REGRESSION BASED POSE ESTIMATION WITH AUTOMATIC OCCLUSION DETECTION AND RECTIFICATION

Ibrahim Radwan[1]*, *Abhinav Dhall*[2]*, *Jyoti Joshi*[1], *and Roland Goecke*[1,2]

University of Canberra[1], Australian National University[2]

ibrahim.radwan@canberra.edu.au, abhinav.dhall@anu.edu.au, jyoti.dhall@canberra.edu.au, roland.goecke@ieee.org

ABSTRACT

Human pose estimation is a classic problem in computer vision. Statistical models based on part-based modelling and the pictorial structure framework have been widely used recently for articulated human pose estimation. However, the performance of these models has been limited due to the presence of self-occlusion. This paper presents a learning-based framework to automatically detect and recover self-occluded body parts. We learn two different models: one for detecting occluded parts in the upper body and another one for the lower body. To solve the key problem of knowing which parts are occluded, we construct Gaussian Process Regression (GPR) models to learn the parameters of the occluded body parts from their corresponding ground truth parameters. Using these models, the pictorial structure of the occluded parts in unseen images is automatically rectified. The proposed framework outperforms a state-of-the-art pictorial structure approach for human pose estimation on 3 different datasets.

Index Terms— Pictorial Structure, Articulated Pose Estimation, Occlusion Sensitive Rectification, Gaussian Process Recognition, Pose Search

1. INTRODUCTION

With the availability of cheap digital camera technology and public online databases for sharing digital images and videos, such as Flickr, Instagaram or Picasa, massive amounts of digital image data are now available. However, given the sheer amount of data, it is prohibitive to manually annotate these for retrieval purposes. It is therefore essential to have efficient automatic annotation and retrieval approaches at hand to enable users to find the data they are interested in. One such approach is based on annotating digital images with the human body pose of persons shown. To this end, articulated human pose estimation is a long studied problem in computer vision. This paper proposes a robust Pictorial Structure (PS) based framework, which results in better pose estimation in the case of self-occlusion in unconstrained images.

Articulated pose estimation based on the PS framework has attracted much attention in developing a large variety of applications such as automotive safety (pedestrian detection), surveillance, pose search and video indexing. PS models represent an object as a graph, where each node represents a body part and edges between nodes encode the kinematic constraints between connected pair of parts. Significant progress has been achieved [1, 2, 3, 4, 5, 6, 7], but highly articulated objects (e.g. human body) lead to many self-occluded parts, resulting in less accurate pose estimation and detection. There are two types of occlusion: 1) *Self-occlusion* caused by the object itself due large degrees of freedom, different camera views or different poses; 2) *Inter-occlusion* between different objects in the same image. In this paper, we focus on the former and propose a robust learning-based framework to rectify the human pose estimation in highly self-occluded scenes. The contributions of this paper are solutions to the following three key questions: 1) How can we detect whether there is occlusion in a given image? 2) If there is occlusion, how can we identify the body parts responsible for that occlusion? 3) How can we rectify the occluded part's position? To this end, we introduce (1) a general framework for self-occlusion detection, which reduces the search space of occluded parts, and (2) an approach for rectifying PS parameters of occluded parts in highly articulated poses that can work with any PS model, making it more robust to self-occlusion and allowing us to

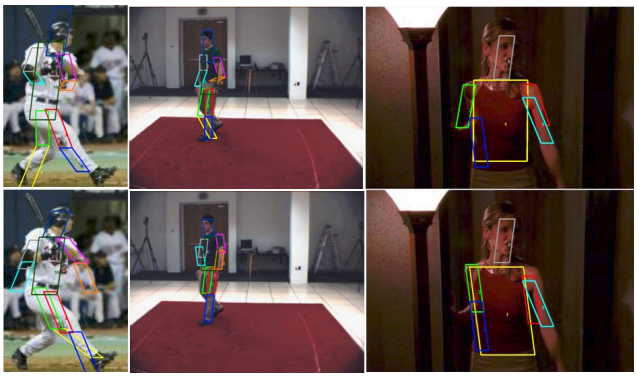

Fig. 1. Sample results from the People, HumanEva and Buffy datasets: (*Top*) Rectified pictorial structures from proposed approach. (*Bottom*) Andriluka *et al.* [1] pictorial structures.

