Articulated Human Pose Estimation in Unconstrained Images and Videos

Umar Iqbal

1 Introduction

Recent advances in Artificial Intelligence (AI) have resulted in the integration of an ever-increasing number of AI technologies in our daily lives. One of the quests in this direction is to develop autonomous systems that can integrate into our daily lives and interact with us as naturally as another human could do. Humans interact with each other and with their environment in complex ways. Even if we do not speak, our bodies transmit a lot of information such as our behavior, intention, or activity via hand and body gestures, facial expressions and body movements. While the humans can effortlessly detect and interpret such information from subtle and complex body signals, we need to endow these autonomous systems with similar capabilities to integrate them naturally in real-world human environments. Supporting such detailed reasoning would eventually need highly sophisticated computer vision systems – able to estimate human body pose, hand articulations, facial landmarks, and perceive their motion, all in the three-dimensional world space.

Besides, the developed methods can also be utilized in numerous practical applications e.g., robotics, activity recognition, surveillance systems, self-driving cars, driver-assistance, sign-language recognition, human monitoring, sports video analytic, gaming, entertainment industry, etc. Some example applications of the human body understanding can be seen in Figure 1. However, to be applicable in complex real-world scenarios, the developed methods must operate under varied background and lighting conditions, scenes with an unknown number of persons, sever body part occlusions and truncations, and large appearance variations exhibited by the human body. Further, they should be robust against noise due to sensing modalities such as motion blur, depth and scale ambiguities, low image resolution, and compression artifacts. The small functional parts of the human body, such as hands, present additional challenges due to the small size and heavy occlusions.

While the problem of articulated human pose estimation has been studied for a long time in the literature, most of the earlier works make strong assumptions and ignore the aforementioned challenges posed by the realistic scenarios. For example, a large number of earlier works for 2D body pose estimation assumes that only a single, pre-localized person is visible in the image [Felzenszwalb and Huttenlocher, 2005, Peng et al., 2018]. For 3D body pose estimation, the used training data is often recorded in indoor settings. Hence, the trained methods are only applicable in similar constrained scenarios. In contrast to body pose estimation, hand pose estimation has received much less attention in the literature and only a few works exist that address this challenging problem from only RGB images [Zimmermann and Brox, 2017, Mueller et al., 2018].

The goal of this thesis, therefore, is to address the challenges mentioned above and fill the gaps in the literature. To this end, the thesis focuses on developing computational methods for articulated pose estimation of humans from completely unconstrained images and videos. It starts by presenting a method for 2D body pose estimation of multiple, potentially truncated and occluded, people (Section 2). As a natural next step, it then addresses the problem of multi-person pose estimation and tracking in videos. The methods developed for pose estimation in images cannot be applied directly to this problem since in this case persons should also be associated over time. To this end, it presents an approach that jointly models pose estimation and tracking in a single formulation (Section 3). Furthermore, it introduces the PoseTrack benchmark which is a new large-scale benchmark for video-based multi-person human pose estimation and tracking. To date, the proposed dataset contains more than 1300 annotated videos and comprises more than 215K pose annotations. It also presents an extensive experimental study on the recent approaches for multi-person pose estimation and tracking and provides an analysis of the strengths and weaknesses of the state-of-the-art (Section 4). Given the 2D pose trajectories of a person, the thesis then presents a method to refine the pose trajectories by exploiting the information about human activities (Section 5).
Figure 1: Example applications of the computational methods for human pose estimation. Humanoid-robots can utilize body pose information to naturally interact with humans or for assisting people who need help with their day-to-day tasks or interactions e.g., elderly people or children with developmental disabilities (a). A self-driving car can anticipate the intentions of the pedestrians or cyclists by looking at their body and hand pose to make timely decisions (b). The information about the body pose can be used to recognize the activity of the persons, and to detect abnormal behaviors in surveillance systems (c,d). In sports video analytic it can be used to generate game statistics or to provide personalized training to the players (e). In the entertainment industry, body pose information can be used for gaming, or to replace expensive motion capture systems to apply special effects without the need for specialized suits (f).

Subsequently, given the estimated 2D poses, the problem of lifting the 2D poses to 3D is addressed. For this, an efficient and robust method for 3D pose retrieval and reconstruction is presented(Section 6). Finally, the thesis also addresses the challenging problem of hand pose estimation from RGB images and presents an approach which can work on images taken from the wild (Section 7).

