Machine Vision and Applications manuscript No.
(will be inserted by the editor)

Stereo Pictorial Structure for 2D Articulated Human Pose Estimation

Manuel I. López-Quintero · Manuel J. Marín-Jiménez · Rafael Muñoz-Salinas · Francisco J. Madrid-Cuevas · Rafael Medina-Carnicer

Received: date / Accepted: November 2015

Abstract In this paper, we consider the problem of 2D human pose estimation on stereo image pairs. In particular, we aim at estimating the location, orientation and scale of upper-body parts of people detected in stereo image pairs from realistic stereo videos that can be found in the Internet. To address this task, we propose a novel pictorial structure model to exploit the stereo information included in such stereo image pairs: the Stereo Pictorial Structure (SPS). To validate our proposed model, we contribute a new annotated dataset of stereo image pairs, the Stereo Human Pose Estimation Dataset (SHPED), obtained from YouTube stereoscopic video sequences, depicting people in challenging poses and diverse indoor and outdoor scenarios. The experimental results on SHPED indicates that SPS improves on state-of-the-art monocular models thanks to the appropriate use of the stereo information.

Fig. 1 Objective of this work. Our goal is to estimate the 2D pose of people in stereo videos. **(Top row)** Stereo pair from a video hosted in YouTube. **(Bottom row)** From left to right, disparity map computed from the stereo pair, and estimated pose of the upper-body (represented by sticks).

1 Introduction

Articulated Human Pose Estimation (HPE) is the task of obtaining the spatial configuration of human body parts from images. There is an increasing interest in HPE in highly uncontrolled imaging conditions [21,16], but despite the recent advances achieved, the problem is still open. Recovering the human pose from a single image is the most popular technique to predict positions of body joints [6]. However, several research areas use other methods such as single depth images [60,45], silhouettes for 3D reconstructed pose [2,62], Latent Variable Models [51], gradient combination and color segmentation cues [26], or Human Pose Coestimation [15].

Recently, the popularity of stereo video has grown significantly, and it may become an interesting market for home users in coming years [48]. Stereo image pairs provide extra information that can be employed to improve the results obtained by monocular approaches. This paper proposes an extension of the Eichner *et al.* [16] method to calculate 2D pose estimations in stereo image pairs as shown in Fig. 1 – we coin this problem Stereo Human Pose Estimation (SHPE).

The contribution of this paper is twofold. First, we propose a new technique to automatically detect and estimate the 2D pose of humans in stereo image pairs. The proposed method is based on a similarity constraint that promotes a collaboration between two pose estimators. We show experimentally that our SHPE proposal improves the accuracy of the estimated poses when compared to standard HPE techniques running independently on each image (see Fig. 2).

Department of Computing and Numerical Analysis. Maimonides Institute for Biomedical Research (IMIBIC).
University of Córdoba, 14071, Córdoba (Spain). Tel.:(+34)957212255
E-mail: i32loqum@uco.es, mjmarin@uco.es, rmsalinas@uco.es, ma1macuf@uco.es, rmedina@uco.es

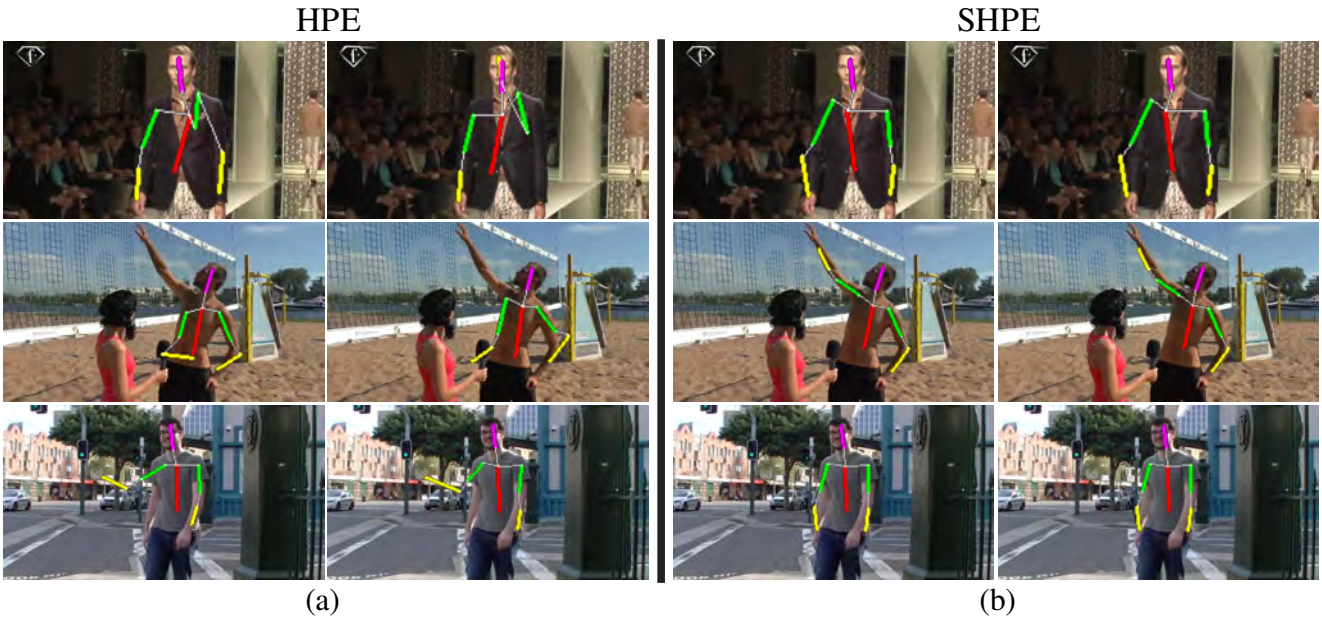


Fig. 2 Monocular HPE vs Stereo HPE. **(a)** Poses estimated by a monocular HPE model [16], where each view is processed independently. Note that a different pose is found for each image of the stereo pair. **(b)** Poses estimated by our stereo HPE model. In this case, a common pose is estimated for both images of the stereo pair.

See for example top row of Fig. 2.a where a state-of-the-art HPE method (i.e. [16]) fails on one of the images of the stereo pair but it is correctly estimated by our approach, Fig 2.b.

Our second contribution is a dataset for the SHPE problem. To experimentally validate our approach, we have created a new annotated dataset of 630 stereo image pairs from stereo videos depicting people in different backgrounds, clothing, lighting or locations in the image frames. The dataset covers upright people with a great variety of arms poses, covering the space of possible configurations quite uniformly.

The remaining of this paper is organized as follows: in section 2 we discuss related previous work; in section 3 we describe the basis of Pictorial Structures (PS) and Eichner *et al.*'s framework [16]; the methodology is outlined in section 4, proposing stereo adapted models for people detection, foreground-highlighting and inference; section 5 describes and presents the experimental results; and, finally, the conclusions are outlined in section 6.

2 Related works

Human Pose Estimation has been intensively studied in the field of Computer Vision for the last 20 years [35]. However, the problem of Human Pose Estimation in uncontrolled environments is still an open and challenging problem. Several approaches have been reported, and significant improvements have been obtained in both data representation and model design. Holistic shapes, silhouettes in particular, are common features for pose estimation. Current approaches achieve state-of-the-art performance by combining silhouettes with new features or constraints, including motion templates [38], pedestrian detectors [5], shape-contents [1] and user interaction [23].

A well-known approach consists on assembling body part detectors in a consistent configuration with the body structure. Such configuration is not defined by physical constraints but is described by soft restrictions. Pictorial Structures (PS) [18] are generative arrangements of parts, where each part is detected with its specific detector. A popular framework for HPE using PS models is the progressive search space reduction by Ferrari *et al.* [21]. This framework, based on the work of Ramanan [37], progressively reduces the search space for body parts to greatly increase the chances for correct pose estimation. Eichner and Ferrari [13] extend the previous approach by improving the person-specific appearance model used in the PS. Besides improving the appearance of body parts with densely sampled shape context descriptors, Andriluka *et al.* [4] propose to relate body parts in the PS model by using Gaussian distributions. Previous results with PS-based models can be improved by using adaptive pose priors [40] and cascades of PS [41]. Yang and Ramanan propose in [56] a tree-structured model with discriminatively trained parts that allows both people detection and human pose estimation simultaneously, in contrast to some of previous approaches (e.g. [21]) that rely on the detection of upper-bodies as a preprocessing stage. Zuffi *et al.* [63] introduce the Deformable Structures model where body parts can suffer non rigid deformations. We will see

later that we also adopt the use of PS models in our proposal.

Pictorial structures have been recently extended to deal with multiple views for articulated 3D human pose estimation. Amin *et al.*present in [3] a mixture of PS for 2D HPE on monocular setups that is also generalized for 3D HPE with multiple cameras. Nearly at the same time, Burenius *et al.* [11] introduce the concept of 3D Pictorial Structures as an extension of PS for estimating 3D human pose given a set of calibrated cameras. Many works rely on several views to improve the results. A popular approach for solving the problem of pose ambiguity in multi-view pose estimation is to increase the field of view with multiple images taken simultaneously using calibrated cameras [47,57,28,59]. Although they achieve excellent accuracy, potential applications are restricted to a fixed, calibrated multi-camera system.