*Equal first authors

978-0-7695-4711-4/12 $26.00 © 2012 IEEE
DOI 10.1109/ICME.2012.160

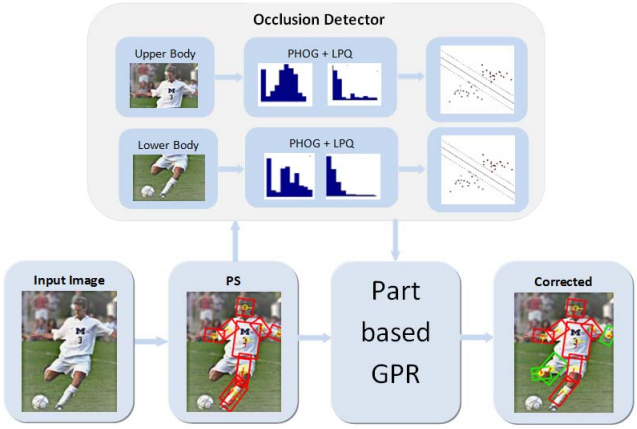

Fig. 2. Overall framework architecture for both occlusion detection and occluded parts rectification

accurately estimate the pose from monocular images (Fig. 1).

We use two binary discriminative non-linear SVM classifiers to detect the occluded parts in the upper and lower body regions. Iteratively, we select parts with a low posterior score in the region of the detected occlusion to rectify the PS parameters. Our rectification step is based on learning a set of mapping functions between the PS parameters and the ground truth from labelled training images. We employ Gaussian Process Regression (GPR) [8] for that purpose, which constructs a Bayesian model $p(\Sigma|\mathcal{D},\mathcal{D}') = p(\mathcal{D}|\Sigma,\mathcal{D}')p(\Sigma|\mathcal{D})'$ for learning the correlation between the correspondence parameters, where $p(\mathcal{D}|\Sigma,\mathcal{D}')$ is the likelihood of PS parameters $\mathcal{D}$ given the ground truth positions $\mathcal{D}'$ and the covariance function Σ, and $p(\Sigma|\mathcal{D}')$ is the prior of the covariance function for the PS parameters. To rectify an occluded part i, we use the model: $p(\mathcal{D}'_i|\Sigma_i,\mathcal{D}) = p(\Sigma_i|\mathcal{D},\mathcal{D}'_i)p(D'_i|D)$.

2. RELATED WORK

While there is a plethora of literature on articulated pose estimation and occlusion manipulation, we focus here on the recent and widely used PS models and methods for handling occlusion. The original PS by Felzenszwalb *et al.* [2] is based on a simple appearance model and requires background subtraction, which hence does not work well in cluttered and dynamic background scenes. Andriluka *et al.* [1] overcame this problem by using a discriminative appearance model. Starting from the original PS method [9] for discriminatively detecting each body part, they interpret the normalised margin of each part as the appearance likelihood for that part. Although this produces a general framework for both object detection and articulated pose estimation, their model is not able to estimate the human pose in highly occluded scenes. Our proposed algorithm is inspired by their work and extends it by providing a robust framework for rectifying PS in occluded scenes for human body pose estimation and detection.

[10, 11, 12] provide frameworks for handling occlusions between *multiple objects* in an image by estimating each object's pose based in the PS framework. In contrast, our approach focusses on *self-occlusion*. While all of the above methods are modelled to estimate poses from still images, there exists only limited research on the same task in videos. Kumar *et al.* [13] used structural learning of the PS parameters from videos. Their model is based on background subtraction from consecutive frames to define which parts are occluded. Ramanan *et al.* [7] employ the PS idea to find stylised poses, such as walking persons, by learning the difference between the background and foreground from consecutive frames and tracking the person in the followomg video frames. Their approach works well in videos for poses with little self-occlusion.