1.1 Problem Formulation

We describe the pose by the locations of $J$ number of keypoints predefined using an anatomical structure, and develop methods for both 2D and 3D pose estimation. We define the 2D pose as $\mathbf{p} = \{\mathbf{x}_j\}_{j \in J}$ and 3D pose as $\mathbf{P} = \{\mathbf{X}_j\}_{j \in J}$, where $\mathbf{x}_j = (x_j, y_j) \in \mathbb{R}^2$ represents the 2D pixel coordinates of the body keypoint $j$ in image $I$ and $\mathbf{X}_j = (X_j, Y_j, Z_j) \in \mathbb{R}^3$ denotes the location of the keypoint in the 3D camera coordinate frame measured in millimeters.

Figure 2: (a) An example of the anatomical structure used to represent 2D body pose in this thesis. (b) The 3D body pose corresponding to the 2D pose shown in (a). (c) An example of multi-person 2D pose estimation, a unique color represents a unique individual. (d) An example of the anatomical structure used to represent 2D hand pose. (e) The 3D hand pose corresponding to the 2D pose shown in (d).

In this thesis, we do not make any assumption on the environment or the number of persons, and assume RGB input captured using a single camera, for both 2D and 3D pose estimation. We use a pose representation consisting of a sparse set of anatomical landmarks i.e., $J \in [13, 15]$ for body pose, and $J = 21$ for hand pose (see Figure 2). Such a sparse representation has been shown to be sufficient for the human visual system to recognize, for example, the activity [Johansson 1973], gender [Kozlowski and Cutting 1977], and sign-language [Poizner et al. 1981].

The rest of this summary describes the contributions of the thesis in further detail where each section corresponds to a chapter in the original thesis.
2 Multi-Person 2D Pose Estimation in Images

Figure 3: Overview of the proposed method. We detect persons in an image using a person detector (a). A set of joint candidates is generated for each detected person (b). The candidates build a fully connected graph (c) and the final pose estimates are obtained by integer linear programming (d). (best viewed in color)

Single person pose estimation has made a remarkable progress over the past few years. This is mainly due to the availability of deep learning based methods for detecting joints [Wei et al., 2016]. These approaches, however, assume that only a single person is visible in the image and the location of the person is known a-priori. Moreover, the number of parts are defined by the network, e.g., full body or upper body, and cannot be changed. For realistic scenarios such assumptions are too strong and the methods cannot be applied to images that contain a number of overlapping and truncated persons. In comparison to single person human pose estimation benchmarks, multi-person pose estimation introduces new challenges. The number of persons in an image is unknown and needs to be correctly estimated, the persons occlude each other and might be truncated, and the joints need to be associated to the correct person.

This section of the thesis presents an efficient method that estimates the body poses of an unknown number of persons in an image in which a person can be occluded by another person or might be truncated. For this, we cast the problem of multi-person pose estimation as joint-to-person association problem. Since solving joint-to-person association jointly for all persons in an image is an NP-hard problem and even approximations are expensive, the proposed approach solves the problem locally for each person. It starts by detecting all the persons visible in the image using a person detector and crops an image region for each detected person such that the cropped region contains sufficient context, but only the joints of persons that are very close. Subsequently, a fully connected graph is constructed from a set of detected joint candidates in an image region and the pose of the person at the center of the cropped region is estimated by solving the joint-to-person association using integer linear programming (ILP). The labeling of the joints and non-maxima suppression are directly performed by a convolutional neural network. An overview of the proposed approach can be seen in Figure 3. The proposed approach is evaluated on the challenging MPII Human Pose Dataset for multiple persons where it improves the accuracy of the competing approach [Pishchulin et al., 2016] while reducing the runtime by a factor between 6,000 and 19,000. Some qualitative results can be seen in Figure 4.

Figure 4: Some qualitative results for the MPII Multi-Person Pose Dataset.
3 Joint Multi-Person Pose Estimation and Tracking

Figure 5: Top row shows the results of the single-frame method presented in Section 2. Same color represents same person across frames. Note the different identities (color) assigned to same persons across frames. Bottom row shows the results of the approach presented in this section. We can see that the proposed method tracks the persons overtime while also estimating their body poses. (best viewed in color)

Section 2 presented an approach that can estimate the poses of multiple people in unconstrained images. The proposed method does not make any assumption regarding the appearance of the people, nor does it assume that the number of people is known. However, it cannot be applied to videos directly since it also requires to solve the problem of person association over time in addition to the pose estimation for each person (see. Figure 5).

