Mixing Body-Parts Model for 2D Human Pose Estimation in Stereo Videos

Manuel I. López-Quintero[1], Manuel J. Marín-Jiménez[1,2], Rafael Muñoz-Salinas[1,2,*], Rafael Medina-Carnicer[1,2]

[1]Dept. Computing and Numerical Analysis, University of Córdoba, Campus de Rabanales, 14071, Córdoba, Spain
[2]Maimonides Institute for Biomedical Research (IMIBIC), University of Córdoba, 14004, Córdoba, Spain
[*]rmsalinas@uco.es

Abstract: This work targets 2D articulated human pose estimation (*i.e.* localization of body limbs) in stereo videos. Although in recent years depth-based devices (*e.g.* Microsoft Kinect) have gained popularity, as they perform very well in controlled indoor environments (*e.g.* living rooms, operating theatres or gyms), they suffer clear problems in outdoor scenarios and, therefore, human pose estimation is still an interesting unsolved problem. We propose here a novel approach that is able to localize upper-body keypoints (*i.e.* shoulders, elbows, wrists) in temporal sequences of stereo image pairs. Our method starts by locating and segmenting people in the image pairs by using disparity and appearance information. Then, a set of candidate body poses is computed for each view independently. Finally, temporal and stereo consistency is applied to estimate a final 2D pose. We validate our model on three challenging datasets: "Stereo Human Pose Estimation Dataset", "Poses in the Wild" and "INRIA 3DMovie". The experimental results show that our model not only establishes new state-of-the-art results on stereo sequences, but also brings improvements in monocular sequences.

1. Introduction

Given an image or set of images, the goal of 'Human Pose Estimation' (HPE) is to obtain the spatial location of human body limbs in the target images. The output of HPE can be useful for higher level recognition tasks as, for example, human activity recognition[1, 2] or pose-based video retrieval[3]. Most of the state-of-the-art approaches focus on monocular images, however, in recent years the so called '3D cameras' have lowered in price, increasing its adoption at consumer level. Therefore, the amount of video sequences with stereo information has grown rapidly in popular video hosting websites as YouTube. In parallel, devices that record depth information (*e.g.* Microsoft Kinect) are broadly used for human-machine interaction (HMI). This kind of HMI has led to the development of several remarkable techniques based on depth information [4, 5]. Nevertheless, Kinect-like devices are known not to work properly outdoors due to sunlight. In addition, the working distance is limited, which makes difficult its use in some applications. Therefore, the goal of this work is to estimate human body parts in stereo video from 3D cameras by taking advantage of both color and disparity information. The main idea of our approach is illustrated in Fig. 1: people segmentation by using disparity and appearance information; estimation of body parts per view by

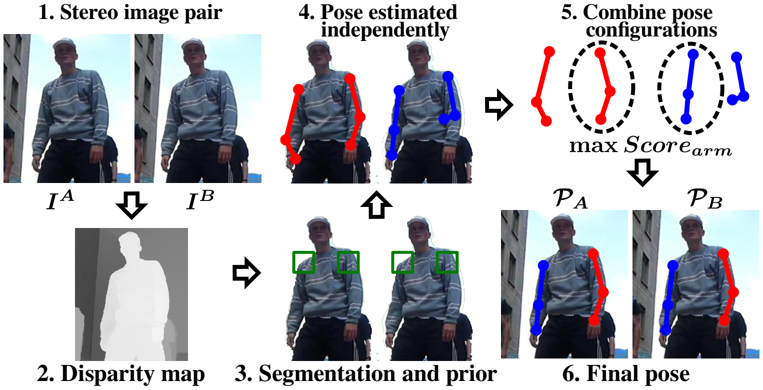


Fig. 1. *Given a stereo video sequence, the goal of this work is to localize the position and orientation of human body parts. Disparity and temporal information is used to select and combine the most stable configurations along the sequence. (Best viewed in colour.)*

using a shoulder prior; and, final combination of body parts.

The main contributions of our paper can be summarized as follows: *(i)* a full-pipeline to estimate human poses in stereo pairs with multiple people; *(ii)* a set of steps to reduce the search space of body parts in stereo images; and, *(iii)* two new models to combine body parts in stereo sequences.

A thorough experimental study carried out on three state-of-the-art datasets (*i.e.* 'INRIA 3DMovie', 'Poses In the Wild', 'Stereo Human Pose Estimation Dataset') shows that our proposal improves consistently on previously published results on the considered datasets.

The rest of the paper is organized as follows. We start by reviewing related work in Sec. 2. Our approach for estimating body parts in stereo sequences is described in Sec. 3. Then, Sec. 4 contains the experiments and results. Finally, we present the conclusions in Sec. 5.

2. Related Work

We summarize in this section the HPE works that are most closely related to our approach. For an extensive summary on HPE, we refer the reader to the excellent survey presented by Pérez-Sala *et al.* [6].

2.1. Color-based monocular approaches.

The work of Yang and Ramanan[7] defines a deformable part model for estimating human poses in static images. They use a local mixture of non-oriented parts to model body articulation. One extension of the previous work is the proposal of Chen and Tan [8], where they use local contextual information to reduce the noise and non-local contextual information to detect the leaf parts with more accuracy. Another relevant extension of the method of Yang and Ramanan[7] is the proposal of Cherian *et al.* [9], where they estimate the human pose in videos using temporal links in their model. Thanks to the advances in stereo consistency [10, 11] we propose here novel ideas that allow to improve the part recombination model on stereo videos by using disparity information in two ways: for people segmentation (*i.e.* removing background pixels) and for imposing stereo matching between the per-view estimated poses. In addition, we show that the recombination-based monocular case can be improved by removing background pixels based on appearance information [12, 13, 14, 15, 16] and by imposing a new shoulder prior given a person detection window. Recently, remarkable results have been obtained with the method of Pfister *et al.* [17] for HPE, referred as F-CNN. This method, which is based on the recently successful Convolutional Neural Networks, assumes a square input image where just one person is depicted.

2.2. Depth-based monocular approaches.

Concerning human pose estimation from depth images, Shotton *et al.* [4] propose a 3D joint model in a per-pixel classification problem. To take advantage of temporal consistency, Sun *et al.* [5] introduce a global latent variable associated to torso orientation or person height which increases the accuracy of body joint location prediction. Convolutional neural networks are used in the work of Jiu *et al.* [18] for the estimation of human poses from depth information. In contrast to many popular approaches, they do not include pairwise pixel terms in their energy function to reduce computation times.