Sigal and Black [4] modelled self-occlusion handling in the PS framework as a set of constraints on the occluded parts, which are extracted after performing background subtraction which renders it unsuitable for dynamic background scenes. Our work follows both [1, 4] by producing a framework for articulated pose estimation robust to cluttered backgrounds and self occlusion without relying on background subtraction models. The step of rectifying occluded body parts via a GPR model is inspired by recent work by Asthana *et al.* [14], who used GPR for modelling parametric correspondences betweem face models of different people. Our problem is more difficult because the human body includes more parameters to be rectified and has more degrees of freedom than faces.

3. THE PROPOSED APPROACH

We start with a brief introduction of the *Pictorial Structure* framework for articulated pose estimation, then discuss its shortcomings in the presence of self-occlusion, before proposing a novel approach to rectify the PS.

3.1. Pictorial Structures

In the PS framework [2], the human body is represented as a graph with n vertices $V = v_1, \ldots, v_n$ for the body parts and a set of edges E where each $(v_i, v_j) \in E$ pair encodes the spatial relationship between parts i and j. For a given image I, PS learns two models. The first one learns the evidence of each part as an *appearance model*, where each part is parameterised by its location (x, y), orientation θ, scale s, and sometimes foreshortening. All of these parameters $\mathcal{D}$ are learned from exemplars and produce a likelihood of that image. The second model learns the kinematic constraints between each pair of parts in a *prior configuration model*. Given those two models for an image I, the posterior distribution over all the set of part locations is

$$p(\mathcal{L}|\mathcal{I},\mathcal{D}) \propto p(\mathcal{I}|\mathcal{L},\mathcal{D})p(\mathcal{L}|\mathcal{D}) \quad (1)$$

where $p(\mathcal{I}|\mathcal{L},\mathcal{D})$ measures the likelihood of representing the image in a particular configuration and $p(\mathcal{L}|\mathcal{D})$ is the kinematic prior configuration. Finding a maximum *a posteriori* probability (MAP) is equivalent to estimating the maximum likelihood for all parts. The best spatial relationship between pairs of parts, L^* for an image I is

$$\mathcal{L}^* = \arg\max_{L} P(\mathcal{L}|\mathcal{I},\mathcal{D}) \tag{2}$$

One of the major problems of this framework is the low contributions of the parts when they are occluded, resulting in either wrong or missing detections of these parts, which in turn leads to inaccurate pose estimation. To overome this, we propose a robust self-occlusion model, which works with any pictorial structure approach and can produce a robust pose estimate for articulated objects in scenes with cluttered background and self-occlusion.

3.2. Self-Occlusion Detector

Firstly, we introduce a novel self-occlusion detection approach that, unlike [4], does not rely on background subtraction for input images or the identification of occlusion relationships as a set of constraints between pairs of parts. Instead, we learn two binary models corresponding to the upper and lower body, respectively. Firstly, *Pyramids of Histogram of Gradients (PHOG)* are computed. The PHOG descriptor is an extension of Dalal *et al.*'s [15] HOG descriptor and has been extensively used for various computer vision problems such as object recognition [16] and facial expression analysis [17]. The upper and lower body regions of interest are divided into patches and a $3x3$ Sobel mask is applied to the edge contours for calculating the orientation gradients. Then, the gradients of each grid are joined together at each pyramid level. Secondly, we also compute the *Local Phase Quantisation (LPQ)* [18] descriptor, which belongs to the class of Local Binary Patterns (LBP) [19]. LPQ computes the short-term Fourier transform on a patch and has been empirically shown to better handle blur and illumination than LBP [18].