The emergence of new active depth sensors (such as Time of Flight Cameras and the Kinect sensor) has led to the development of novel techniques exploiting this type of information. One approach is based on decision forests that are a classic method for inductive inference which have recently regained popularity. One of the key contributions in this context is the work of Shotton *et al.* [45], where the segmentation of the human body into parts is carried out by using a forest on single depth images. Other works use a random forest classifier to deal with the variation in appearance of body parts in 2D images [28] or a conditional regression forest model [49] that integrates dependency relationships between output variables by using a global latent variable. Lastly, Pons-Moll *et al.*'s work [36] introduce the Metric Space Information Gain (MSIG), a new decision forest training objective designed to directly optimize the entropy of distributions in a metric space. Other methods recognise 3D human poses from depth images, using techniques such as point cloud matching [8,58], constrained optimization [61] or constrained inverse kinematics [42]. Despite the good precision achieved by active stereo sensors, they require a controlled illumination since ambient light interferes with the sensor.

It is well known that the filming industry has recently adopted the 3D video as a new standard. In contrast to depth sensors, 3D video is normally recorded using a traditional stereo vision scheme: a pair of horizontal video cameras with short baseline. This configuration, that is initially selected to produce the visual perception of depth in viewers, can be exploited to improve state of the art HPE estimators using monocular images. The goal of this work is to extend the Eichner *et al.*'s HPE method [16] (summarized in Sec. 3.2) to take advantage of the additional information available in modern 3D films so as to detect and estimate the 2D pose of humans more reliably.

In general, most previous works carry out the estimation of the human pose for each person independently (including works based on other paradigms than Pictorial Structures [33,34]). The exceptions are [14], which models the occlusion interactions between nearby people in an image, [44], which estimates the human pose in a stereo image pair but is restricted to a single person and the dataset (H2view) is limited to eight subjects in three locations and, finally, [15] which deals with multiple people that are in a common, but unknown, pose.

Although the main goal of this paper is 2D pose estimation, some works use stereo cameras to derive 3D poses. However, its use is limited to very particular scenarios. The authors of [24] address the problem of 3D HPE by combining silhouettes, from two synchronized wide-baseline cameras, in a Bayesian Mixture Expert framework, where the models are trained from synthetic data. In [55] the 3D body pose is estimated from sequences of silhouettes represented by the silhouette history image descriptor, which helps to relate binary silhouettes with depth images within a hierarchy of clusters of body poses. In both previous works ([24, 55]), the use of silhouettes as input data limits its range of applicability to scenarios with static cameras, in contrast to our approach which can handle multiple people in scenarios with dynamic backgrounds. The system presented in [50] for 3D HPE on stereo sequences relies on the computation of 3D coordinates of body points based on the estimated disparity, and an ellipsoid-based body model that is fitted to data by using a Variational Expectation- Maximization approach. Since the authors work on video sequences, the body pose estimated in the previous frame is used to initialize the current pose. Random forests are used in [31] to estimate the 3D body pose from stereo images in two steps. Firstly, a grid-based shape descriptor is computed from depth maps to predict the orientation of the body. Then, such orientation information is used to select a pretrained random forest which is specialized in such body orientation. Although the authors present promising results on their custom dataset of ten people performing controlled movements, we cannot predict the behaviour of their method on uncontrolled scenarios, as the ones used in this paper. Note that neither [50] nor [31] make use of RGB data during the pose estimation process and, therefore, a good quality of the estimated depth maps is required to obtain accurate results, even in controlled indoor scenarios. In contrast, in this paper, the combination of both RGB and depth information allows us to deal with a wide range of body poses and very challenging imaging conditions.

In our work, we deal with stereo videos available on the Internet and, instead of using either calibrated cameras or depth sensors (as required by other approaches), we estimate disparity between the stereo image pairs to isolate people from background (in combination with an upper-body de-

tector) and, then, we apply a new Stereo Pictorial Structure model that simultaneously estimates the body pose in both viewpoints.

3 Human Pose Estimation using Pictorial Structures

This section provides the basis involved in HPE using Pictorial Structures (PS).

3.1 Pictorial Structures for HPE

Let us consider that the body parts of a person are represented by a Conditional Random Field [30] as proposed in [18]. Each part l_p is represented by a rectangular image patch, whose position is parametrized by its spatial location (x,y), orientation θ, scale s, and sometimes foreshortening [18, 10]. The tuple (x,y,θ,s) constitutes the state-space of the nodes. The posterior $P(L|I)$ of a configuration of parts L given an image I is defined as

$$P(L|I) \propto \exp\left(\sum_{(p,q)\in\varepsilon} \Psi_{pq}(l_p,l_q) + \sum_{p} \Phi_p(I|l_p)\right). \quad (1)$$

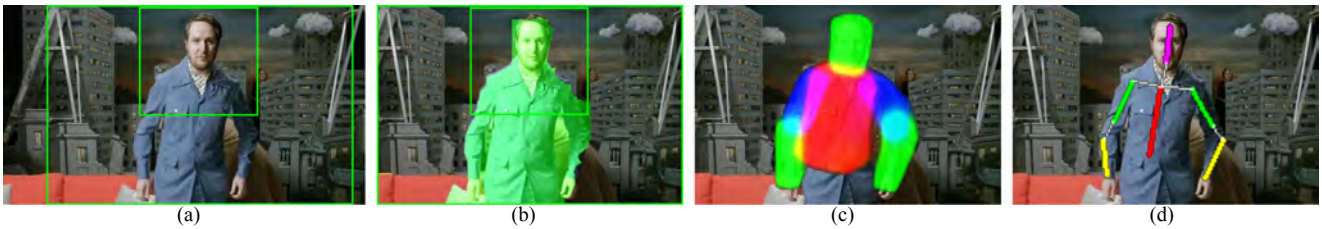
In the previous equation, $\Phi_p(I|l_p)$ is the unary potential associated to part l_p and encodes the local image evidence for such part in a particular position (likelihood). It depends on appearance models describing how parts look like. The success of PS for HPE depends strongly on having good appearance models, which limits the image positions likely to contain a part. Among the best performing models we find generic models based on gradients [4] and superpixels [41], as well as person-specific models derived automatically from the image [13,37]. Kinematic constraints (e.g. the lower arms must be attached to the upper arms) are encoded by the pairwise potential $\Psi_{pq}(l_p,l_q)$ (i.e. a prior on the relative position of two parts). In addition to kinematic constraints, the pairwise potentials can encode complex relations as parts coordination [32] or self-occlusion constraints [46].

Inference on the model returns either the single MAP configuration $L^* = \text{argmax}_L P(L|I)$ [4,18] or the posterior marginal distribution for each part [37]. Exact inference is possible when the model is a tree [18,37,21,4], however, some works have explored more complex topologies [22] or mixtures of trees [27].

3.2 Reducing the Search Space for HPE

Eichner *et al.* [16] propose a pipeline that progressively reduces the search space for body parts to increase the chances of correct 2D pose estimation – assuming that the torso is restricted to be nearly vertical and non-profile. This involves a generic detector using a weak model of pose to substantially reduce the full pose search space; and employ Grabcut on detected regions, proposed by the weak model, to further prune the search space. Also, they rely on the human parsing technique of Ramanan [37], on which they build directly. This model can be summarized in the following stages:

1. *Human Detection and Tracking*. Firstly, human upper-bodies (i.e. head and shoulders) are detected in every image (see Fig. 3.a), using a sliding window detector based on the part-based model of Felzenszwalb et al. [20]. In case video frames are processed, upper-body (UB) detections are grouped over time and each resulting track connects the detections of a different person in every video shot. Detections contains information about the rough position and scale of people in the image. Thanks to such information, the set of possible (x,y) locations of the body parts is reduced, and by fixing their scale, a dimension of the Pictorial Structures' state space is removed entirely. In practice, for each detected person, the state space is limited to a region of the image around the detection, covering the possible arms extent of the person. This image region is called the *enlarged window*.
2. *Foreground Highlighting*. In the second stage the search for body parts is limited to the so called enlarged window. The search area is further reduced by exploiting prior knowledge about the structure of the detection window, where some areas are very likely to contain body parts, whereas other areas are very unlikely. This allows the initialization of a GrabCut segmentation [39] to remove part of the background (see Fig. 3.b). Therefore, the search space will be limited to the (x,y) locations that lie within the foreground area determined by the GrabCut segmentation.
3. *Appearance Model Estimation*. In the third stage, a person-specific appearance model [13] is learnt from a single image based on two observations: (i) certain body parts have rather stable location with respect to the detection window (e.g. head and torso); (ii) often a person's body parts share similar appearance (e.g. upper arms).
4. *Parsing*. A body pose is estimated by running inference with generic appearance models (edges) and person-specific appearance models (computed in the third stage). The image area to be parsed is restricted to the region output of foreground highlighting (second stage). Explicit search for body parts over scales is not necessary as the person's scale has been fixed during the first stage. For each person detected in the image, this parsing stage delivers the posterior marginal distribution $P_i(x,y,\theta)$ for every body part (see Fig. 3.c–d).



(a) (b) (c) (d)

Fig. 3 Eichner et al. [16] overview. **(a)** *Upper body detection*: the output of the human upper-body detector (smaller rectangle) is enlarged (bigger rectangle) for the next processing steps. **(b)** *Foreground highlighting*: the result of the segmentation considerably eliminates most of the background clutter, what facilitates the later search of body parts. **(c)** *Inference*: the remaining pixels are labelled into body parts or background. Red specifies the torso, blue the upper arms, green the lower arms and the head. Frequently, the colors are overlapped, in that case yellow specifies the combination between lower-arm and torso, purple between upper-arm and torso, etc. **(d)** *Stick fitting*: the body pose is represented by straight line segments (sticks) that are obtained from the body part segmentations in (c).