2.3. Multicamera approaches.

The method of Shen *et al.* [19] estimates the human pose given several camera views. They work on 3D voxel reconstructions computed from 2D foreground silhouettes instead of image data. Template fitting is used to predict the head and torso locations from these 3D voxels and, then, a hierarchical fitting method estimates the remaining body parts. Burenius *et al.* [20] use 3D pictorial structures to estimate 3D pose of humans from images obtained from multiple calibrated cameras. To address the problem of discretization of the search space, they establish view, skeleton, and joint angle constraints from different views to define a discrete search grid. In addition, one notable use of the proposed discrete search grid from multiple and calibrated cameras is the work of Kazemi *et al.* [21].

2.4. Stereo approaches.

Seguin *et al.* [22] introduce a graphical model for joint segmentation and pose estimation of multiple people in stereo video sequences. In addition, a new dataset named 'INRIA 3DMovie Dataset', built from sequences of 3D movies, is delivered for evaluating their proposal. In the work of López-Quintero *et al.* [23], the authors propose the Stereo Pictorial Structure for human pose estimation in stereo pairs, where disparity information is used to obtain common poses between pairs of images. They validate their approach on stereo images extracted from videos hosted in YouTube, delivering a new dataset named 'Stereo Human Pose Estimation Dataset'. In contrast to this previous work, which deals with isolated stereo image pairs (*i.e.* not a joint estimation for the whole video sequence), in this work, among other improvements (*e.g.* multimodal segmentation), we take advantage of the temporal information provided by stereo sequences to improve the previously published results on the dataset.

3. Proposed approach

We describe in this section the pipeline we follow for estimating 2D poses in stereo video sequences. The first step consists in estimating the rectification parameters so as to obtain good results from stereo block-matching algorithms. For that purpose, the uncalibrated stereo image rectification method available online at Mathworks' website[1] is employed. In essence, it collects

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

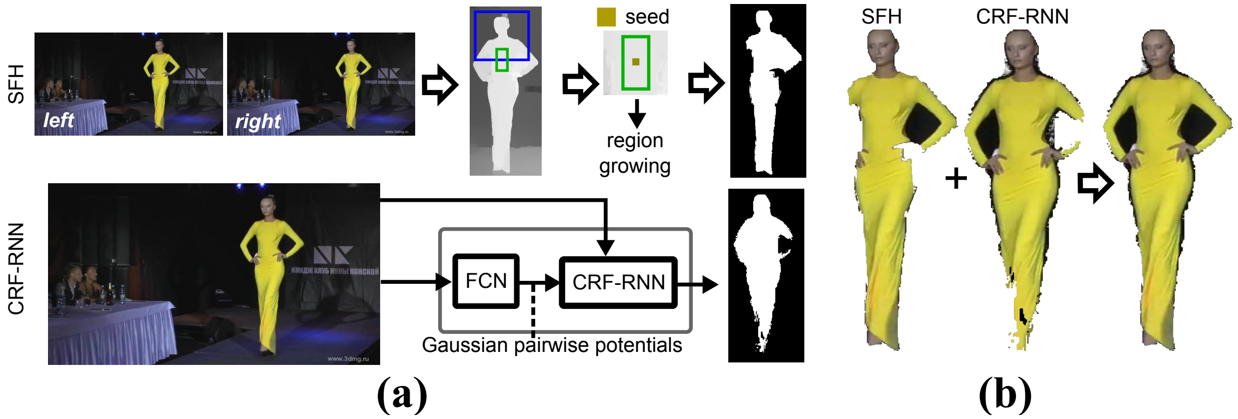


Fig. 2. *Segmentation scheme (best viewed in colour)*
a Two segmentation methods are applied: one based on disparity (SFH) and other based on appearance (CRF-RNN)
b To prevent missing foreground pixels, a logical OR operator is applied to the obtained binary masks

interest points from image pairs (using SURF) and then finds putative correspondences filtered by epipolar constraints. Correspondences are used to compute the rectification transform producing proper horizontal alignment. This process is required only once per video (using the first frame) and then applied to the remaining frames.

Then, given an image window containing a single person, *i.e.* returned by a generic person detector (Sec. 3.1), the steps of the pipeline we propose for estimating the 2D pose are: (*i*) to remove background pixels by using disparity and appearance information (Sec. 3.2); (*ii*) to estimate body poses independently in each image of the stereo pair sequences and, then, (*iii*) to combine the independently estimated body pose sequences into a common single sequence by imposing constraints given by the stereo data (Sec. 3.3) and by location priors (Sec. 3.4).

In the following subsections, we present and discuss the novelties introduced in each stage of the proposed pipeline.

3.1. People detection

The first step of our method is to delimit the image region where the target person is located, allowing our method to deal with images containing multiple people, in contrast to other approaches [9]. For that purpose, we run the recent generic person detector proposed in the work of Shaoqing *et al.* [24], ‘Faster R-CNN’, which is based on Convolutional Neural Networks and the use of region proposals to hypothesize the location of the objects. The output of the detector is a set of rectangular bounding-boxes (BB). In our case, each BB is slightly enlarged with the idea of covering diverse arm configurations. The subsequent stages of the HPE pipeline will be performed only within the enlarged image window. In addition, we run within the image window the person detector of Eichner *et al.* [3] to localize the upper-body (UB) of people in images. This will be used for initializing the depth-based segmentation (Sec. 3.2) and for defining priors on shoulders keypoint location (Sec. 3.4).

3.2. Foreground highlighting

We adopt here the term 'foreground highlighting', coined by Ferrari *et al.* [13], to designate the process of removing background pixels from the target image window to limit the search space of the body limbs in the image. We propose here two complementary approaches for that purpose. The first one is based on the estimated disparity per pixel. The idea is to apply a thresholding method on the disparity map to remove as many background pixels as possible. For that purpose, we follow the Stereo Foreground Highlighting (SFH) method developed in previous works [23]. The method employs a region growing algorithm on the estimated disparity map in which the seed is placed in the center of a prior region on the torso, given by an upper-body detector (Sec. 3.1). It is assumed that the disparity values of the person body follows a normal distribution, estimating the mean and variance values from the predefined region. An example of application of this method can be seen at the top of Fig. 2.(a). The output is a binary mask where foreground pixels are *'on'* (*i.e.* white) and the background pixels are *'off'* (*i.e.* black). The second approach proposed to reduce the search space is based on color information. The method of Zheng *et al.* [16] allows to assign a class label (from a set of predefined ones) to each pixel in the target image window. This method, named CRF-RNN, builds on top of a Convolutional Neural Network, combined with a Conditional Random Field. In our case, we are interested in the pixels automatically labelled by the method as *'person'*. We show in Fig. 2.(a) (bottom) a typical case of application of this method. If we pay attention to the output of both methods in Fig. 2, we notice that SFH has removed the right arm of the person. Whereas CRF-RNN has removed part of the left arm. However, the desired solution would be the combination of both segmentation masks. Therefore, we combine both segmentation masks by using the logical inclusive *OR* operator, *i.e.* the union of both sets of foreground pixels, that it is applied to each pair of corresponding pixels. The final result can be seen in Fig. 2.(b). Note that this choice is quite conservative, with the idea of removing the minimum amount of pixels belonging to the person. Using *AND* operator would be an alternative to eliminate more background pixels. However, in practice, there are many situations, as in Fig. 2.(b), where some of the pixels of body parts are labeled incorrectly and the final segmentation would have removed those body parts. Note that the removed foreground pixels cannot be recovered in the following stages.