The output from the two descriptors is combined (separately for the upper and lower body). A non-linear *Support Vector Machine (SVM)* [20] is learnt and optimum parameters are found via fivefold cross-validation. As described in Fig. 2, standard PS is estimated for each image I, before the two ROIs (upper and lower body) are passed to the self-occlusion detection step. If there is a self-occlusion part i detected (e.g. left lower leg), the configuration parameters of that part are changed from $\mathcal{D}_i$ to $\mathcal{D}'_i$, where $\mathcal{D}'_i = (x', y', \theta')$, representing the rectified location and orientation of part i via *hallucination*. Based on this hallucination step the *eXtended Pictorial Structure (XPS)* model can be defined as

$$p(\hat{\mathcal{L}}|\mathcal{I},\mathcal{D}) \propto p(\mathcal{I}|\hat{\mathcal{L}},\mathcal{D}')p(\hat{\mathcal{L}}|\mathcal{D}') \tag{3}$$

In the next section, we discuss how to detect and rectify the occluded parts from the ROIs, i.e. how to map $\mathcal{D}$ to $\mathcal{D}'$

Algorithm 1: Part-by-part rectification via GPR

[Training]
input : $\{\mathcal{D}\}$ and $\{\mathcal{D}'\}$ matrices for training images
output: a Model $\mathcal{M}^i$ for each part p in the occluded ROI
begin
  $P \longleftarrow$ parts in ROI
  for $i \in P$ **do**
    $x \longleftarrow \mathcal{D}$
    $y \longleftarrow \mathcal{D}'^i$, `// ` i^{th} `column of` $\mathcal{D}'$
    Use GPR to estimate the prior covariance function
    Construct and save model $\mathcal{M}^i$ for part i
  end
end
[Prediction]
input : PS parameters d for occluded image I
begin
  $P \longleftarrow$ parts in ROI
  $S \longleftarrow$ PS score for all parts in ROI
  $i \longleftarrow 0$, `// i here means iteration number`
  $p(\hat{L}_i) \longleftarrow p(L)$ from eq. 1
  for $i \leq size\{P\}$ **do**
    $i \longleftarrow i + 1$
    Select $part$ with minimum Score
    Load GPR model for that part
    Predict d' for that part
    Estimate $\hat{L}_i$ for whole image I
    if $p(\hat{L}_i) \leq p(\hat{L_{i-1}})$ **then**
      $break$
    end if
  end
end

3.3. Rectifying Hallucinated Occluded Body Parts

Let $\mathcal{D} \in \mathbb{R}^{n \times m}$ represent the PS parameters matrix for m training Images with falsely detected parts due to self-occlusion. Let $\mathcal{D}' \in \mathbb{R}^{q \times m}$ be the corresponding ground truth parameters in those images. Our aim is to learn a mapping between these two set of parameters. To compute the mapping function $\mathcal{W} : \mathbf{D} \rightarrow \mathbf{D}'$, we propose a part-by-part hallucination method.

Part-by-Part Hallucination. Let,$\mathcal{D} = [d_1 \dots d_m]$ and $\mathcal{D}' = [d'_1 \dots d'_m]$, where each d_j is the j^{th} column, which represents the position and orientation parameters of all parts in a training image $\mathcal{I}$ and d'_i, j is $(i,j)^{th}$ element of matrix $\mathcal{D}'$. Formally, the training set is

$$\tau_i = \{(\mathbf{d}_j, d'_{i,j})\}_{j=1}^{m} \qquad i = 1, \dots, q \tag{4}$$

where $\mathbf{d} \in \mathcal{D}$ (the set of multivariate inputs) and $d' \in \mathcal{D}'$ (the set of outputs/targets). Here, a simple approach is to learn a non-linear mapping function $w_i : \mathcal{D}^{n \times m} \rightarrow \mathcal{D}'^{q \times m}$, where $i = 1, \dots, q$, that results in the mapping function $\mathcal{W} = [w_1\, w_2 \dots w_q]$, resulting in q models, which can be