4 Stereo Human Pose Estimation using PS

The main drawback of Eichner *et al.*'s method [16] (summarized in Sec. 3.2) can be found in the *Foreground Highlighting* stage. This stage is crucial since a percentage of the pixels are removed for further processing. Removing partial or full body parts during that stage prevents the pose estimator from being able to correctly localize such body parts. This usually happens when the color distribution of the background is similar to some of the body parts.

As already indicated, the extra information available in stereo sequences can be used in order to overcome these problems. We propose an extension of the previous method based on stereo information. First a person detector (Sec. 4.1) is run on both images of the stereo pair, in all video frames. Then, a temporal association algorithm is applied to remove false positives. This process is run independently on each stereo pair. Thus, we then match tracks using a measure based on the degree of overlapping of the detected bounding-boxes (BB). Then, for each detected person, disparity information is computed in the detected regions. Disparity is computed only in the subregions of the image with people to speed up the processing. Using the computed disparity, a segmentation method (Sec. 4.2) is employed to separate body pixels from background in both images. Finally, we apply our stereo parsing model (Sec. 4.3) to infer the pose.

Note that, although the input of the proposed pipeline is a single stereo pair, the detection rate would improve and the amount of false positives would decrease [16] by exploiting temporal smoothness of the stereo video. However, we choose not to restrict our proposal to video sequences by including temporal constraints in the PS model, thus making the model ready for stereo pictures.

The rest of this section provides a detailed explanation of the stages summarized above.

4.1 People Detection and Tracking

As firstly done by Ferrari *et al.*in [21], we start by detecting human upper-bodies in every frame to reduce the search space. We use the upper-body detector released by the authors of [16] at [52]. This upper-body detector is based on the successful Deformable Parts Model (DPM) of Felzenszwalb *et al.* [19]. A DPM contains several Histogram of Oriented Gradients filters [12] related by deformable edges.

We run the upper-body detector on each frame of the stereo pair independently. Then, in order to remove false positives, we carry out a tracking-by-detection process, as in [16], which generates one track per potential person in each stereo pair independently. Tracks are given a score based on their length and detection score. Low scored tracks are discarded for the subsequent stages. Finally, possible gaps in tracks (due to misdetections, e.g. low contrast or profile viewpoints) are filled in by interpolation. Fig. 4 shows the raw upper-body detections (top) and the output of the tracker (bottom) on one camera of the stereo pair for a given sequence. Note how the false positives are removed after the tracking.

Using the above calculated (most reliable) tracks independently for each camera, it is necessary to match the tracks in the left camera with the ones in the right camera. To do so, for a given instant of time, we compute the intersection-over-union (IoU) [17] of all the possible pairs of bounding-boxes belonging to a different stereo pair. Then, we match the tracks whose sum of IoU is maximum. In case the cameras diverge, a more sophisticated procedure, as matching of histograms of colour, should be used. However, our IoU-based matching works since the majority of commercial stereo cameras used for recording user-consuming videos have a short base line.

4.2 Stereo Foreground-Highlighting

As reported by [21], the location and scale information provided by an upper-body detection greatly constrains the space

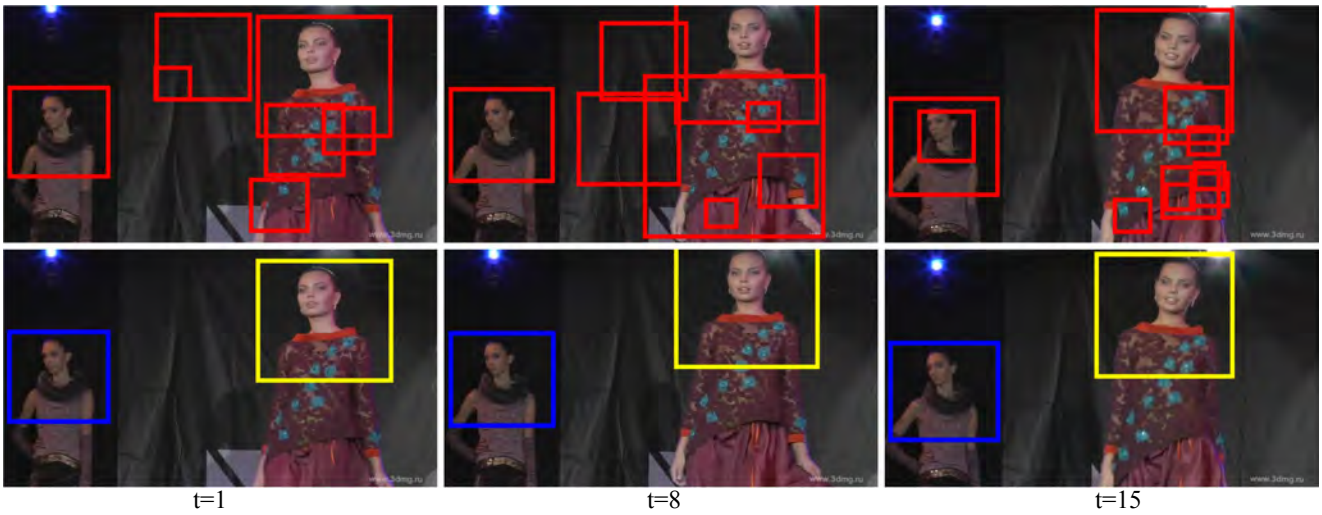


Fig. 4 People detection and tracking. Bounding-boxes returned by the upper-body detector **(top row)** and final tracks **(bottom row)** in frames 1 **(left column)**, 8 **(middle column)** and 15 **(right column)** of a stereo sequence. After the tracking process, false positive detections are removed, and a single track is assigned to each person for the whole stereo sequence.

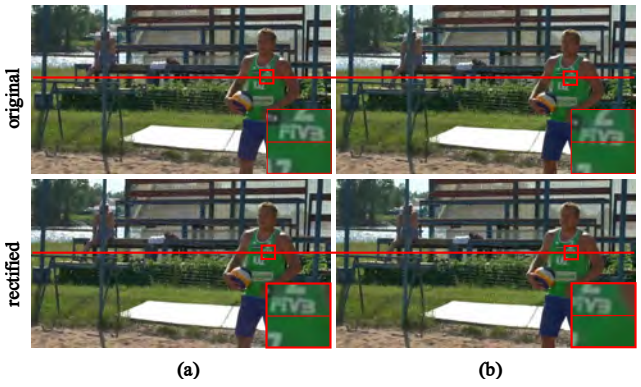


Fig. 5 Stereo image pair rectification. Images in the top row are the original left (a) and right (b) image pairs from one of the sequences tested, and images in the bottom row show the result of rectification. Please focus on the regions enclosed in red squares. It can be seen that the line passing below number 3 in original left image crosses the number in the original right image. That vertical misaligment causes inaccuracies in the block-matching stereo algorithm. However, the aligment errors are greatly reduced after rectification.

of possible body parts. In order to reduce even more the search for the inference, image foreground/background segmentation algorithms, i.e. GrabCut [39], are extended to extract body parts from the clutter [16]. However, to segment all possible arm poses is still a hard problem. Therefore, we propose a new strategy to extract the person from the background, helping to solve the arm segmentation problem. Most of the image segmentation algorithms are based on a Gaussian Mixture Model of a two-class image. It has the potential for effective segmentation provided that the histogram of the image approximates a Gaussian mixture and the parameters of the model can be estimated accurately.

Our proposal, coined Stereo Foreground Highlighting (SFH), exploits stereo information to separate the image pixels of people from the background. Since we deal with stereo videos downloaded from the Internet, recorded by commercial cameras aimed at 3D filming purposes, perfect horizontal alignment cannot be guaranteed since human brain does not need it to estimate three-dimensional information (see Fig. 5 for a real example). However, special care must be taken on the alignment so as to obtain good results from stereo block-matching algorithms. To solve that problem, we apply the uncalibrated stereo image rectification method available online at Mathworks' website [1]. It consists in collecting interest point from a image pair (using SURF [9]) and then finding putative correspondences filtered by epipolar constraints [25]. The correspondences are employed to compute the rectification transformations that produces a proper horizontal alignment (see Fig. 5). The estimation of the transformations is done once per video (using the first frame), and then applied to the rest of the frames.

Then, disparity information [29] is computed only in the enlarged image subregions (Fig. 6.a) where the person has been detected (instead of computing it in the whole image), so as to speed up computation. The computed disparity map D (Fig. 6.b) indicates for each pixel $D_{(x,y)}$ of the left image the horizontal displacement required to obtain the same pixel in the right image, i.e., a pixel (x,y) in left image corresponds to pixel $(x+D_{(x,y)},y)$ in the right image.