3.3. Recombining body-part sequences in stereo videos

We start by summarizing the 'Mixing Body-Parts' method (MBP) of Cherian *et al.* [9] which was proposed for monocular videos. Then, we introduce and discuss our stereo model.

Monocular sequences. Let $\mathcal{I} = (I_1, I_2, ..., I_T)$ be a video sequence of length T, where I_i represents a video frame (*i.e.* image). The goal of MBP is to estimate the body parts (*i.e.* head, shoulders, elbows and wrists) of the human bodies detected in the video sequence (Fig. 3.(a)). For that purpose, a graphical model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined, where $\mathcal{V}$ are the vertices (*i.e.* body parts) and $\mathcal{E}$ are the edges (*i.e.* connections between pairs of body parts). A pose p with regard to $\mathcal{G}$ is defined as a set of 2D coordinates representing the positions of the body parts in the image: $p = \{p^u = (x^u, y^u) \in \mathbb{R}^2 : \forall u \in \mathcal{V}\}$. This formulation leads to the following cost function $C(I, p)$ that has to be minimized in order to estimate the body pose:

$$C(I, p) = \sum_{u \in \mathcal{V}} \phi_u(I, p^u) + \sum_{(u,v) \in \mathcal{E}} \psi_{u,v}(p^u - p^v), \qquad (1)$$

where $\phi_u(I, p^u)$ is an appearance term for the body part u at the position p^u, and $\psi_{u,v}(p^u - p^v)$ is a

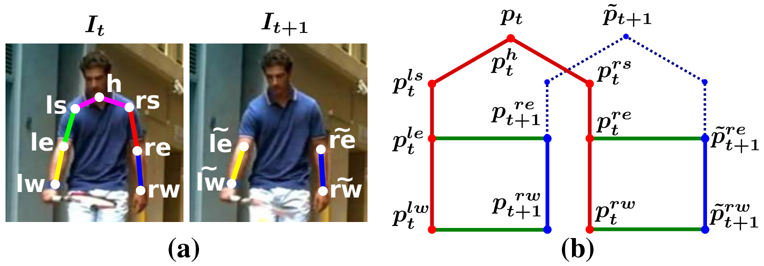


Fig. 3. *Graphical model of Mixing Body-Part Sequences (best viewed in colour)*
a Estimation of body parts (*i.e.* (h)ead, (l)eft and (r)ight (s)houlders, (e)lbows and (w)rists) of I_t using temporal links on I_{t+1} of elbows and wrists
b Graphical model connecting keypoints in two consecutive frames (I_t, I_{t+1})

deformation cost for body parts (u, v). These terms have been adopted from the work of Yang and Ramanan [7]. In addition to the edges defined between body parts, and in order to impose temporal consistency of the body poses, a temporal edge between pairs of parts p_t^u and p_{t+1}^u is introduced. Therefore, the new cost function for the whole video sequence is rewritten as:

$$C(I_T, p_T) + \sum_{t=1}^{T-1} C(I_t, p_t) + \lambda_1 \theta(p_t, p_{t+1}, I_t, I_{t+1}), \tag{2}$$

where λ_1 is a regularization parameter and function θ measures consistency between the poses in two consecutive frames. In particular, θ is defined as:

$$\theta(p_t, p_{t+1}, I_t, I_{t+1}) = \sum_{u \in \mathcal{V}} ||p_{t+1}^u - p_t^u - f_t(p_t^u)||_2^2, \tag{3}$$

where $f_t(p_t^u)$ corresponds to the optical flow between frames I_t and I_{t+1} at position p_t^u. As the proposed graphical model contains loops and, therefore, is a hard – and sometimes intractable – inference problem, they propose both a simplification of the model (see Fig.3.(b)) and a two-stage approach: (*i*) to generate a set of candidate poses in each frame; and, (*ii*) to decompose the candidate poses into limbs, generating limb sequences along time, and then, to recompose the complete pose by mixing those body part sequences.

Stereo sequences. Let $\mathcal{S} = (\mathcal{I}^A, \mathcal{I}^B)$ be a stereoscopic video sequence, where $\mathcal{I}^A = (I_1^A, I_2^A, ..., I_T^A)$ and $\mathcal{I}^B = (I_1^B, I_2^B, ..., I_T^B)$ correspond to the left and right image sequences of the stereo pair, respectively. The goal here is to find a common body pose at time t for each image of the stereo pair. A given body part $p_A^u = (x_A^u, y_A^u)$ in the left image will be related to the body part $p_B^u = (x_B^u, y_B^u)$ in the right image by the disparity δ at that position: $p_B^u = (x_A^u + \delta, y_A^u)$. We discuss below two ways of using the image pairs in order to improve the body part estimation along the video sequence.

3.3.1. Stereo Limb Recombination: Given a stereo video sequence, we generate for each stereo view in frame t a set of K candidate poses $\mathcal{P}_t^A$ and $\mathcal{P}_t^B$ by using the n-best algorithm of Park and Ramanan [25]. Then, in order to recombine the pose candidates from each view from the stereo image pair, we propose to combine the best candidates of each view to obtain a new set $\mathcal{P}_t = \left\{ \mathcal{P'}_t^A \bigcup (\mathcal{P'}_t^B + \delta) \right\}$, where $\mathcal{P}_t$ is the union of the the best $K/2$ candidate poses from each view of the stereo pair in frame t; and $\mathcal{P'}_t^A$ and $\mathcal{P'}_t^B$ are the corresponding subsets of best poses. We