In order to address this problem, this section introduces the challenging problem of pose estimation and tracking of an unknown number of persons in unconstrained videos. This means that we have to deal with large pose and scale variations, fast motions, and a varying number of persons and visible body parts due to occlusion or truncation in videos. In contrast to previous works, we aim to solve the association of each person across the video and the pose estimation together. For this, we present a novel method that jointly models multi-person pose estimation and tracking in a single formulation. The proposed approach builds on the work for multi-person pose estimation in images described in Section 2, and represents body joint detections in a video by a spatio-temporal graph. The spatio-temporal joint-to-person association is then performed by solving an integer linear program whose feasible solution partitions the graph into valid body pose trajectories of any unknown number of persons. The proposed approach implicitly handles occlusion and truncation of the persons, and can work on videos taken from the wild. An overview of the proposed method can be seen in Figure 6.

Since the problem of joint multi-person pose estimation and tracking had not been addressed quantitatively in the literature, we also present a new challenging “Multi-Person PoseTrack” dataset with sixty fully-annotated videos. The dataset provides pose annotations for multiple persons in each video to measure pose estimation accuracy, and also provides a unique ID for each of the annotated persons to benchmark multi-person pose tracking. The proposed dataset introduces new challenges to the field of human pose estimation and tracking since it contains a large amount of appearance and pose variations, body part occlusion and truncation, large scale variations, fast camera and person movements, motion blur, and a sufficiently large number of persons per video. In order to jointly evaluate the pose estimation and tracking accuracy, we also introduce a comprehensive evaluation protocol. The evaluation protocol is devised such that the developed algorithms cannot make any assumption about the number, size, or location of the persons. Finally, we evaluated our proposed approach and several baseline methods on the new dataset. The experimental results demonstrate that the proposed method outperforms other baseline methods. Some qualitative results of our approach can be seen in Figure 7. The source code, pre-trained models and the dataset are publicly available.

http://pages.iai.uni-bonn.de/iqbal_umar/PoseTrack/
Figure 6: The proposed method jointly solves the problem of multi-person pose estimation and tracking for all persons appearing in a video together. It first generates a set of joint detection candidates in each video frame (top). From the detections, it builds a graph consisting of spatial edges (blue) connecting the detections within a frame and temporal edges (red) connecting detections of the same joint type over frames (middle). Finally, it solves the problem using integer linear programming whose feasible solution provides the pose estimate for each person in all video frames, and also performs person association across frames (bottom).

Figure 7: Qualitative Results. Visualization of the pose estimation and tracking results of the proposed approach on the newly proposed Multi-Person Pose-Track dataset. Every fifth frame for each video clip is shown. Note that the proposed approach can estimate poses under severe occlusions and truncations, clutter, complex background and large scale variation.
4 A Benchmark for Human Pose Estimation and Tracking

The significant progress in single-frame human pose estimation has been facilitated by the use of deep learning-based architectures [Simonyan and Zisserman, 2014, He et al., 2016] and by the availability of large-scale benchmark datasets such as “MPII Human Pose” [Andriluka et al., 2014] and “MS COCO” [Lin et al., 2014]. In Section 3, we presented a dataset for multi-person pose estimation in videos. However, as compared to the datasets available for multi-person pose estimation in images, the dataset is at a very small scale and provides only 60 videos with most sequences containing only 41 frames. While the dataset makes a first step toward solving the problem at hand, it is certainly not enough to cover a large range of real-world scenarios and to learn stronger pose estimation models. In this section, we aim to fill this gap by extending the dataset proposed in Section 3 to a large-scale, high-quality benchmark for video-based multi-person pose estimation and tracking. For this, we collected new videos and annotated them in an accurate and unified manner. The videos were chosen such that they represent challenging real-world scenarios i.e., they contain a large amount of body pose, appearance, and scale variation, as well as body part occlusion and truncation. Further, it was ensured that the videos contain severe body motion, i.e., people occlude each other, re-appear after complete occlusion, vary in scale across the video, and also significantly change their body pose. Compared to 16K poses in the previous version, the extended dataset contains more than 150K pose annotations.