Given that the upper part of the torso is detected in the previous phase, a smaller rectangular region in the center of the bounding box is selected as a representative sample of the disparity distribution for the whole body in the left image

[1] `http://es.mathworks.com/help/vision/examples/uncalibrated-stereo-image-rectification.html`

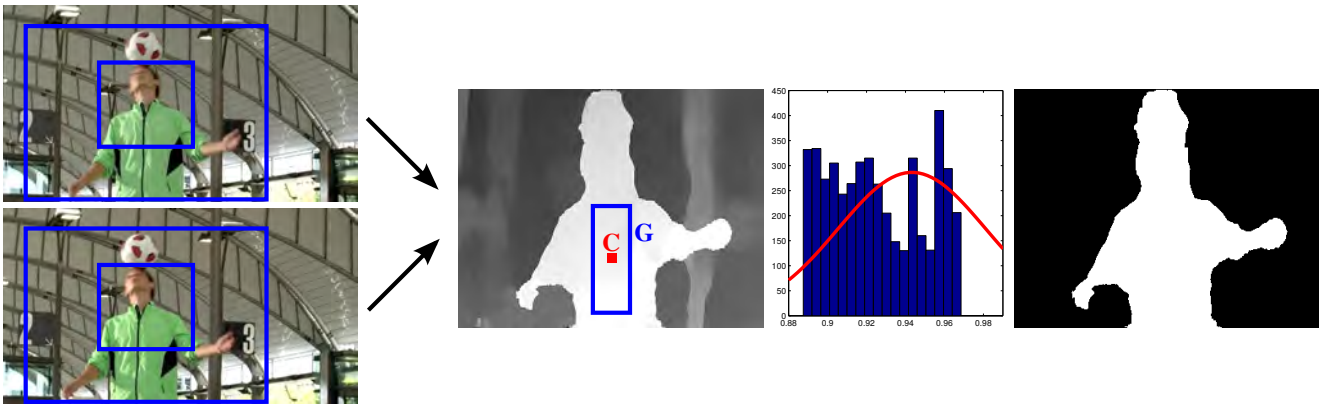


Fig. 6 People detection and stereo foreground highlighting. *Stereo upper body detection:* **(a)** The two enlarged bounding-boxes (from each view) are averaged. Therefore, the resulting bounding boxes for both images are exactly the same. *Stereo foreground highlighting*: **(b)** Firstly, the disparity map is computed. Brighter pixels indicate objects which are closer to the camera. Then, we establish a rectangular region G on the torso which is used as a prior for segmentation. Point C is the seed selected to initialize the algorithm. **(c)** We assume that the disparity values follow a normal distribution, which parameter μ is estimated from region G. **(d)** Finally, the previously learnt distribution is used to output a binary mask from the disparity map by region growing from seed C.

(Fig. 6.b). Assuming that most of the person body configurations can be modelled as a normal distribution $\mathcal{N}(\mu,\sigma)$ (Fig. 6.c), μ is calculated as the mean of disparity values in the torso region. Then, σ is selected considering the average dimensions of people so that extended arms fits into the distribution. Consequently, the segmentation problem is tackled as a region growing algorithm with seeds selected in the torso region (see point C in Fig. 6.b), and using the previous normal distribution to determine the probability of adding points to the segmented region. This approach allows for extended arms to be properly added as part of the person. Figure 6.d shows the result of applying the proposed method to one of the images in our dataset. As in [16], we also add to foreground a rectangular region, defined as a function of the upper-body bounding box, that covers the head and part of the torso, exploiting in this way prior information provided by the UB detection.

The generated mask corresponds to the person's upper body in the left image I^A. In order to obtain the equivalent mask for the right image I^B, the location of pixels from I^A are calculated in the right image by employing the disparity. Please notice that although our Stereo Foreground Highlighting shares some ideas with the method proposed by Sheasby *et al.*in [43], the latter runs a human pose estimator [56] to define two starting seeds for their region growing algorithm. What for them is a stage of their method, for us is our final goal.

Both the disparity estimation and the foreground segmentation are carried out independently for each upper-body enlarged region. Therefore, situations such as people standing next to each other, or people pointing towards or away from the camera are treated satisfactorily, see for example rows 1, 4 and 5 of Fig. 7. We have included in column (c) the segmentation mask obtained by applying the original foreground highlighting algorithm of Eichner *et al.* [16]. Note how it frequently loses part of the arms, as in rows 2, 4 and 5, in contrast to our disparity-based proposal that satisfactorily keeps them. However, in the example depicted in row 6, due to a not very accurate estimation of the disparity, our method removes fewer background pixels than the GrabCut-based approach.

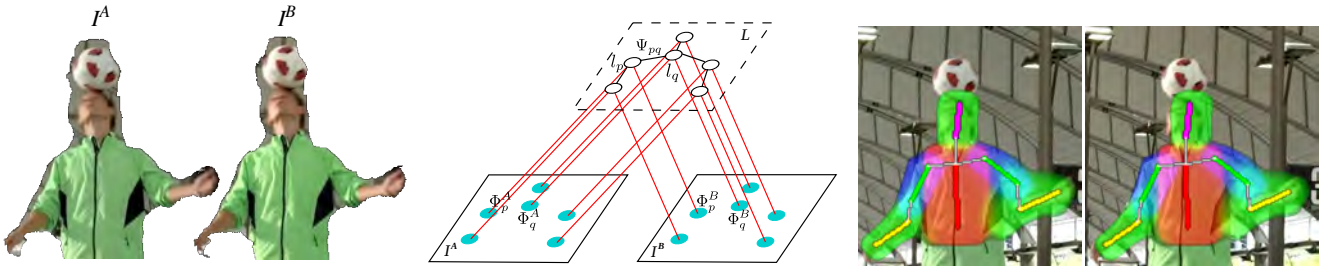
4.3 Stereo Pictorial Structure: SPS

Our proposal to add stereo information in pictorial models is based on the model of Eichner *et al.*'s [16] summarized in Sec. 3.2. Among other things, [16] extends Ramanan's model [37] with orientation priors Υ of the torso and the head to be nearly vertical. We also benefit from the algorithm of Eichner and Ferrari [13] to generate person-specific appearance models Φ for each image. Kinematic constraints Ψ are as in Ramanan's work [37]: for the relative position (x,y) we use a truncated cost, giving an uniform probability close to the joint location and zero elsewhere, and for the relative orientation θ we use a histogram of orientations learnt from training data [37].

Let I^A and I^B be the segmented images after applying SFH to the stereo image pair (Fig. 8.a), and l_p^A, l_p^B the upper-body parts of I^A and I^B respectively. Since we are working with stereo pairs, the following relation holds: $l_p^B = \mathscr{D}(l_p^A, D)$ where $\mathscr{D}(l,D)$ is an indicator function that applies the disparity map D to the configuration l.

To take advantage of the appearance information encoded in the images of the stereo pair, we combine the unary potentials Φ_p corresponding to the same body part in each

Fig. 7 Qualitative results of background removal: Stereo Foreground Highlighting vs GrabCut. From left to right: **(a)** estimated disparity map for the target person; **(b)** overlaid foreground mask proposed by SFH; **(c)** overlaid foreground mask proposed by GrabCut as used in [16]. The inner green rectangles in (b) and (c) represent the upper-body detection, whereas the outer green rectangles represent the enlarged window where the body parsing stage will be carried out. Note the different situations that SFH can handle satisfactorily: arms in a different plane of torso (e.g. pointing to the camera in row 1); arms above the head (row 5); multiple people in different depth planes (row 3), etc. In general, SFH removes more background pixels than GrabCut but keeping more actual foreground (e.g. in row 4, one hand is cut by GrabCut). In contrast, bottom row shows an example where three people are in the same depth plane and are all included in the foreground mask of the central person, whereas GrabCut keep the other persons as background.



(a) (b) (c)

Fig. 8 Stereo inference. **(a)** The person segmentation (I^A, I^B), returned by SFH, is used to compute $P(L|I)$. **(b)** In the Stereo Pictorial Structure, each node represents a body part from each stereo pair (head, torso, left/right upper/lower arms). The tree includes edges between every two body parts, parametrized by location (x,y) and orientation θ, which are physically connected by kinematic priors Ψ in the human body. Each hidden node (empty circle) is related to two observed nodes (one per view), represented by blue filled circles, which are associated to unary potentials Φ (i.e. image evidences). **(c)** Configuration of the parts L given by the Stereo Pictorial Structure model. Note that the configuration L is the same for both I^A and I^B, except the displacement in the horizontal axis given by the disparity.

view through a function Ω. Such function will be instantiated later. Therefore, our upper-body stereo pictorial structure (SPS) (depicted in Fig. 8.b) consists of two sub-models (one per view) related by the disparity D and the function Ω. Each sub-model consists of six body parts, namely head, torso, upper and lower arms, connected in a tree structure by the kinematic priors $\Psi(l_p, l_q)$. The probability of a configuration L given the stereo pair $\mathscr{I} =< I^A, I^B >$ and the disparity map D is given the following equation:

$$P(L|\mathscr{I},D) \propto \exp\left\{ \sum_{(p,q)\in\varepsilon} \Psi_{pq}(l_p,l_q) + \sum_p \Omega\Big(\Phi_p(I^A|l_p), \Phi_p(I^B|\mathscr{D}(l_p,D))\Big) + \Upsilon(l_{head}) + \Upsilon(l_{torso}) \right\}. \quad (2)$$

For the sake of clarity, l_k refers to the configuration of limb k on image A.

The selection of the function Ω leads to specific instantiations of the proposed model. One could think of defining Ω as, for example, the sum, the product, the arithmetic mean, etc. of the likelihoods. After carrying out some early experiments, we define function Ω_{max} as the maximum of the two likelihoods Φ^A and Φ^B:

$$\Omega_{max}(\Phi^A, \Phi^B) = \max\left(\Phi^A, \Phi^B\right) \quad (3)$$

This choice relates both viewpoints by giving preference to greater likelihood values between pairs of corresponding points.

4.3.1 Inference

At inference time, SPS finds a configuration of body parts L^* that maximizes $P(L|\mathscr{I},D)$ (see Fig. 8.c):

$$L^* = \underset{L}{\operatorname{argmax}} P(L|\mathscr{I},D). \quad (4)$$

Note that the coordinates (x,y) of the body parts obtained in L^* are defined in the reference system of I^A. Therefore, to obtain the body parts location in I^B, the previously defined function $\mathscr{D}(\cdot,\cdot)$ is employed.

The inference can be performed in an efficient and exact way [37], by sum-product Belief Propagation, since there are no loops in the graphical model (i.e. the model structure is a tree).