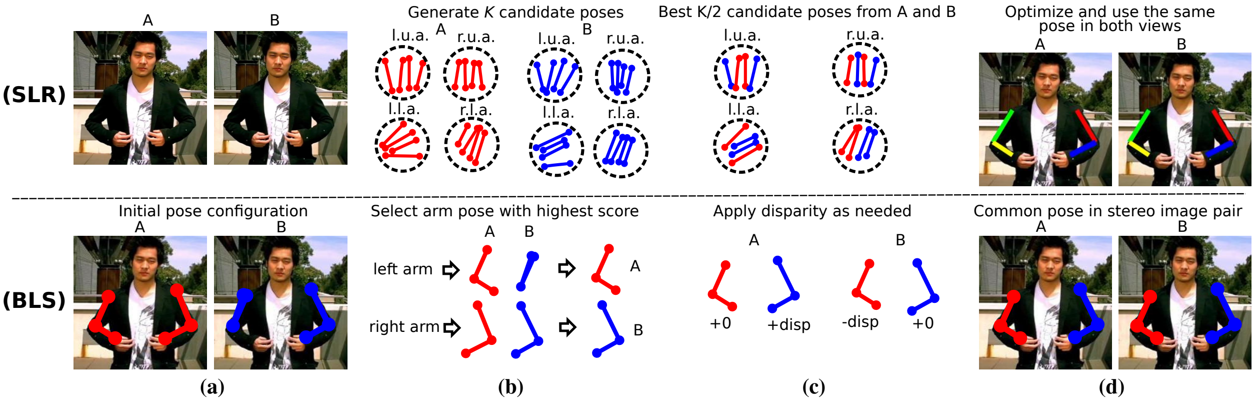

(SLR)
A
B
Generate K candidate poses
A
B
l.u.a.
r.u.a.
l.l.a.
r.l.a.
Best K/2 candidate poses from A and B
Optimize and use the same pose in both views
(BLS)
Initial pose configuration
Select arm pose with highest score
left arm
right arm
Apply disparity as needed
+0
+disp
-disp
Common pose in stereo image pair
(a)
(b)
(c)
(d)

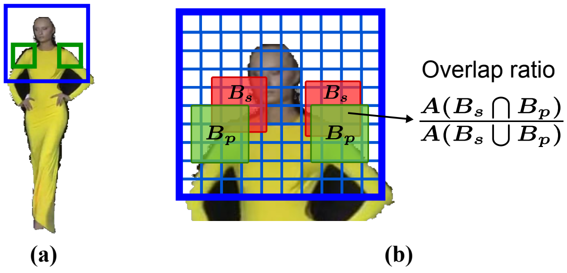
Fig. 4. *Stereo mixing body parts (best viewed in colour)*
top Stereo Limb Recombination: candidate limb poses from each stereo view are combined before global optimization
bottom Best Limb Score: limbs from each stereo view are combined after optimizing each view independently

use the disparity value δ to shift the x coordinates of $\mathcal{P}_t^B$, adjusting the horizontal displacement. Once the best candidate poses from each view are combined into the new set, we use dynamic programming to minimize the cost function described in Eq. 2. Finally, we apply the recombination of limbs proposed for MBP [9] and, then, we correct the shift done in B by subtracting the disparity values added previously. We coin this method ‘Stereo Limb Recombination’ (SLR). It is graphically summarized in Fig. 4 (top).

3.3.2. Best Limb Score: We focus now on the arms, where most variability is found. After running the full monocular procedure on each view, an alternative way of taking advantage of the stereo information is to select the candidate locations with the highest scores from each view. The idea here is to independently select the best configuration of these limbs in each view and to combine them together into a single configuration for the stereo pair. For that purpose, for each limb we select from the left and right views the proposed $\mathcal{P}_{arm}$ with the highest score. Finally, for each view we shift the x coordinate of each $\mathcal{P}_{arm}$ with the corresponding disparity value.This method, coined ‘Best Limb Score’ (BLS), is graphically summarized in Fig. 4 (bottom).

3.4. Keypoint priors

The BB returned by the UB detector allows us to constrain the spatial areas where ‘shoulders’ joints should be located. We set two rectangular BB priors located close to the bottom corners of this BB as seen in Fig. 5.(a). The size of those BB has been estimated as approximately 0.3 the size of the UB. Such prior information is transferred to the estimation process by computing a score between the automatically proposed position and the prior one. In particular, we compute the overlap ratio υ between one BB defined from the estimated position (B_s) and other BB defined from the prior (B_p) as follows: $\upsilon = \frac{area(B_s \cap B_p)}{area(B_s \cup B_p)}$, where B_s and B_p correspond to the estimations and the shoulder prior, respectively. The value of υ is in $[0, 1]$ (*i.e.* value 1 means perfect overlap). Finally, we add this value υ to the score (the greater the better) of a pose candidate p, returned by the n-best algorithm of Park and Ramanan [25], increasing its probability of being selected. An example of this overlap can be seen in Fig. 5.(b).



(a) **(b)**

Fig. 5. *Shoulder prior (best viewed in colour)*
a Given a bounding box of the upper-body detector, we set two shoulder priors (green)
b We overlap our prior bounding boxes B_p with the estimation boxes B_s and calculate the overlap ratio

4. Experimental Results

We describe here the datasets and experiments we carry out to validate our approach for HPE on stereo videos. For the sake of clarity, we summarize the main acronyms used in this paper in Tab. 1.

4.1. Datasets

We run our experiments on three state-of-the-art challenging datasets for 2D human pose estimation. The details of each one are given below.

4.1.1. SHPED: 'Stereo Human Pose Estimation Dataset' (SHPED), released by López-Quintero *et al.* [23], consists of 630 stereo image pairs (*i.e.* 1260 images) classified into 42 video clips of 15 frames each. The stereo image pairs were extracted from 26 stereoscopic videos, downloaded from YouTube (found by using the 3D tag filter). The dataset is suitable for pose estimation of multiple people depicted in a stereo image pair. In general, all upper-body parts are almost visible, and the viewpoint of the body is non-profile. SHPED includes 1470 stickmen upper-body annotations belonging to 49 persons. In addition, we have added body keypoint annotations (*i.e.* shoulders, elbows, wrists). Some examples of this dataset can be seen in Fig. 6.(a).

4.1.2. PIW: 'Poses in the Wild' (PIW) dataset [9] includes 830 single images grouped into 30 sequences of about 30 frames each. This dataset is only prepared for detecting one person per image and hence it contains only 830 annotations. Some examples of this dataset can be seen in the Fig. 6.(b).

4.1.3. INRIA 3DMovie: The 'INRIA 3DMovie' dataset [22] includes 587 annotations of human poses, 1158 bounding boxes, and 686 pixel-wise person's segmentations in total, both for training and test. Particularly, 103 stereo image pairs (*i.e.* 206 images) belong to the test set in the category of pose estimation. These stereo image pairs do not follow a particular temporal sequence and thus the set is composed of still frames. Multiple people are annotated per image pair, therefore, the total amount of annotated poses is 149. Some examples of this dataset can be seen in Fig. 6.(c).