Furthermore, to attract the attention of the research community towards this challenging problem, we organized PoseTrack challenge and workshop at ICCV’17 (later also at ECCV’18) and invited the research community to submit novel solutions for multi-person pose estimation and tracking. We also developed two baseline methods. For this, we modified our approach from Section 3 to incorporate stronger pose estimation model [Cao et al., 2017] and developed an additional baseline method based on [Insafutdinov et al., 2017]. We evaluated the approaches submitted to PoseTrack challenge and our baselines on the new dataset and conducted extensive performance analysis. We highlighted the strengths and weaknesses of the state-of-the-art approaches and highlighted the most promising future research directions. To prevent over-fitting and to ensure that all methods will be evaluated using the same ground-truth and evaluation scripts, we kept the test data with-held and developed an online evaluation server. We are continuously in the process of extending the dataset with more challenging videos, and to-date the dataset contains more than 1300 videos with more than 250K pose annotations. A comparison of the proposed dataset with other publicly available datasets can be seen in Table 1. To-date, more than 400 researchers from ~ 380 institutes have registered at our online server and more than 75 methods have been evaluated on our evaluation server.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of person</th>
<th>Multi-person</th>
<th>Video-labeled poses</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSP</td>
<td>2,000</td>
<td></td>
<td></td>
<td>sports (8 act.)</td>
</tr>
<tr>
<td>LSP Extended</td>
<td>10,000</td>
<td></td>
<td></td>
<td>sports (11 act.)</td>
</tr>
<tr>
<td>MPII Single Person</td>
<td>26,429</td>
<td></td>
<td></td>
<td>diverse (491 act.)</td>
</tr>
<tr>
<td>FLIC</td>
<td>5,003</td>
<td></td>
<td></td>
<td>feature movies</td>
</tr>
<tr>
<td>FashionPose</td>
<td>7,305</td>
<td></td>
<td></td>
<td>fashion blogs</td>
</tr>
<tr>
<td>We are family</td>
<td>3,131</td>
<td>✓</td>
<td></td>
<td>group photos</td>
</tr>
<tr>
<td>MPII Multi-Person</td>
<td>14,993</td>
<td>✓</td>
<td></td>
<td>diverse (491 act.)</td>
</tr>
<tr>
<td>MS COCO Keypoints</td>
<td>105,698</td>
<td></td>
<td></td>
<td>diverse</td>
</tr>
<tr>
<td>Penn Action</td>
<td>159,633</td>
<td>✓</td>
<td></td>
<td>sports (15 act.)</td>
</tr>
<tr>
<td>JHMDB</td>
<td>31,838</td>
<td>✓</td>
<td></td>
<td>diverse (21 act.)</td>
</tr>
<tr>
<td>YouTube Pose</td>
<td>5,000</td>
<td>✓</td>
<td></td>
<td>diverse</td>
</tr>
<tr>
<td>Video Pose 2.0</td>
<td>1,286</td>
<td>✓</td>
<td></td>
<td>TV series</td>
</tr>
<tr>
<td>FYDP</td>
<td>1680</td>
<td>✓</td>
<td></td>
<td>dance</td>
</tr>
<tr>
<td>UYDP</td>
<td>2000</td>
<td>✓</td>
<td></td>
<td>dance</td>
</tr>
<tr>
<td>PoseTrack 2016</td>
<td>16,219</td>
<td>✓</td>
<td>✓</td>
<td>diverse</td>
</tr>
<tr>
<td>PoseTrack 2017</td>
<td>153,615</td>
<td>✓</td>
<td>✓</td>
<td>diverse</td>
</tr>
<tr>
<td>PoseTrack 2018</td>
<td>261,255</td>
<td>✓</td>
<td>✓</td>
<td>diverse</td>
</tr>
</tbody>
</table>

Table 1: Overview of publicly available datasets for articulated human pose estimation in single frames and video. For each dataset the number of annotated poses, availability of video pose labels and multiple annotated persons per frame, as well as types of data are reported.
5 Action Priors for Human Pose Estimation

Figure 8: Overview of the proposed framework. We propose an action conditioned pictorial structure model for human pose estimation (2). Both the unaries $\phi$ and the binaries $\psi$ of the model are conditioned on the distribution of action classes $p_A$. While the pairwise terms are modeled by Gaussians conditioned on $p_A$, the unaries are learned by a regression forest conditioned on $p_A$ (1). Given an input video, we do not have any prior knowledge about the action and use a uniform prior $p_A$. We then predict the pose for each frame independently (3). Based on the estimated poses, the probabilities of the action classes $p_A$ are estimated for the entire video (4). Pose estimation is repeated with the updated action prior $p_A$ to obtain better pose estimates (5).

In Section 3 & 4 we discussed several methods for multi-person pose estimation and tracking. The methods provide pose trajectories for each person visible in the video. Such body pose trajectories have been shown to be useful for activity recognition in the literature [Jhuang et al., 2013]. Intuitively, the information about the activity of a person can also provide a strong cue about the pose. Hence, in this section, we propose to utilize information about human actions to improve pose estimation.