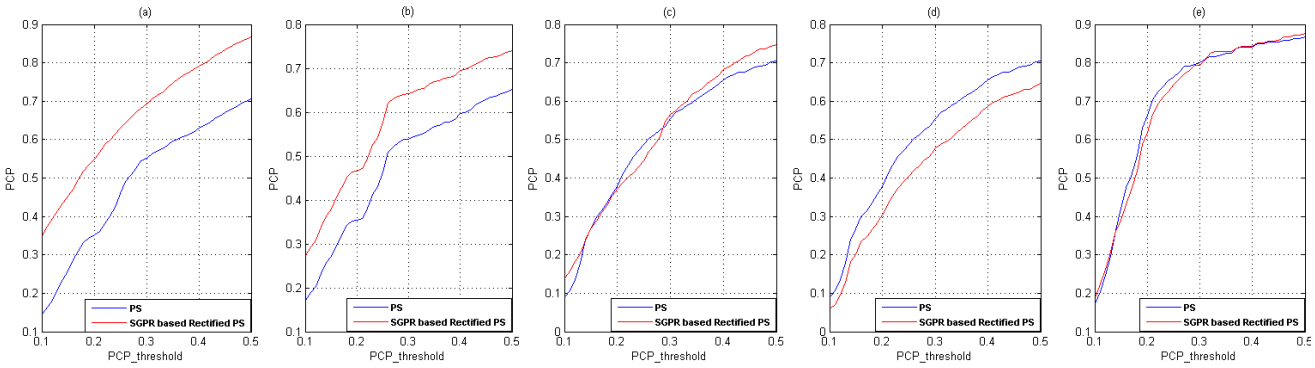

Fig. 3. PCP performance of our 2 frameworks against Andriluka *et al.* [1] (PS): (a) HumanEva dataset - full body, (b) HumanEva dataset - upper body, (c) People dataset - full body, (d) People dataset - upper body, (e) Buffy datatset - upper body

used to rectify q parts. We employ GPR [8] to compute the mapping function $\mathcal{W}$.

For unseen images, we rectify those occluded parts, which decrease the appearance and configuration likelihood in estimating the posterior value L of Eq. 1. We argue that the presence of occlusion in a part affects the likelihood of that part, which in turn affects the posterior probability of the whole body in the PS model. This argument motivates us to sort the likelihood of each part existing in the ROI and select the minimum likelihood value. Then, we use its pre-learnt model to rectify its correspondence and use it to estimate a new $p(\hat{L})$ for all parts. A comparison with the previous value of $p(\hat{L})$ is performed, and if it is improved, we select the second smallest value in the list and so on until the improvement of the pose estimation stops (see Algorithm 1).

4. EXPERIMENTS AND DISCUSSION

In this section, we evaluate the capability of the proposed approach to model the parametric correspondence between PS and the ground truth (from labelled data), and use them for localising the occluded parts in unseen sequences.

4.1. Datasets

Selecting suitable datasets was one of the challenges we faced in this work because of a lack of data with self-occlusion. Many commonly used datasets contain only a small amount of self-occluded body parts in their images. Therefore, we collected our training, validating and testing data from 3 different public databases: the People dataset [21], the BUFFY dataset [22] and the HumanEva dataset [23]. For evaluating the full body pose, the human body was divided into 10 parts: torso, head, left / right upper / lower arms, and left / right upper / lower legs. When evaluating the upper body pose only, we used only the first 6 parts.

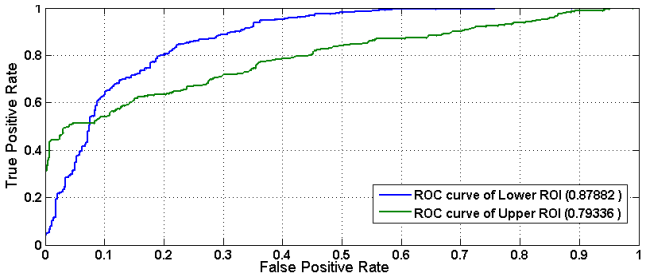

Fig. 4. ROC curves for our occlusion detector based for upper and lower body occluded parts