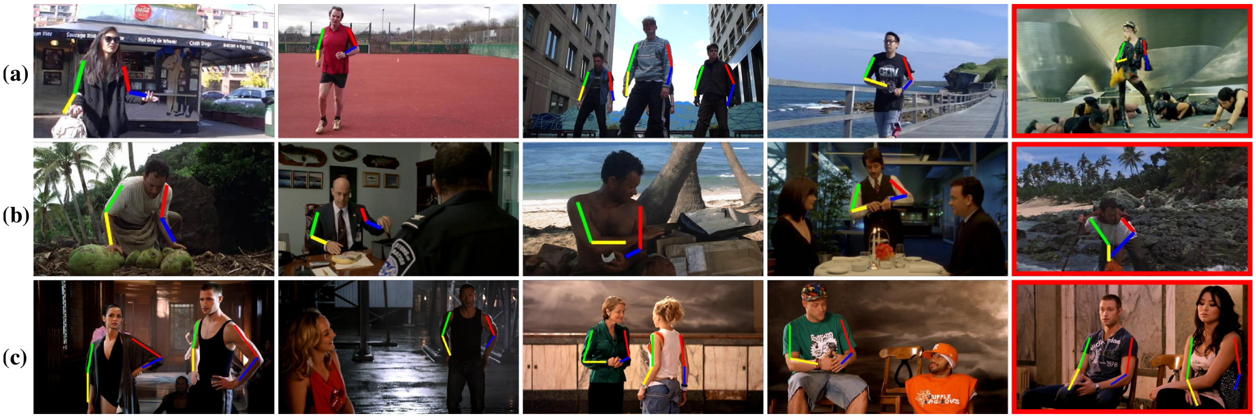


Fig. 6. *Estimated pose on different datasets. The used method is in parentheses; frames with red border contain inaccurate estimations. (Best viewed in colour.)*
a SHPED (‘S + P + SLR’): contains stereo sequences from YouTube videos
b PIW (‘S (Color) + P’): contains monocular video sequences from Hollywood movies
c INRIA 3DMovie (‘S + P + BLS’): contains isolated stereo images from Hollywood movies

Table 1 Summary of the main acronyms used in this paper.

Acronym	*Full name*
F-CNN	Flowing ConvNets*et al.* [17]
FMP	Flexible Mixtures of Parts [7]
HOGcomb	Pose Estimation and Segmentation [22]
MBP	Mixing Body-Parts [9]
PIW	Poses in the Wild dataset [9]
SFH	Stereo Foreground Highlighting [23]
SHPED	Stereo Human Pose Estimation Dataset [23]
SPS	Stereo Pictorial Structure [23]

4.2. Implementation Details

4.2.1. Model parameters: We use the pre-trained appearance terms and deformation costs of Cherian *et al.* [9], available online along with the source code of their method. This will allow a direct comparison with them.

4.2.2. Optical Flow computation: To calculate the optical flow of a sequence of monocular images we use the technique of Brox and Malik[26] as proposed in Cherian *et al.* [9]. This approach integrates correspondences from a keypoint descriptor matching into a variational method. Then, for an accurate estimation of the disparity maps on the stereo pairs, we use the method of Ayvaci *et al.* [27]. Recall that the disparity maps are used both for segmenting people with the SFH method (Sec. 3.2) and for adjusting the horizontal shift of estimated body poses (Sec. 3.3) in a given stereo pair.

4.2.3. Evaluation Metrics: We use two evaluation metrics to compare the different state-of-the-art methods, as not all of them have used the same one. For SHPED and PIW, we apply the *Keypoint Localization Error* (KLE) used in the work of Sapp *et al.* [28]. It computes the percentage of keypoints that fall within a given distance from the ground-truth keypoint. For each symmetric body part (*e.g.* elbow, wrist), we evaluate the average percentage (*avg*) of left and right sides, and the maximum percentage (*max*) of these sides. We show these values as a pair (*avg*, *max*). Note that the *max* version is used in the work of Cherian *et al.* [9], which in our opinion is not fair, as a

Table 2 SHPED results. Comparison of body-part localization accuracy ($\%$) (at 15 pixel error threshold). Each entry indicates (*avg*, *max*), where *avg* is the average accuracy on left and right sides of a body part, and *max* is the maximum accuracy among them. Column Δ corresponds to the increment with regard to MBP+BB.

Method	Shoulders	Elbows	Wrists	*Avg*	Δ
MBP+BB	(74.3, 74.8)	(55.0, 56.5)	(47.9, 52.4)	(59.1, 61.2)	(0.0, 0.0)
S	(74.7, 75.8)	(53.7, 56.8)	(49.0, 57.1)	(59.2, 63.2)	(+0.1, +2.0)
S+BLS	(74.5, 75.3)	(54.3, 58.2)	(49.4, 58.1)	(59.4, 63.9)	(+0.3, +2.7)
S+SLR	(75.1, 75.6)	(54.6, 58.6)	(49.1, 58.6)	(59.6, 64.3)	(+0.5, +3.1)
S+P	(83.0, 84.1)	(59.5, 61.2)	(50.4, 54.5)	(64.3, 66.6)	(+5.2, +5.4)
S+P+BLS	(**83.9**, 85.0)	(**60.5**, 62.0)	(50.9, 55.4)	(65.1, 67.5)	(+6.0, +6.3)
S+P+SLR	(83.6, 84.8)	(59.3, 60.9)	(**53.2**, 59.6)	(**65.4**, 68.4)	(**+6.3**, +7.2)
FMP	(**79.5**, 82.0)	(50.5, 51.8)	(40.3, 41.6)	(56.7, 58.5)	(-2.4, -2.7)
SPS	(67.4, 68.4)	(58.1, 61.4)	(45.6, 48.0)	(57.0, 59.3)	(-2.1, -1.9)
F-CNN+BB	(72.2, 74.8)	(**62.1**, 64.9)	(**57.8**, 60.8)	(**64.1**, 66.8)	(**+5.0**, +5.6)

very good estimation of one of the limbs hides a wrongly estimated one. However, for comparison purposes we have added *max* to our results. For INRIA 3DMovie, we use the *Average Precision of Keypoints* (APK) of Yang and Ramanan[7], as done in the work of Seguin *et al.* [22]. In this case, it is assumed that a keypoint is defined by a bounding box in the ground-truth and a candidate keypoint will be correct if it falls within $\gamma \cdot \max(h, w)$ pixels of the ground-truth keypoint, where γ is a thresholding parameter and h and w are the height and width of the bounding box, respectively. Then, a precision-recall curve can be computed, delivering the average precision.