4.3.2 Implementation details

Unary potentials We use the unary potentials described in [16]. The edge images (Sec. 3.2) are convolved with the person-generic part templates released by the author of [37]. A total of 24 discretized orientations are used during the convolution to deal with the rotation of the limbs. For the color-based unary potentials (Sec. 3.2), the CIE-Lab color space is used to compute color histograms with dimensionality $8 \times 16 \times 16$. For each body part l_i, we have a probability distribution of color c for both foreground and background, which will be used as likelihood: $P_i(c|fg)$ and $P_i(c|bg)$. The color-based posterior probability (i.e. probability of belonging to part i given the color pixel c) is computed by using Bayes' rule (assuming $P_i(fg) = P_i(bg)$):

$$P_i(fg|c) = \frac{P_i(c|fg)}{P_i(c|fg) + P_i(c|bg)}$$

As in [37], the edge-based unary potential is used to initialize the color model. Then, both kind of unary potentials are added to compute Φ_p. Reader is referred to [37] for further details.

Binary potentials For the binary potential $\Psi_{pq}(l_p, l_q)$, we use a truncated cost as in [16], giving 0 probability to invalid configurations and an uniform probability to valid configurations, defined by the kinematic model (e.g. head must be attached to the torso). A configuration is said to be valid

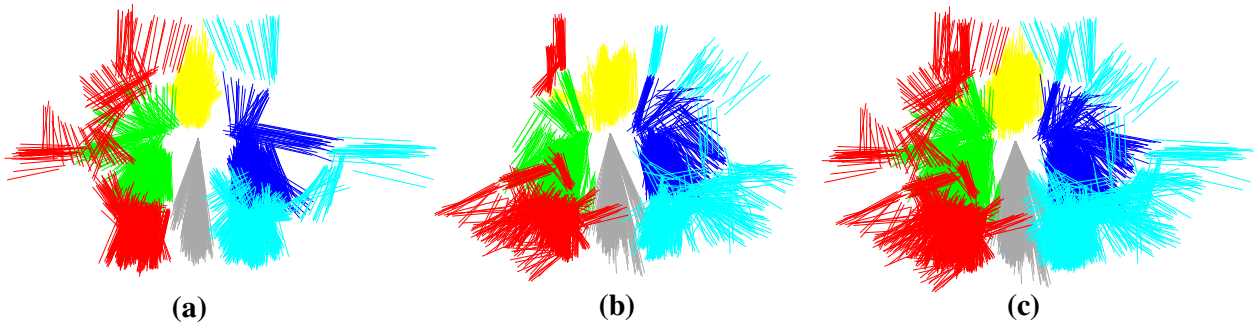


Fig. 9 Distribution of the ground-truth poses in the SHPE dataset. Colour coding of sticks: head in yellow; torso in gray; upper-arms in green and dark blue; and, lower-arms in red and light blue. **(a)** Partition A. **(b)** Partition B. **(c)** Whole dataset. (Best viewed in colour.)

if both the relative location of the limbs and their relative orientation are within the intervals learnt during training. In particular, we use the pretrained data provided by the author of [37].

5 Experimental Results

This section aims at validating the proposed method and to compare its results with state-of-the-art methods in the field. Since using stereo information for human pose estimation in uncontrolled Internet stereo videos is a new task, as far as we know, there are no available datasets for comparison. As a consequence, we start by introducing the dataset that we have created (Sec. 5.1), which contains ground-truth annotations for body parts. Then, we describe how we evaluate the performance (Sec. 5.2). Next, we define the experiments that we have carried out and quantitatively present the results of applying our SHPE framework and the competitors algorithms on SHPED (Sec. 5.3). Finally, we present a discussion of the SHPE performance and analyze the impact of various components of our method (Sec. 5.4).

5.1 Dataset of stereo image pairs for SHPE: SHPED

To analyze the results of the proposed stereo-based framework, we introduce a new dataset for SHPE named *SHPE Dataset* (SHPED)[2]. It contains 630 stereo image pairs (i.e. 1260 images) grouped into 42 video clips of 15 frames each. The clips have been extracted from 26 stereo videos obtained from the popular video-sharing website YouTube[3].

Fig. 13 shows some keyframes extracted from the dataset. The clips included in the dataset depict people in a wide range of variations in appearance, clothing, human pose, illumination and/or background. Since there are many different stereo cameras on the market, we obtained stereo videos with different image quality and baseline separation (distance between the two cameras).

We provide, as in [21], 1470 stickman annotations (i.e. there are sequences with more than one person per image) for the upper-body of people. We have annotated all the people that satisfy the following conditions (as in [16]): up-right position, non-profile viewpoint of the body, and all upper-body parts almost visible along the whole sequence. We have annotated 49 individuals in SHPED with these conditions. In this work, these annotations are used only for evaluation purposes. In addition, we provide a plane projective transformation for every clip, as a pre-processing step for rectifying the stereo image pairs and the stickmen.

In order to allow comparable results for future publications on this dataset, we have defined two disjoint partitions on the clips (i.e. 50% each). These two partitions have been randomly created – just making sure that two clips extracted from the same video were not located in the same partition. Fig. 9 shows the distribution of the ground-truth poses per partition (Fig. 9.a and .b), and the whole dataset. Each stick represents a body part. Note that the clips included in SHPED cover a wide range of spatial locations and orientations of the limbs, what makes a challenging task for HPE methods.

5.2 Evaluation metrics employed

Our SHPE technique estimates a stickman for each detection window computed in SHPED. With this stickman and the manually annotated in SHPED, we evaluate the performance using two measures.

First, we employ the Percentage of Correctly estimated body Parts (PCP) proposed in [16]. An estimated body part is considered correct if its segment endpoints lie within a fraction of the length of the ground-truth segment from their annotated location. By varying the fraction (τ_{PCP}) between 0.1 and 0.5 we obtain a PCP-curve. The lower (τ_{PCP}), the stricter the criterion and the more accurate the estimated body parts are deemed correct. PCP is evaluated only for

[2] SHPED is available at: `http://www.uco.es/investiga/grupos/ava/node/47`

[3] We used the tag *yt3d:enable=true* to find stereo videos in YouTube (http://youtube.com)

Table 1 Summary of acronyms used in this paper.

Acronym	*Full name*
SPS	Stereo Pictorial Structure
SFH	Stereo Foreground Highlighting
SHPE Ω_{max}	SHPE framework using the function Ω_{max}
SHPED	Stereo Human Pose Estimation Dataset
EA [16]	Eichner *et al.*'s framework
FMP [56]	Flexible Mixtures of Parts
PCE [15]	Human Pose Co-Estimation (Direct Model)

stickmen that have been correctly localized by our stereo upper-body detection windows. To allow an easy comparison of results, we use the implementation of the PCP criterion published online by the authors of [16] in [54].

Second, in order to summarize the values obtained for the different τ_{PCP} values used to build the PCP-curve, we compute the Area Under the PCP Curve (AUC-PCP). This area allows to easily compare different algorithms without having to choose any particular operational point τ_{PCP}. For the experimental results reported in Sec. 5.3, we compute AUC-PCP in the interval $\tau_{PCP} = [0.1, 0.5]$.

For evaluation purposes, in both monocular and stereo cases, each image of the stereo pair is considered an independent instance and, therefore, the two PCPs obtained from the pair are not combined by any means.

5.3 Comparative results

Since the dataset defines two disjoint partitions (Sec. 5.1), our experiments follow a two-fold cross validation approach. We report the mean PCP and AUC-PCP over the two partitions. The reader is referred to Tab. 1 for a summary of the main acronyms used in this paper.

5.3.1 Baseline methods

In order to put in context our proposal, we compare it with some state-of-the-art monocular HPE methods. Note that, since each image of the stereo pair has its own stickman annotation, we treat each view independently for evaluation purposes (i.e. PCP for left view may differ from PCP for right view).

Eichner *et al.* [16] (EA) Since we base our SHPE framework on the one proposed by Eichner *et al.* in [16] (see Sec. 3.2), we run their algorithm on SHPED by using their source code [54] and their default parameters. Here, we apply this method to each image of the stereo pair independently. The results of this experiment are summarized in row 'EA' of Tab. 2.

Flexible Mixtures of Parts (FMP) Yang and Ramanan propose a Flexible Mixtures of Parts (FMP) [56] to address the HPE problem. Since FMP is considered one of the state-of-the-art models, we run their code [53] on our stereo dataset for comparison purposes. We use the default parameters included in their software. As done above, we apply this method to each image of the stereo pair independently. The results of this experiment are summarized in row 'FMP' of Tab. 2.

In addition, row 'FMP+BB' shows the results obtained when the FMP method is applied to the same image windows returned by our people detection stage (Sec. 4.1), instead of searching over the whole image. This will allow a much more direct comparison with our SPS model.

Human Pose Co-Estimation (PCE) Human Pose Co-Estimation (PCE) [15] tries to estimate a common pose for a group of people in an image. Since PCE is somehow related to SHPE (in the sense of sharing a common pose), we have implemented, and applied to our stereo dataset, the *Direct Model* presented in the original paper. In this case, we obtain common estimations for each image of the stereo pair. We use the same default parameters of the upper-body detector and foreground highlighting stage used for the baseline EA. The results of this experiment are summarized in row 'PCE' of Tab. 2.