4.2. Performance Evaluation

Occlusion Detection Step: Rectifying the occluded body parts requires to localise them first (see Sec. 3). We built two discriminative SVM binary classifiers to reduce the search space into one of two regions of interest (ROI). The performance of these two binary classifiers was evaluated via the ROC metric, measuring the true positive average of samples against the false positive for both upper body ROI and lower body ROI. The accuracy of the occlusion detector was higher when the occluded body parts were located in the lower body region (Fig. 4). We believe that is due to the degree of freedom in the arms being higher than the degree of freedom in the legs; hence, the arm parts result in more self occlusion.

Rectification Step: To evaluate the performance of our regression based approach, we employ two criteria:

- **Detection rate** indicates number of detected stick figures men (PASCAL VOC criterion [24]).
- **Percentage of Correctly estimated body Parts (PCP)** counts an estimated body part as correct if its segment endpoints lie within $t\%$ of the length of the ground truth segment from their annotated location. PCP is evalu-

Table 1. Comparison of PCP, detection rate and total accuracy between the proposed approach and Andriluka *et al.* [1] (PS)

Database	Type	(PS) PCP	(Our) PCP	(PS) DetRate	(Our) DetRate	(PS) Accuracy	(Our) Accuracy
HumanEva	full body	70.54%	86.70%	85.20%	92.09%	60.10%	79.48%
	upper body	65.19%	74.10%	91.50%	94.00%	59.64%	69.65%
People	full body	70.56%	74.62%	84.71%	94.12%	59.77%	70.23%
	upper body	70.56%	64.55%	84.71%	90.59%	59.72%	58.47%
Buffy	full	–	–	–	–	–	–
	upper body	86.67%	87.72%	85.11%	88.64%	73.76%	77.75%

ated only for those that have been detected (i.e. there is a correct detection window). Overall performance is evaluated by a PCP curve, obtained by varying the accuracy threshold t [25].

Using those two evaluation criteria, we measure the performance of part-by-part rectification method against the classic PS method [1].

To rectify the occluded parts in the whole body, we established two experiments, one for the HumanEva and another one for the People database. In the first experiment, we constructed 10 independent models, one for each part based on single GPR (SGPR) [8], where the first 4 models are from images, which have lower occluded parts, and the 6 other parts are from images with occlusion in the upper body parts. The performance of these models has been evaluated on 200 frame from HumanEva database and 85 images from People database. These two are picturised in Fig. 3.a and 3.c, respectively.

To rectify the occluded parts in the upper body, we constructed three experiments: for HumanEva, for People and for Buffy databases. For frames from HumanEva and images from People we used the same models from the previous experiments which corresponding to the upper parts such that we used the 6 models for the upper body parts based on SGPR. The overall result of those experiments are shown in Fig. 3.b and 3.d. Since the Buffy database contains information about the upper body parts, the last experiment is only for upper body parts. We built the regression models based on frames which do not have occlusion because there are a few number of frames which contain occluded body parts. However, we tested on those which have occluded parts. The comparison between original PS and the rectified ones are shown in Fig. 3.e.

From the experimental results in Fig. 3, we can infer that the proposed approach using SGPR convincingly outperformed the state-of-the-art approach [1] for pose estimation of both full body part and upper body part localisation. In the upper body experiment for the People dataset, we got less accuracy because the regressor has been affected by the occlusion detector performance. A summary of the accuracy results for all experiments is shown in Table 1.

5. CONCLUSIONS AND FUTURE WORK

A GPR-based framework has been proposed to rectify self-occluded human body parts, which results in better articulated pose estimation for both upper and full body. This general framework can work on the output of any PS approach to detect occluded body parts and rectify their PS parameters. We showed that it is suitable for both videos and still images without prior tracking of the body parts, enabling accurate pose search in media databases such as YouTube, Flickr or Picasa. In the future, we plan to investigate other regression methods such as *Multiple output GPR*. Merging temporal information with PS parameters is another avenue to improved pose estimation, reducing the search space of the occluded parts.