4.2.4. Computational time: We report here the computational time of the proposed system, which has been implemented on MATLAB, without any paralelization or mex optimization. These measurements have been taken on a laptop with Intel Core i7-5650U processor with 8 GB of RAM memory, on Linux.

The new BLS and SLR stages proposed in this paper take, on average, 2.24s and 3.37s, respectively on sequences of 15 frames. With regard to the optical flow computation, the computational time is around 434s, per image pair. Whereas the estimation of the disparity takes around 675s. Note that much faster implementations and/or algorithms can be currently found for both cases [29, 30], reducing drastically the needed time at those stages.

4.3. Ablative feature analysis

We start by studying the contribution of each proposed stage in the final performance of the system. For this experiment we mainly focus on SHPED, as it contains all the characteristics needed by our system (*i.e.* temporal sequences of stereo image pairs). The top group of rows of Tab. 2 shows the different configurations that we evaluate, where 'S' indicates that people segmentation is used (Sec. 3.2), 'P' indicates that a prior on the shoulders location is used given the person detector (Sec. 3.4), and, 'BLS' and 'SLR' refer to our limb combination methods (Sec. 3.3.1). Note that row 'MBP+BB' refers to the original MBP method, aided with our person detector (Sec. 3.1) to deal with multiple people in the scene (*i.e.* not supported by the original code release). The results indicate that each of the proposed stages brings improvements on the monocular MBP. The biggest improvement, w.r.t MBP, is given by the shoulder prior ($\approx 5\%$), followed by the use of SLR, which boosts the wrist accuracy ($\approx 7\%$).

Although PIW dataset does not contain stereo image pairs, only monocular images from videos, we also study on it the contribution of several of our stages. The results of this study are summarized in Tab. 3. In this case, we can see that our segmentation stage 'S' (in this case using only color, as disparity is not available) brings a small improvement over all the body parts. And, by applying the shoulder prior ('S+P' row) the improvement is even greater ($\approx 3\%$). Therefore, we can conclude that a monocular HPE system based on MBP can be improved by the use of these two proposed stages. Note that although this is a positive finding, we want to recall that the final goal of this paper is stereo video sequences and, therefore, a possible further study on the monocular case is relegated for future work.

Finally, the top rows of Tab. 4 summarize the results of different variants of our method on the image pairs of INRIA 3DMovie. Note that this dataset does not provide temporal sequences, but isolated image pairs. Therefore, the temporal consistency needed by our method cannot be applied. However, for completeness of our study we show the results obtained by using a simplified model. We can notice that, as previously observed in the two other datasets, the shoulder prior brings the biggest improvement w.r.t. MBP ($\approx 12\%$ on average). As temporal information is not available, BLS and SLR do not help to improve much the results. Some qualitative results are shown in Fig. 6. Note how our method returns correct poses in different and challenging scenarios, even with multiple people.

4.4. State-of-the-art Methods

For comparison purposes, we include in Tables 2, 3 and 4 results reported by several state-of-the-art methods on the different datasets. In some cases, we have run the authors' code, if available online. Those methods are located in the bottom rows of each table. Results on SHPED for several values of the KLE metric threshold are summarized in Fig. 7, where the percentage of correctly localized keypoints is represented. In addition, we report for each method the area-under-the-curve, in parentheses. In the case of monocular HPE approaches, since each image of the stereo pair contains its own ground-truth annotations, we evaluate each image of the stereo pair independently. A visual comparison of the results obtained by our method and other state-of-the-art ones is presented in Fig. 8. Each row shows a crop window from a different dataset. Column 'ours' contains our results, indicating the specific combination used, as required by the characteristics (*i.e.* mono/stereo, image/sequence) of the target dataset. Note that bottom row shows an example where our method failed. In that case, during the segmentation stage, part of the wrongly estimated arm was removed from the foreground set and, therefore, those pixels were not considered during the subsequent stages. Such an inaccurate segmentation is probably due to the blurry and low quality of the image.

We indicate below where and how we have applied each comparative method.

4.4.1. Mixing Body-Part Sequences: We use the source code provided by Cherian *et al.* [9], who released PIW dataset, to run their MBP method on the three considered datasets. In cases where multiple people are shown in an image, *i.e.* Tables 2 and 4, we have aided the method with our people detector bounding-boxes. We can see that in all cases, our proposed method improves on it.

Table 3 PIW dataset. Comparison of body-part localization accuracy (%) (at 15 pixel error threshold). Each entry indicates (*avg*, *max*).

Method	Shoulders	Elbows	Wrists	*Avg*
FMP	(37.4, 43.8)	(26.8, 29.7)	(19.9, 20.0)	(28.0, 31.1)
MBP	(61.2, 62.7)	(49.8, 57.0)	(42.4, 54.3)	(51.1, 58.0)
S (Color)	(65.9, 73.0)	(47.7, 58.4)	(**42.7**, 47.8)	(52.1, 59.7)
S (Color) + P	(**71.5**, 77.3)	(**49.5**, 57.3)	(41.6, 47.8)	(**54.2**, 60.8)
F-CNN+BB	(**68.1**, 73.7)	(**55.4**, 64.0)	(**56.3**, 58.5)	(**59.9**, 65.4)

Table 4 3DMovie results. Comparison of average precision of keypoints (APK) ($\gamma = 0.2$ threshold) on the INRIA 3D Movie Dataset. The rows containing the word 'paper' indicate that those results have been directly extracted from the original paper.

Method	Shoulders	Elbows	Wrists	*Avg*
MBP+BB	0.449	0.288	0.142	0.293
S	0.500	0.310	0.134	0.315
S+P	0.706	**0.393**	**0.144**	0.415
S+P+BLS	**0.758**	0.383	**0.144**	0.428
S+P+SLR	**0.758**	0.389	0.139	**0.429**
F-CNN+BB	0.361	0.193	0.136	0.230
FMP (paper)	0.935	0.658	0.298	0.630
HOGcomb (paper)	**0.969**	**0.784**	**0.400**	**0.718**

4.4.2. Stereo Pictorial Structure: The Stereo Pictorial Structure (SPS) proposed in López-Quintero *et al.* [23] is only applied on SHPED, the dataset released by the authors of that paper, as it is designed for stereo pairs. Tab. 2 shows that our proposed method improves on it more than 7%, on average.

4.4.3. Flexible Mixtures of Parts: This method, referred as FMP, was proposed in the work of Yang and Ramanan[7] for monocular images. We can see that it works especially well on INRIA 3DMovie dataset, surpassing the MBP-based methods. However, it has clear limitations on SHPED and PIW.