5.3.2 SPS evaluation

For evaluating our proposed pipeline (Sec. 4), we set the free parameters so as to maximize AUC-PCP on the training set – this is repeated for each partition of the dataset. In particular, we have to set the value of σ for the stereo foreground-highlighting step (Sec. 4.2). We carry out a grid-search in the interval $\sigma = [0.190, 0.250]$. As in the protocol of [13], PCP and AUC-PCP are computed only on correct detections (i.e. covering a ground-truth stickman). In our case, these detections cover the 100% of the ground-truth for SHPED. Note that we use as input of the HPE algorithms the same set of detection windows for all the methods but the one represented in Tab. 2 as 'FMP'.

Row 'SHPE Ω_{max}' of Tab. 2 show the results of our SPS model by using Eq. 3. Note that we report the AUC-PCP results for both the whole upper-body parts (UBP) and only the arms. Moreover, we report the PCP values for values 0.2 and 0.5 of τ_{PCP}. To evaluate the contribution of the proposed Stereo Foreground Highlighting stage in the performance of the system, instead of using the SPS model for inference, we use the monocular approach of Eichner *et al.* [16] on top of image pairs segmented by the SFH. The results of this case are shown in row 'EA+SFH' of Tab. 2.

Using our stickman ground-truth annotations, we try to evaluate the percentage of foreground pixels that are missing after the SFH stage. In partition *A* we miss around 4.8%, whereas in partition *B* we miss around 3.1% of the foreground pixels. Two examples are represented in Fig. 10, where only few pixels are missing in top row (white pixels in column (c)), but a significant percentage of important

Table 2 Comparative with the state-of-the-art. We report the quantitative results after applying the different algorithms on SHPED. Each entry reports either AUC-PCP or PCP values for both the whole upper-body (*UBP*) and only the arms (*Arms*). Column *AUC* represents the AUC-PCP value, Δ denotes the difference of AUC-PCP values between target algorithm and EA (baseline), and column % shows the previous difference in terms of percentage. With regard to PCP results, τ_{PCP} is the threshold used in the PCP curve. Note that SHPE Ω_{max} clearly improves on the results offered by the baseline [16], particularly on the arms. The highest results are marked in bold.

	AUC-PCP						*PCP* (%)			
Algorithm	UBP			Arms			UBP τ_{PCP}		Arms τ_{PCP}	
	AUC	Δ	%	*AUC*	Δ	%	*0.2*	*0.5*	*0.2*	*0.5*
SHPE Ω_{max}	**0.633**	0.047	**8.0**	**0.531**	0.066	**14.2**	**50.0**	**85.6**	36.6	**79.7**
EA [16] + SFH	0.627	0.041	7.0	0.524	0.059	12.7	49.4	84.6	36.1	78.2
FMP [56]	0.599	0.013	2.2	0.509	0.044	9.5	45.7	81.4	**36.7**	72.9
FMP [56] + BB	0.589	0.003	0.5	0.505	0.040	8.6	44.2	81.3	36.2	73.0
PCE [15]	0.579	-0.007	-1.2	0.469	0.004	0.9	47.2	78.0	34.2	68.7
EA [16] (baseline)	0.586	0	0	0.465	0	0	47.3	78.6	33.1	69.4

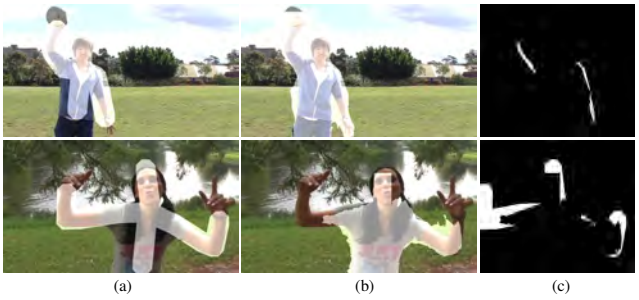


Fig. 10 Derived ground-truth for Stereo Foreground-Highlighting evaluation. **(a)** Coarse ground-truth region derived from stickmen annotations. **(b)** Output of our SFH. **(c)** Positive differences of mask in (a) minus mask in (b). White pixels represent ground-truth pixels not included in the SFH mask. The top row example is considered a good segmentation, whereas bottom row example represents missing pixels in the arms.

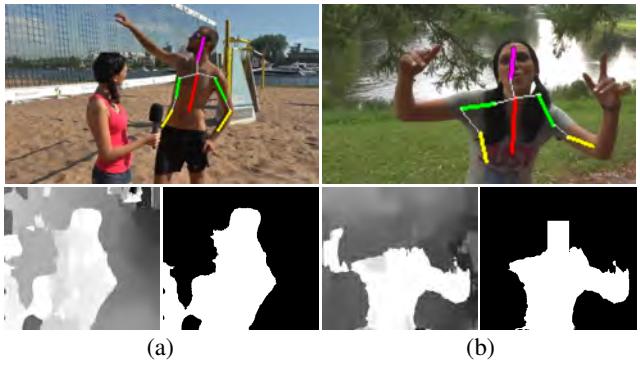


Fig. 11 Failure cases with SHPE. **(Top)** Estimated pose for a single person. **(Bottom)**. Estimated disparity map for the region of interest and derived foreground mask. **(a)** The right arm of the man is not included in the foreground mask due to a wrong estimation of the disparity. **(b)** The left hand and the right arm are not included in the foreground mask delivered by SFH. The head+torso prior is clearly visible in this example in the head region.

pixels are not included by SFH in the foreground of the example represented in the bottom row.

In general, wrong pose estimations with SHPE are due to inaccurate estimations of the disparity maps. Two frames of representative failure cases are shown in Fig. 11. In both cases, the SPS model is unable of correctly estimate the location of body parts that have been previously removed from the background area: one arm in the case of Fig. 11.a, and one arm plus a hand in case of Fig. 11.b.

In addition to the previously described tables, Fig. 12 presents a comparative of the PCP curves obtained from the evaluation of each method on set A and set B showed in left column and right column respectively. Top row of figure corresponds to the whole upper-body (UBP), whereas bottom row corresponds to the four body parts related to arms (Arms). AUC-PCP for each method is shown into parenthesis in the legend of the plots.

Fig. 13 shows some examples of correctly estimated poses in diverse challenging situations, as well as one estimation not as accurate as desired and two failures in difficult situations (i.e. 5–b contains a difficult arm pose estimation, and 5–c shows blurry arms due to motion).

5.4 Discussion

We discuss here the results obtained in the previous experiments.

If we compare, in terms of AUC-PCP, row 'EA' to 'SHPE Ω_{max}' of Tab. 2, we can see how our stereo pipeline greatly contributes to the final performance of the system: 8%. If we focus only on the estimation of the arms (i.e. 4 parts out of 6), the improvement is even larger: 14.2%. In our opinion, the clear improvement shown by SHPE is due to the stereo foreground highlighting stage, where the removal of actual background pixels is more precise thanks to the use of disparity. In turn, making easier the success of subsequent stages. This fact is reflected in row 'EA+SFH' where the disparity-based segmentation boosts the performance up to 7% with regard to the baseline 'EA'.

