.
[34] Y.J. Lee, J. Kim, K. Grauman, Key-segments for video object segmentation, in: IEEE International Conference on Computer Vision, 2011, pp. 1995–2002.
[35] P. Anestis, F. Vittorio, Fast object segmentation in unconstrained video, in: Proceedings of the International Conference on Computer Vision, 2013.
[36] H. Wang, D. Koller, Multi-level inference by relaxed dual decomposition for human pose segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 2433–2440.
[37] J. Puwein, L. Ballan, R. Ziegler, M. Pollefeys, Foreground consistent human pose estimation using branch and bound, in: European Conference on Computer Vision, 2014.
[38] D. Stavens, S. Thrun, Unsupervised learning of invariant features using video, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 1649–1656.
[39] B. Sapp, A. Toshev, B. Taskar, Cascaded models for articulated pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 406–420.
[40] X. Chen, A.L. Yuille, Articulated pose estimation by a graphical model with image dependent pairwise relations, 2014.
[41] J. Tompson, A. Jain, Y. LeCun, C. Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, 2014.
[42] L.D. Bourdev, J. Malik, Poselets: Body part detectors trained using 3d human pose annotations, in: IEEE International Conference on Computer Vision, 2009, pp. 1365–1372.
[43] F. Wang, Y. Li, Beyond physical connections: Tree models in human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 596–603.
[44] Y. Wang, D. Tran, Z. Liao, Learning hierarchical poselets for human parsing, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1705–1712.
[45] D. Tran, Y. Wang, D.A. Forsyth, Human parsing with a cascade of hierarchical poselet based pruners, in: International Conference on Multimedia and Expo, 2014, pp. 1–6.
[46] P. Srinivasan, J. Shi, Bottom-up recognition and parsing of the human body, in: A.L. Yuille, S.C. Zhu, D. Cremers, Y. Wang (Eds.), Energy Minimization Methods in Computer Vision and Pattern Recognition, vol. 4679, 2007, pp. 153–168.
[47] M. Dantone, J. Gall, C. Leistner, L. Van Gool, Human pose estimation using body parts dependent joint regressors, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3041–3048.
[48] B. Yao, L. Fei-Fei, Modeling mutual context of object and human pose in human-object interaction activities, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 17–24.
[49] V. Ramakrishna, D. Munoz, M. Hebert, J.A. Bagnell, Y. Sheikh, Pose machines: Articulated pose estimation via inference machines, in: European Conference on Computer Vision, 2014.

[50] D. Tran, D. Forsyth, Improved human parsing with a full relational model, in: European Conference on Computer Vision, 2010, pp. 227–240.
[51] L. Karlinsky, S. Ullman, Using linking features in learning non-parametric part models, in: European Conference on Computer Vision, 2012, pp. 326–339.
[52] M.A. Fischler, R.A. Elschlager, The representation and matching of pictorial structures, IEEE Trans. Comput. 22 (1) (1973) 67–92.
[53] P.F. Felzenszwalb, D.P. Huttenlocher, Pictorial structures for object recognition, Int. J. Comput. Vision 61 (1) (2005) 55–79.
[54] M. Eichner, V. Ferrari, Better appearance models for pictorial structures, in: The British Machine Vision Conference, 2009, pp. 1–11.
[55] Y. Yang, D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1385–1392.
[56] B. Sapp, B. Taskar, Modec: Multimodal decomposable models for human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3674–3681.
[57] M. Kiefel, P.V. Gehler, Human pose estimation with fields of parts, in: European Conference on Computer Vision, 2014.
[58] L. Pishchulin, M. Andriluka, P. Gehler, B. Schiele, Poselet conditioned pictorial structures, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1–8.
[59] L. Pishchulin, M. Andriluka, P. Gehler, B. Schiele, Strong appearance and expressive spatial models for human pose estimation, in: IEEE International Conference on Computer Vision, 2013, pp. 1–8.
[60] M. Sun, S. Savarese, Articulated part-based model for joint object detection and pose estimation, in: IEEE International Conference on Computer Vision, 2011, pp. 723–730.
[61] Y. Yang, S. Baker, A. Kannan, D. Ramanan, Recognizing proxemics in personal photos, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3522–3529.
[62] D. Weinland, R. Ronfard, E. Boyer, Free viewpoint action recognition using motion history volumes, Comput. Vis. Image Underst. 104 (2) (2006) 249–257.
[63] S. Zuffi, J. Romero, C. Schmid, M.J. Black, Estimating human pose with flowing puppets, in: Proceedings of the 2013 IEEE International Conference on Computer Vision, 2013, pp. 3312–3319.
[64] G. Gkioxari, P. Arbelaez, L. Bourdev, J. Malik, Articulated pose estimation using discriminative armlet classifiers, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3342–3349.
[65] D.J. Weiss, B. Taskar, Learning adaptive value of information for structured prediction, in: Advances in Neural Information Processing Systems, 2013, pp. 953–961.
[66] R. Tokola, W. Choi, S. Savarese, Breaking the chain: liberation from the temporal markov assumption for tracking human poses, in: International Conference on Computer Vision, 2013, pp. 2424–2431.
[67] A. Fathi, G. Mori, Human pose estimation using motion exemplars, in: IEEE International Conference on Computer Vision, 2007, pp. 1–8.
[68] R. Navaratnam, A. Thayananthan, P.H.S. Torr, R. Cipolla, Hierarchical part-based human body pose estimation, in: The British Machine Vision Conference, 2005.
[69] B. Sapp, C. Jordan, B. Taskar, Adaptive pose priors for pictorial structures, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 422–429.
[70] L. Bo, C. Sminchisescu, Twin gaussian processes for structured prediction, Int. J. Comput. Vision 87 (1-2) (2010) 28–52.