4.4.4. Flowing ConvNets: The method of Pfister *et al.* [17], referred as F-CNN, is based on Convolutional Neural Networks, and targets HPE on monocular images. This method assumes a square input image where just one person is depicted. Therefore, for experiments on SHPED where multiple people can be in the same image, we have aided the method with our person detector BB. The results in row 'F-CNN+BB' of Tab. 2 indicate that this method is more accurate than the MBP-based ones for elbows and wrists, although our full system works on average better than it. A similar behaviour is observed on PIW (Tab. 3). As this method assumes temporal sequences, we can see in Tab. 4 that it obtains low results on single images.

4.4.5. Pose Estimation and Segmentation: The leading results on the INRIA 3DMovie dataset are the ones reported by the authors of the dataset [22]. Their proposed model is designed for stereo images where simultaneous estimation and segmentation of people is performed. As indicated in Sec. 4.1 of this paper, they do not release annotated stereo sequences, but isolated image pairs. This fact is reflected in the results in Tab. 4 obtained by the MBP-based methods, which rely on temporal sequences for an optimal performance, obtaining the 'HOGcomb' approach of Seguin *et al.* [22] the best results on this dataset.

5. Conclusions

This paper has presented a new approach for human pose estimation in stereo video sequences. The proposed pipeline starts by constraining the possible location of body joints by exploiting color and disparity information, and adding location priors to the most structured joints (*i.e.* shoulders). Finally, a body limbs recombination method is applied along the stereo sequence to obtain the best configuration of the body joints. The approach is tested on three state-of-the-art datasets: 'Poses In the Wild' that contains monocular video sequences, 'INRIA 3DMovie' that contains stereo image pairs; and, 'Stereo Human Pose Estimation Dataset' that contains stereo video sequences.

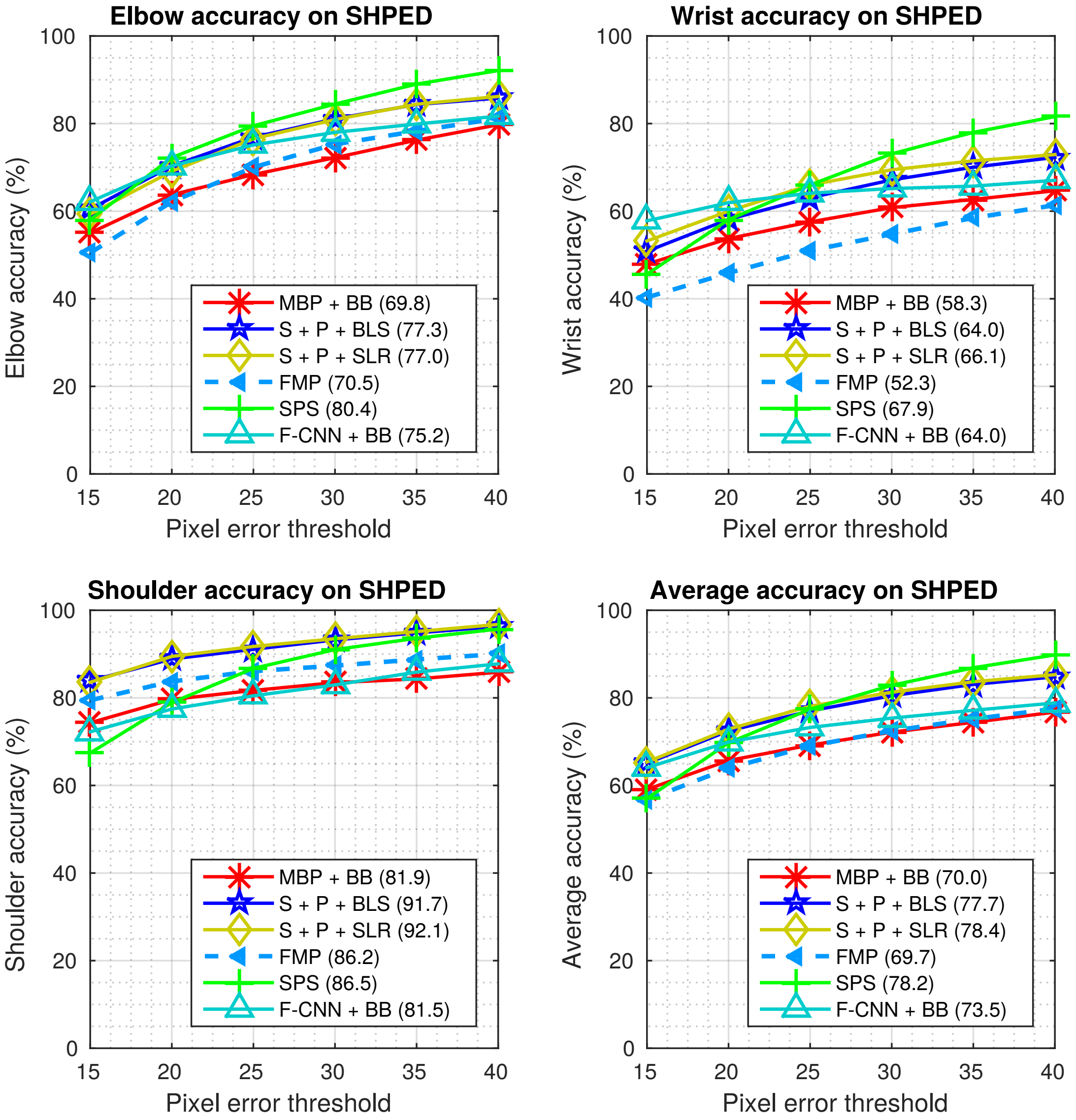


Fig. 7. *Comparison of accuracy on* ***SHPED (avg.)*** *dataset. For each curve, in parentheses, we show the area-under-the-curve. The higher, the better. (Best viewed in colour.)*

The latter has been used to fully evaluate the impact of our approach, whereas the remaining two datasets have been used for comparison with the state-of-the-art approaches on them. The results show that our method obtains better average results than the compared ones on SHPED, establishing new state-of-the-art results. Moreover, several stages of our method have shown to be helpful even on the monocular case (*e.g.* shoulders prior).

6. Acknowledgments

This project has been funded under projects TIN2016-75279-P (Spanish Ministry of Science and Technology) and RTC-2016-5661-1 Volumea (Spanish Ministry of Economy, Industry and Competitiveness).