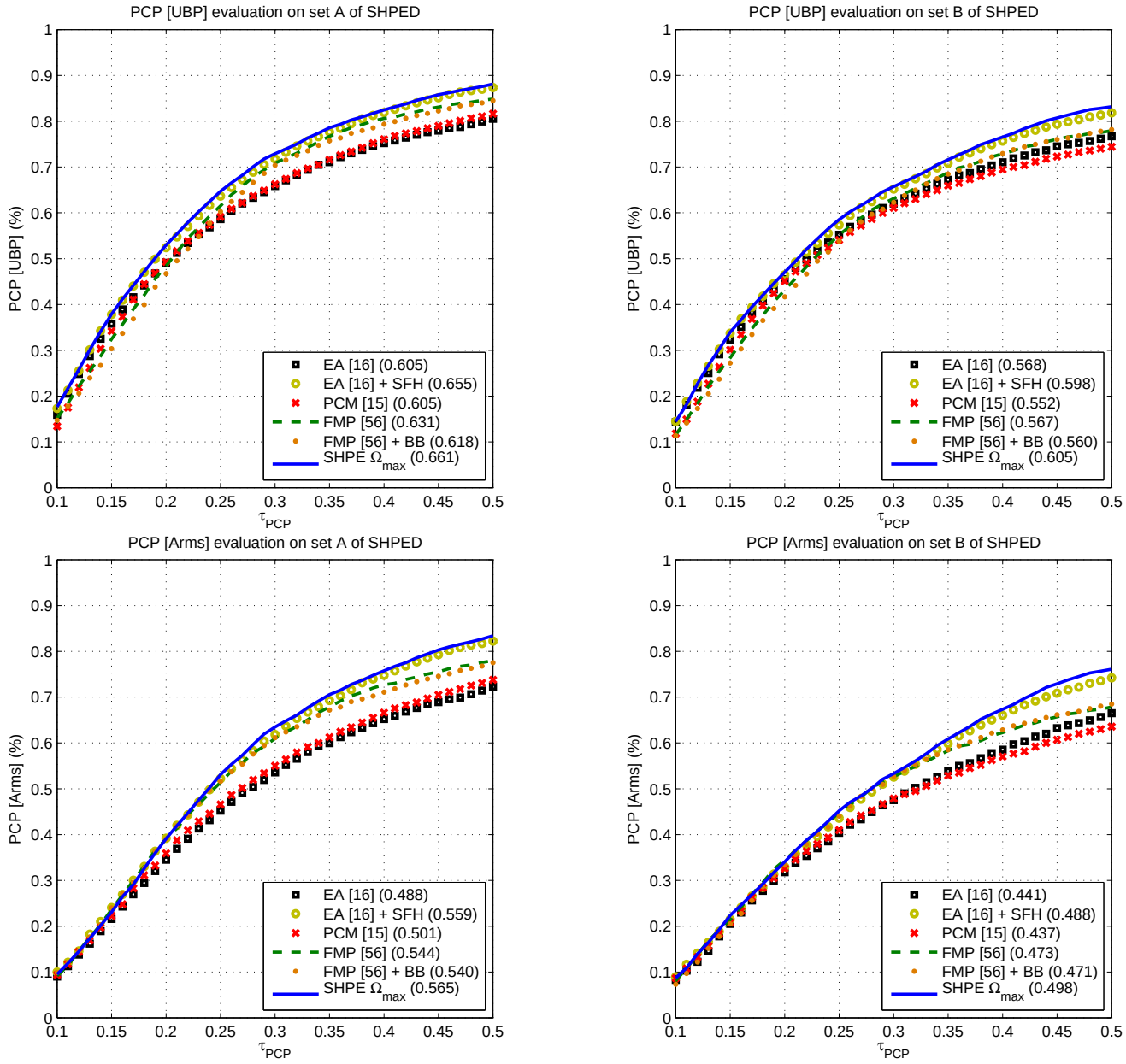


Fig. 12 Comparison of PCP curves for stereo pose estimation on sets A and B of SHPED. The performance of the framework of Eichner *et al.*(EA) [16] is shown in this figure as a baseline (black squares). The remaining curves represent our proposal (solid blue line) and the competitor methods. Additionally, AUC-PCP values for each method are shown into parentheses. All models are independently run on set A **(left column)** and set B **(right column)** of SHPED, categorized into two groups of upper-body parts: UBP (all upper-body parts) **(top row)** and arm parts **(bottom row)**. On average, the best result is returned by our SHPE Ω_{max}. This is especially relevant in the case of the arms at $\tau_{PCP} = 0.5$.

The simplest model proposed in [15] (Sec. 5.3.1) for co-estimation is used in our comparison (row 'PCE' of Tab.2). However, the results obtained in this case are even lower than the ones achieved with the model in [16] (row 'EA'). In our opinion, although the detection windows should be aligned due to the structure of the upper-body model used for detection, the slight displacements existing between the theoretically corresponding points in the two viewpoints lead to a poor combination of appearance-based potentials in the PS model during the inference.

We also compare with the successful model of Yang and Ramanan [56] (row 'FMP' of Tab. 2). We verify that FMP improves over the baseline in 2.2%. However, our SPS model improves over FMP around 5.7% on the whole upper-body, in terms of AUC-PCP, and a modest 4.3% on the arms. Note that FMP uses an articulated model with finer-grained parts than our model, which allows a more accurate estimation of the arms, but with a significant increase in the computational cost. In addition, their body parts can be stretched and shrunk independently, in contrast to ours that have a com-

Fig. 13 Qualitative results on SHPED with Ω_{max}. Rows 1 to 4 show successful examples, while 5*a* shows an almost successful example, and 5*b* and 5*c* show two examples of failures. Note the variety of image conditions and arm poses where our method works (e.g., **1a–c, 2c, 3c, 4b**). Images are often very cluttered, and a person might cover only a small proportion of the image area **(2b, 4a)**, as they can appear at any scale **(3b)**. Illumination varies over a wide range **(2a)**. Sometimes, there is poor contrast between people and background, preventing the use of background subtraction techniques **(3a, 4c)**. Examining failure cases, we found that our model can be confused by excessive bent arms **(5b)** (both the upper and lower arms almost occupy the same image region), and when the camera is moving fast, causing an intense motion blur **(5c)**.

mon scale given by the scale of the upper-body detection. Comparing row 'FMP' to 'FMP+BB', we can say that FMP does not benefit from the usage of a initial people detection stage to limit its search, as both results are very similar. In Fig. 14, we can visually compare some results obtained by FMP and SHPE models. Note that most cases of stretched out arms are correctly handled by SHPE in contrast to FMP.

In terms of PCP (right most columns of Tab. 2), our SPS model achieves 85.6% at 0.5. This value is superior to the one achieved by both 'EA' (78.6%) and 'FMP' (81.4%). Focusing on arms, FMP and SPS behaves quite similar at the strict operating point 0.2.

Finally, in Fig. 12, we can visually observe that *(i)* the curve corresponding to SPS model is in general above the other baseline methods (i.e. better PCP); *(ii)* FMP offers

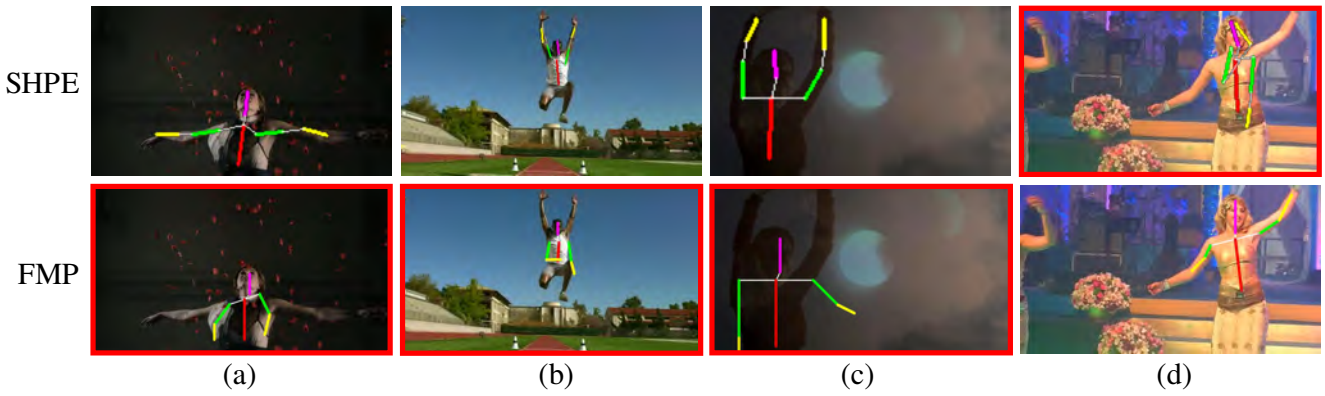


Fig. 14 Qualitative comparison of SHPE versus FMP. **(Top)** Poses obtained by our method. **(Bottom)** Poses obtained by FMP model. Wrong estimations are marked with a red border.

better estimates than both EA and PCE; and, *(iii)* the pose estimation in set B is more difficult than in set A.

5.5 Computational time

We show here a breakdown of the computational time of the different stages of our proposed algorithm given a target upper-body window. The implementation has been done using as basis the source code of HPE framework of Eichner *et al.* [16] released by its authors. Therefore, most code is written in Matlab with a few mex functions. The non parallelized and unoptimized code has been run on a Linux desktop (Ubuntu 12.04 LTS) with 6 GB of RAM and a CPU at 3.4 GHz. On average, for a standard enlarged image window of size 160×140, the Stereo Foreground Highlighting stage takes 2.2 seconds and the parsing stage with the Stereo Pictorial Structure model takes 8.6 seconds.

In addition, the image rectification carried out on each pair of video frames (with a frame size of 1280×720 pixels) takes, on average, around 1.8 seconds. Note that this is not performed per person but per frame. The estimation of the parameters of the transformation is computed once per sequence and takes, on average, 16.6 seconds.

6 Conclusions

This work has presented a novel technique to automatically estimate the 2D human pose of upper-bodies in stereo image pairs extracted from realistic stereo videos. Our proposal extends the monocular Eichner *et al.*'s framework in three ways: *(i)* an adapted people detection and grouping approach for sequences of stereo pairs; *(ii)* a new stereo foreground-highlighting algorithm to segment people by using disparity maps; and, *(iii)* a new Stereo Pictorial Structure model that runs over the two images to find the single most likely upper-body pose. In order to test the proposed method, the Stereo Human Pose Estimation Dataset has been created with ground-truth annotations.

The results obtained in our dataset show that our proposal compares favourably with state-of-the-art techniques such as [16] and [56]. Our method merges the information from the two images obtaining a better approximation of the upper body pose than monocular HPE techniques that run independently on each image.

Finally, it must be indicated that although the proposed approach has been defined and evaluated on upper-bodies, it could be easily extended to full-bodies.

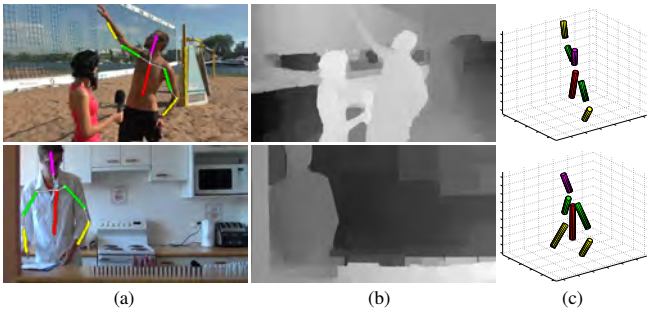


Fig. 15 Qualitative example of 3D estimation from 2D stickman and disparity. **(a)** Estimated 2D pose. **(b)** Estimated disparity [7]. **(c)** Proposed 3D pose obtained by approximating the depth of the limbs from the estimated disparity.

As future work, we plan to exploit the stereo information in such a way that coarse 3D information of the pose can be recovered. In Fig. 15 we show a qualitative example of an early experiment, where the disparity information is used to propose Z coordinates for the estimated 2D limbs.

Acknowledgements

This work was partially supported by the Research Projects TIN2012-32952 and BROCA, both financed by the Spanish

Ministry of Science and Technology and FEDER. We also thank the invaluable help of Marcin Eichner during the implementation of his Direct-PCE method.