6. ACKNOWLEDGEMENT

We would like to thank Akshay Asthana for the useful discussions.

7. REFERENCES

[1] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *CVPR*, 2009, pp. 1014–1021. 1, 2, 4, 5, 6

[2] P.F. Felzenszwalb and D.P. Huttenlocher, "Pictorial Structures for Object Recognition," *IJCV*, vol. 61, no. 1, pp. 55–79, Jan. 2005. 1, 2

[3] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Pose search: Retrieving people using their pose," in *CVPR*, 2009, DOI: 10.1109/CVPR.2009.5206495. 1

[4] L. Sigal and M.J. Black, "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation," in *CVPR*, 2006, pp. 2041–2048. 1, 2, 3

[5] H. Bhaskar, L. Mihaylova, and S. Maskell, "Human body parts tracking using pictorial structures and a genetic algorithm," in *4th Int. Conf. Intelligent Systems*, 2008, vol. 2, pp. 10–2 – 10–6. 1

[6] P.F. Felzenszwalb, R.B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *CVPR*, 2010, pp. 2241–2248. 1

[7] D. Ramanan, D. A. Forsyth, and A. Zisserman, "Strike a pose: tracking people by finding stylized poses," in *CVPR*, 2005, vol. 1, pp. 271–278. 1, 2

[8] C.E. Rasmussen and C.K.I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2005. 2, 4, 5

[9] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *CVPR*, 2008, DOI: 10.1109/CVPR.2008.4587597. 2

[10] M Eichner and V Ferrari, "We Are Family: Joint Pose Estimation of Multiple Persons," in *ECCV 2010, LNCS 6311*, pp. 228–242. 2010. 2

[11] X. Wang, T.X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *CVPR*, 2009, pp. 32–39. 2

[12] P. Fihl and T.B. Moeslund, "Pose Estimation of Interacting People using Pictorial Structures," in *AVSS*, 2010, pp. 462–468. 2

[13] M. Pawan Kumar, P.H.S. Torr, and A. Zisserman, "Learning Layered Pictorial Structures from Video," in *ICVGIP*, 2004, pp. 148–153. 2

[14] A. Asthana, M. Delahunty, A. Dhall, and R. Goecke, "Facial Performance Transfer via Deformable Models and Parametric Correspondence," *IEEE Trans. Vis. and Comp. Graph.*, vol. 99, 2011. 2

[15] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *CVPR*, 2005, pp. 886–893. 3

[16] A. Bosch, A. Zisserman, and X. Munoz, "Representing Shape with a Spatial Pyramid Kernel," in *CIVR*, 2007. 3

[17] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *FG, FERA Workshop*, 2011, pp. 878–883. 3

[18] V. Ojansivu and J. Heikkilä, "Blur Insensitive Texture Classification Using Local Phase Quantization," in *ICISP*, 2008, pp. 236–243. 3

[19] T. Ojala, M. Pietikäinen, and Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *PAMI*, pp. 971–987, 2002. 3

[20] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," 2001, http://www.csie.ntu.edu.tw/˜cjlin/libsvm. 3

[21] D. Ramanan, "Learning to Parse Images of Articulated Bodies," in *NIPS*, 2006. 4

[22] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive Search Space Reduction for Human Pose Estimation," in *CVPR*, 2008. 4

[23] L. Sigal, A. Balan, and M. Black, "HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion," *IJCV*, 2010. 4

[24] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results," http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html. 4

[25] M Eichner, M Marin-Jimenez, A Zisserman, and V Ferrari, "Articulated Human Pose Estimation and Search in (Almost) Unconstrained Still Images," Tech. Rep., ETH Zurich, 2010. 5