[71] S. Sedai, M. Bennamoun, D.Q. Huynh, Localized fusion of shape and appearance features for 3d human pose estimation, in: The British Machine Vision Conference, 2010, pp. 51.1–51.10.
[72] M. Straka, S. Hauswiesner, M. Rüther, H. Bischof, Skeletal graph based human pose estimation in real-time, in: The British Machine Vision Conference, 2011, pp. 69.1–69.12.
[73] C. Chen, Y. Zhuang, F. Nie, Y. Yang, F. Wu, J. Xiao, Learning a 3d human pose distance metric from geometric pose descriptor, IEEE Trans. Visual Comput. Graphics 17 (11) (2011) 1676–1689.
[74] M. Jiu, C. Wolf, G.W. Taylor, A. Baskurt, Human body part estimation from depth images via spatially-constrained deep learning, Pattern Recogn. Lett. 50 (2014) 122–129.
[75] M. Salzmann, R. Urtasun, Implicitly constrained gaussian process regression for monocular non-rigid pose estimation, in: Neural Information Processing Systems, 2010, pp. 2065–2073.
[76] C. Plagemann, V. Ganapathi, D. Koller, S. Thrun, Realtime identification and localization of body parts from depth images, in: Proc. International Conferences on Robotics and Automation, 2010.
[77] X. Wei, J. Chai, Modeling 3d human poses from uncalibrated monocular images, in: IEEE International Conference on Computer Vision, 2009, pp. 1873–1880.
[78] K. Buys, C. Cagniart, A. Baksheev, T.D. Laet, J.D. Schutter, C. Pantofaru, An adaptable system for rgb-d based human body detection and pose estimation, J. Visual Commun. Image Represent. 25 (1) (2014) 39–52.
[79] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, in: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1297–1304.
[80] T. Jonathan, S. Jamie, S. Toby, F. Andrew, The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 103–110.
[81] Kaliamoorthi, K. Ramakrishna, Parametric annealing: a stochastic search method for human pose tracking, Pattern Recogn. 46 (46) (2013) 1501–1510.
[82] N.-G. Cho, A.L. Yuille, S.-W. Lee, Adaptive occlusion state estimation for human pose tracking under self-occlusions, Pattern Recogn. 46 (3) (2012) 649–661.
[83] G. Mori, J. Malik, Recovering 3d human body configurations using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 28 (7) (2006) 1052–1062.
[84] V. Ramakrishna, T. Kanade, Y.A. Sheikh, Reconstructing 3d human pose from 2d image landmarks, in: European Conference on Computer Vision, 2012.
[85] C. Yu, K. Tae-Kyun, C. Roberto, Inferring 3d shapes and deformations from single views, in: European Conference on Computer Vision, 2010, pp. 300–313.
[86] K. Grochow, S.L. Marchtin, A. Hertzmann, Z. Popović, Style-based inverse kinematics, in: ACM SIGGRAPH Conference, 2004, pp. 522–531.
[87] M. Burenius, J. Sullivan, S. Carlsson, 3d pictorial structures for multiple view articulated pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3618–3625.
[88] V. Kazemi, M. Burenius, H. Azizpour, J. Sullivan, Multi-view body part recognition with random forests, in: British Machine Vision Conference, 2013.
[89] L. Rui, T. Tai-Peng, S. Stan, Y. Ming-Hsuan, 3d human motion tracking with a coordinated mixture of factor analyzers, Int. J. Comput. Vision 87 (1-2) (2010) 170–190.
[90] L. Sigal, A.O. Balan, M.J. Black, Humaneva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, Int. J. Comput. Vision 87 (1-2) (2010) 4–27.
[91] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2d human pose estimation: new benchmark and state of the art analysis, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[92] M. Everingham, L. Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vision 88 (2) (2010) 303–338.
[93] D. Ramanan, Learning to parse images of articulated bodies, in: Neural Information Processing Systems, 2006, pp. 1129–1136.
[94] S. Johnson, M. Everingham, Clustered pose and nonlinear appearance models for human pose estimation, in: The British Machine Vision Conference, 2010, pp. 12.1–12.11.
[95] A. Cherian, J. Mairal, K. Alahari, C. Schmid, Mixing body-part sequences for human pose estimation, in: Conference on Computer Vision and Pattern Recognition, 2014, pp. 2361–2368.
[96] T. Helten, A. Baak, G. Bharaj, M. Mller, H.-P. Seidel, C. Theobalt, Personalization and evaluation of a real-time depth-based full body tracker, in: 3DV, 2013, pp. 279–286.
[97] V. Ganapathi, C. Plagemann, D. Koller, S. Thrun, Real time motion capture using a single time-of-flight camera, in: Conference on Computer Vision and Pattern Recognition, 2010, pp. 755–762.
[98] V. Ganapathi, C. Plagemann, D. Koller, S. Thrun, Real-time human pose tracking from range data, in: European Conference on Computer Vision, 2012, pp. 738–751.
[99] S. Johnson, M. Everingham, Learning effective human pose estimation from inaccurate annotation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1465–1472.
[100] V. Ferrari, M. Marín-Jiménez, A. Zisserman, Progressive search space reduction for human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[101] P. Guan, O. Freifeld, M.J. Black, A 2d human body model dressed in eigen clothing, in: European Conference on Computer Vision, 2010, pp. 285–298.
[102] C. Ionescu, F. Li, C. Sminchisescu, Latent structured models for human pose estimation, in: IEEE International Conference on Computer Vision, 2011, pp. 2220–2227.
[103] B. Micusik, T. Pajdla, Simultaneous surveillance camera calibration and foot-head homology estimation from human detections, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 1562–1569.
[104] C. Desai, D. Ramanan, Detecting actions, poses, and objects with relational phraselets, in: European Conference on Computer Vision, 2012.