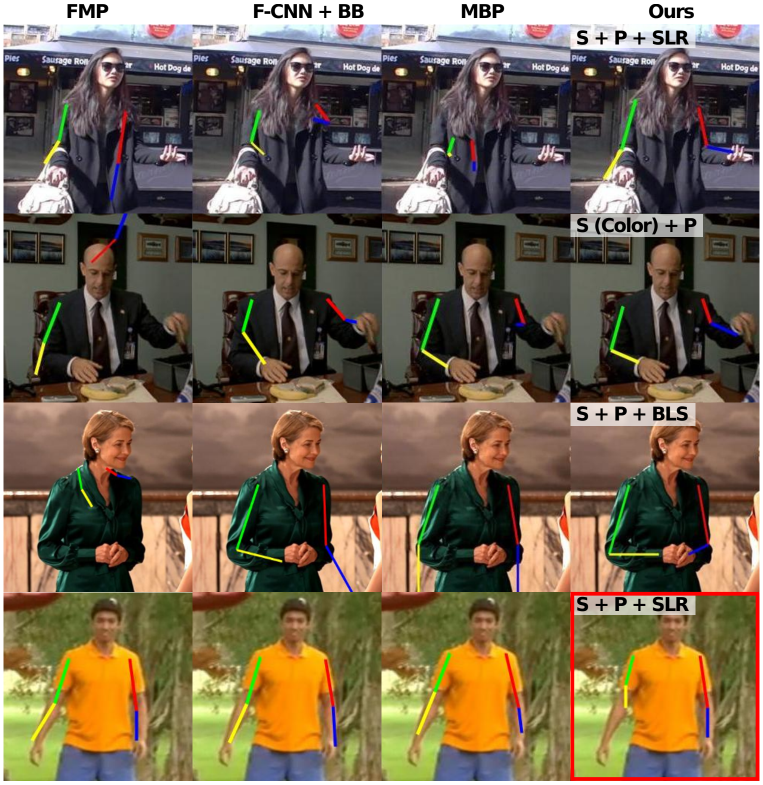


Fig. 8. *Comparative HPE results. Column 'Ours' contains our results, indicating the corresponding configuration. Note how our method is able to deal with different arm poses in both monocular and stereo images. Bottom row shows a failure case due to wrong segmentation of the arm. (Best viewed in colour.)*

References

[1] C. Wang, Y. Wang, and A. Yuille, "An approach to pose-based action recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 915–922, 2013.

[2] W. Zhou, C. Wang, B. Xiao, and Z. Zhang, "Human action recognition using weighted pooling," *IET Computer Vision*, vol. 8, pp. 579–587(8), December 2014.

[3] M. Eichner, M. J. Marín-Jiménez, A. Zisserman, and V. Ferrari, "2D articulated human pose estimation and retrieval in (almost) unconstrained still images," *Intl. Journal of Computer Vision*, vol. 99, no. 2, pp. 190–214, 2012.

[4] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from a single depth image," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1297–1304, 2011.

[5] M. Sun, P. Kohli, and J. Shotton, "Conditional regression forests for human pose estimation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3394–3401, 2012.

[6] X. Pérez-Sala, S. Escalera, C. Angulo, and J. González, "A survey on model based approaches for 2D and 3D visual human pose recovery," *Sensors*, vol. 14, pp. 4189–4210, 2014.

[7] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 2878–2890, Dec. 2013.

[8] M. Chen and X. Tan, "Part-based pose estimation with local and non-local contextual information," *IET Computer Vision*, vol. 8, pp. 475–486(11), December 2014.

[9] A. Cherian, J. Mairal, K. Alahari, and C. Schmid, "Mixing body-part sequences for human pose estimation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2361–2368, Jun. 2014.

[10] C. Zhang, Z. Li, Y. Cheng, R. Cai, H. Chao, and Y. Rui, "MeshStereo: A global stereo model with mesh alignment regularization for view interpolation," in *Proc. of the Intl. Conf. on Computer Vision (ICCV)*, pp. 2057–2065, Dec. 2015.

[11] J. Žbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *Journal of Machine Learning Research*, vol. 17, no. 65, pp. 1–32, 2016.

[12] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM Transactions on Graphics (TOG)*, vol. 23, pp. 309–314, ACM, 2004.

[13] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, June 2008.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *International Conference on Learning Representations*, May 2015.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, Jun. 2015.

[16] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr, "Conditional random fields as recurrent neural networks," in *Proc. of the Intl. Conf. on Computer Vision (ICCV)*, pp. 1529–1537, Dec. 2015.

[17] T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos," in *Proc. of the Intl. Conf. on Computer Vision (ICCV)*, pp. 1913–1921, Dec. 2015.

[18] M. Jiu, C. Wolf, G. Taylor, and A. Baskurt, "Human body part estimation from depth images via spatially-constrained deep learning," *Pattern Recognition Letters*, vol. 50, pp. 122 – 129, 2014.

[19] J. Shen, W. Yang, and Q. Liao, "Multiview human pose estimation with unconstrained motions," *Pattern Recognition Letters*, vol. 32, no. 15, pp. 2025 – 2035, 2011.

[20] M. Burenius, J. Sullivan, and S. Carlsson, "3D pictorial structures for multiple view articulated pose estimation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3618–3625, June 2013.

[21] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan, "Multi-view body part recognition with random forests," in *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 48.1–48.11, September 2013.

[22] G. Seguin, K. Alahari, J. Sivic, and I. Laptev, "Pose estimation and segmentation of multiple people in stereoscopic movies," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 8, pp. 1643–1655, 2015.

[23] M. López-Quintero, M. Marín-Jiménez, R. Muñoz-Salinas, F. Madrid-Cuevas, and R. Medina-Carnicer, "Stereo pictorial structure for 2D articulated human pose estimation," *Machine Vision and Applications*, vol. 27, no. 2, pp. 157–174, 2015.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 91–99, Dec. 2015.

[25] D. Park and D. Ramanan, "N-best maximal decoders for part models," in *Proc. of the Intl. Conf. on Computer Vision (ICCV)*, pp. 2627–2634, 2011.

[26] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 500–513, Mar. 2011.

[27] A. Ayvaci, M. Raptis, and S. Soatto, "Sparse occlusion detection with optical flow," *Intl. Journal of Computer Vision*, vol. 97, pp. 322–338, Oct. 2011.

[28] B. Sapp, D. Weiss, and B. Taskar, "Parsing human motion with stretchable models," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1281–1288, Jun. 2011.

[29] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 4040–4048, 2016.

[30] "OpenCV: Optical Flow computation on CUDA." `http://docs.opencv.org/3.0-beta/modules/cudaoptflow/doc/optflow.html`. Last visit: February 2017.