References

1. Agarwal, A., Triggs, B.: Recovering 3D human pose from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(1), 44–58 (2006)
2. Agarwal, S., Awan, A., Roth, D.: Learning to detect objects in images via a sparse, part-based representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11), 1475–1490 (2004)
3. Amin, S., Andriluka, M., Rohrbach, M., Schiele, B.: Multi-view pictorial structures for 3D human pose estimation. In: Proceedings of the British Machine Vision Conference. Bristol, UK (2013)
4. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1014–1021 (2009)
5. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 623–630 (2010)
6. Andriluka, M., Roth, S., Schiele, B.: Discriminative appearance models for pictorial structures. International Journal of Computer Vision **99**(3) (2012)
7. Ayvaci, A., Raptis, M., Soatto, S.: Sparse occlusion detection with optical flow. International Journal of Computer Vision **97**(3), 322–338 (2012)
8. Baak, A., Müller, M., Bharaj, G., Seidel, H.P., Theobalt, C.: A data-driven approach for real-time full body pose reconstruction from a depth camera. In: Proceedings of the International Conference on Computer Vision, pp. 1092–1099. IEEE (2011)
9. Bay, H., Ess, A., Tuytelaars, T., van Gool, L.: SURF: Speeded up robust features. Computer Vision and Image Understanding pp. 346–359 (2008)
10. Buehler, P., Everingham, M., Huttenlocher, D.P., Zisserman, A.: Long term arm and hand tracking for continuous sign language TV broadcasts. In: Proceedings of the British Machine Vision Conference, pp. 110.1–110.10 (2008)
11. Burenius, M., Sullivan, J., Carlsson, S.: 3d pictorial structures for multiple view articulated pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3618–3625 (2013)
12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893 (2005)
13. Eichner, M., Ferrari, V.: Better appearance models for pictorial structures. In: Proceedings of the British Machine Vision Conference, pp. 3.1–3.11 (2009)
14. Eichner, M., Ferrari, V.: We are family: Joint pose estimation of multiple persons. In: Proceedings of the European Conference on Computer Vision, pp. 228–242 (2010)
15. Eichner, M., Ferrari, V.: Human pose co-estimation and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2282–2288 (2012)
16. Eichner, M., Marín-Jiménez, M.J., Zisserman, A., Ferrari, V.: 2D articulated human pose estimation and retrieval in (almost) unconstrained still images. International Journal of Computer Vision **99**(2), 190–214 (2012)
17. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. International Journal of Computer Vision **88**(2), 303–338 (2010)
18. Feltzenswalb, P., Hutenlocher, D.: Pictorial structures for object recognition. International Journal of Computer Vision **61**, 55–79 (2005)
19. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9) (2010)
20. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2008)
21. Ferrari, V., Marin, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2008)
22. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Pose search: retrieving people using their pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2009)
23. Guan, P., Weiss, A., Balan, A., Black, M.J.: Estimating human shape and pose from a single image. In: Proceedings of the International Conference on Computer Vision, pp. 1381–1388 (2009)
24. Guo, F., Qian, G.: Human pose inference from stereo cameras. In: IEEE Workshop on Applications of Computer Vision, pp. 37–37 (2007)
25. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, second edn. Cambridge University Press, ISBN: 0521540518 (2004)
26. Johnson, S., Everingham, M.: Combining discriminative appearance and segmentation cues for articulated human pose estimation. In: ICCV Workshops: Machine Learning for Vision-based Motion Analysis (2009)
27. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: Proceedings of the British Machine Vision Conference, pp. 12.1–11 (2010)
28. Kazemi, V., Burenius, M., Azizpour, H., Sullivan, J.: Multi-view body part recognition with random forests. In: Proceedings of the British Machine Vision Conference, pp. 48.1–48.11 (2013)
29. Konolige, K.: Small vision systems: Hardware and implementation. In: Y. Shirai, S. Hirose (eds.) Robotics Research, pp. 203–212. Springer London (1998)
30. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the International Conference on Machine Learning, pp. 282–289 (2001)
31. Lallemand, J., Szczot, M., Ilic, S.: Human pose estimation in stereo images. In: Articulated Motion and Deformable Objects, pp. 10–19 (2014)
32. Lan, X., Huttenlocher, D.: Beyond trees: Common-factor models for 2D human pose recovery. In: Proceedings of the International Conference on Computer Vision, vol. 1, pp. 470–477 (2005)
33. Lee, M., Cohen, I.: Human upper body pose estimation in static images. In: Proceedings of the European Conference on Computer Vision, pp. 126–138 (2004)
34. Mori, G., Ren, X., Efros, A., Malik, J.: Recovering human body configurations: Combining segmentation and recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. II–326–II–333 (2004)
35. Pérez-Sala, X., Escalera, S., Angulo, C., González, J.: A survey on model based approaches for 2D and 3D visual human pose recovery. Sensors pp. 4189–4210 (2014)
36. Pons-Moll, G., Taylor, J., Shotton, J., Hertzmann, A., Fitzgibbon, A.: Metric regression forests for human pose estimation. In: Proceedings of the British Machine Vision Conference, pp. 4.1–4.11 (2013)
37. Ramanan, D.: Learning to parse images of articulated bodies. In: Advances in Neural Information Processing Systems, pp. 1129 – 1136. MIT Press (2006)

38. Rogez, G., Rihan, J., Orrite-Uruñuela, C., Torr, P.H.: Fast human pose detection using randomized hierarchical cascades of rejectors. International Journal of Computer Vision **99**(1), 25–52 (2012)
39. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. In: ACM Transactions on Graphics (TOG), vol. 23, pp. 309–314 (2004)
40. Sapp, B., Jordan, C., Taskar, B.: Adaptive pose priors for pictorial structures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 422–429 (2010)
41. Sapp, B., Toshev, A., Taskar, B.: Cascaded models for articulated pose estimation. In: Proceedings of the European Conference on Computer Vision, pp. 406–420 (2010)
42. Schwarz, L.A., Mkhitaryan, A., Mateus, D., Navab, N.: Human skeleton tracking from depth data using geodesic distances and optical flow. Image and Vision Computing **30**(3), 217 – 226 (2012)
43. Sheasby, G., Valentin, J., Crook, N., Torr, P.: A robust stereo prior for human segmentation. In: Proceedings of the Asian Conference on Computer Vision, pp. 94–107 (2012)
44. Sheasby, G., Warrell, J., Zhang, Y., Crook, N., Torr, P.H.: Simultaneous human segmentation, depth and pose estimation via dual decomposition. In: British Machine Vision Conference, Student Workshop, BMVW (2012)
45. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from a single depth image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1297–1304 (2011)
46. Sigal, L., Black, M.: Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2041–2048 (2006)
47. Sigal, L., Isard, M., Haussecker, H., Black, M.J.: Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. International Journal of Computer Vision **98**(1), 15–48 (2012)
48. Smolic, A., Mueller, K., Merkle, P., Kauff, P., Wiegand, T.: An overview of available and emerging 3D video formats and depth enhanced stereo as efficient generic solution. In: Picture Coding Symposium, pp. 1–4. IEEE (2009)
49. Sun, M., Kohli, P., Shotton, J.: Conditional regression forests for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3394–3401 (2012)
50. Thang, N.D.: Human pose and activity recognition from stereo images using probabilistic parametric inference. Ph.D. thesis, Kyung Hee University, Department of Computer Engineering (2011)
51. Tian, Y., Sigal, L., la Torre, F.D., Jia, Y.: Canonical locality preserving latent variable model for discriminative pose inference. Image and Vision Computing **31**(3), 223 – 230 (2013)
52. website: CALVIN Upper Body Detector. `http://www.vision.ee.ethz.ch/~calvin/calvin_upperbody_detector/` (2010). Last visit: May 2014
53. website: FMP software for HPE. `http://www.ics.uci.edu/~dramanan/software/pose/` (2012). Last visit: May 2014
54. website: HPE software. `http://www.vision.ee.ethz.ch/~calvin/articulated_human_pose_estimation_code/` (2012). Last visit: May 2014
55. Yang, H.D., Lee, S.W.: Reconstruction of 3D human body pose from stereo image sequences based on top-down learning. Pattern Recognition **40**(11), 3120 – 3131 (2007)
56. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1385–1392 (2011)
57. Yao, A., Gall, J., Van Gool, L.: Coupled action recognition and pose estimation from multiple views. International Journal of Computer Vision **100**(1), 16–37 (2012)
58. Ye, M., Wang, X., Yang, R., Ren, L., Pollefeys, M.: Accurate 3d pose estimation from a single depth image. In: Proceedings of the International Conference on Computer Vision, pp. 731–738. IEEE (2011)
59. Yeguas-Bolivar, E., Munoz-Salinas, R., Medina-Carnicer, R., Carmona-Poyato, A.: Comparing evolutionary algorithms and particle filters for markerless human motion capture. Applied Soft Computing **17**, 153–166 (2014)
60. Zhu, Y., Dariush, B., Fujimura, K.: Controlled human pose estimation from depth image streams. In: IEEE Computer Vision and Pattern Recognition Workshops (2008)
61. Zhu, Y., Fujimura, K.: Constrained optimization for human pose estimation from depth sequences. In: Proceedings of the Asian Conference on Computer Vision, pp. 408–418 (2007)
62. Zolfaghari, M., Jourabloo, A., Gozlou, S., Pedrood, B., Manzuri-Shalmani, M.: 3D human pose estimation from image using couple sparse coding. Machine Vision and Applications **25**, 1489–1499 (2014)
63. Zuffi, S., Freifeld, O., Black, M.J.: From pictorial structures to deformable structures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3546–3553 (2012)