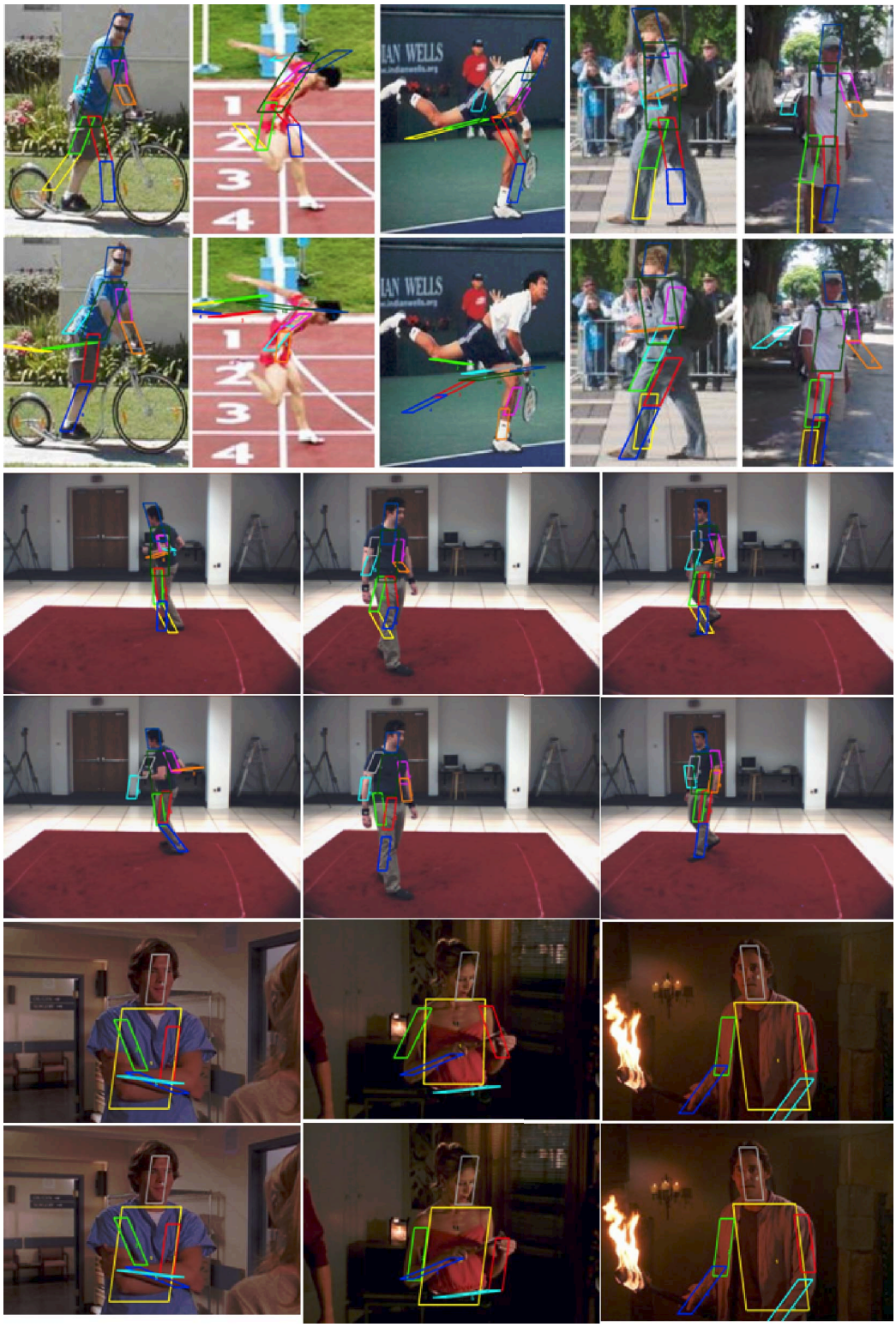

Fig. 5. Sample results for our proposed approach (*top*) and PS [1] (*bottom*) for the People (*top two rows*), HumanEva (*middle two rows*), Buffy (*bottom two rows*) databases