We present a pictorial structure model that exploits high-level information about activities to incorporate higher-order part dependencies by modeling action specific appearance models and pose priors. However, instead of using an additional expensive action recognition framework, the action priors are efficiently estimated by our pose estimation framework, i.e., by using the estimated pose trajectories. The framework of the approach is illustrated in Figure 8. The proposed approach is evaluated on the challenging J-HMDB [Jhuang et al., 2013] and Penn-Action [Zhang et al., 2013] datasets, which consist of videos collected from the Internet and contain large amount of scale and appearance variations. The effectiveness of the proposed approach is demonstrated via extensive experiments for both pose estimation and action recognition. We also show that learning the right amount of appearance sharing among action classes improves the pose estimation. Compared to the competing approach [Nie et al., 2015], the proposed approach improves the pose estimation accuracy by over 30%. The models and source code are publicly available. Some qualitative results can be seen in Figure 9.

Figure 9: Qualitative results as compared to the baseline model. The left part of the images corresponds to the baseline while the right part shows improved poses obtained by the proposed model.

http://pages.iai.uni-bonn.de/iqbal_umar/action4pose/
Many applications such as assistive robots, gaming, or human-computer interaction also need to know body pose information in 3D. All of the methods introduced so far in this thesis, however, focus only on 2D pose estimation. A natural question then is, can we estimate 3D pose from an unconstrained RGB image? To address this question, in this section, we propose an approach for 3D human body pose estimation from RGB images.

Existing methods for 3D pose estimation use deep neural networks to directly regress 3D pose from images. The main bottleneck between these approaches and their applicability in unconstrained environment is the availability of training data, since collecting large number of unconstrained training images with 3D pose annotations is practically infeasible. To address this shortcoming, in this section, we present a dual-source approach that does not require training data with 3D pose annotations, but instead relies on two independent sources of training data both of which are available abundantly. The first source consists of accurate 3D motion capture data, and the second source consists of unconstrained images with annotated 2D poses. We combine both sources in a unified framework that first performs 2D pose estimation, as done in the previous sections, and then lifts it to 3D using a robust approach for 3D pose reconstruction. An overview of the proposed approach can be seen in Figure 10. In contrast to recent approaches that learn deep neural networks to regress 3D pose directly from images, our proposed method does not require training images with 3D pose annotations. This has the advantage that the proposed method is not restricted to the views only present in the training data, which often consist of the views only from the controlled indoor settings. Hence, our method can also perform pose estimation in unconstrained scenes. We extensively evaluate our approach on two popular datasets for 3D pose estimation namely Human3.6M [Ionescu et al., 2014] and HumanEva [Sigal et al., 2010] and provide an in-depth analysis of the proposed approach. Some qualitative results of the proposed approach can be seen in Figure 11.
7 2D and 3D Hand Pose Estimation

Figure 12: Overview of the proposed approach. Given an image of a hand, the proposed CNN architecture produces latent 2.5D heatmaps containing the latent 2D heatmaps $H^{2D}$ and latent depth maps $H^{\hat{z}}$. The latent 2D heatmaps are converted to probability maps $H^{2D}$ using softmax normalization. The depth maps $H^{\hat{z}}$ are obtained by multiplying the latent depth maps $H^{\hat{z}}$ with the 2D heatmaps. The 2D pose $p$ is obtained by applying spatial softargmax on the 2D heatmaps, whereas the normalized depth values $\hat{Z}$ are obtained by the summation of depth maps. The final 3D pose is then estimated by the proposed approach for reconstructing 3D pose from 2.5D.

All of the approaches presented so far in this thesis focus only on human body and completely ignore hands. Whereas, estimating the pose of a hand is an essential part of the human-computer interaction. Therefore, in this last chapter, we shift our attention to the challenging problem of hand pose estimation.

In the previous chapter, we presented an approach for 3D pose reconstruction from 2D poses. The approach has the advantage that it does not require training images with 3D pose annotation. However, relying only on 2D pose information has one major drawback that it can fall prey to re-projection ambiguities e.g., the classic turning ballerina optical illusion. Therefore, it is essential to exploit the image information effectively to resolve challenging ambiguous cases. To this end, we present a carefully designed novel 2.5D pose representation which brings-forth several advantages. First, it is invariant to scale and translation, therefore, can be estimated easily from an RGB image using a CNN. Second, it allows training the network in a multi-task setup hence, multiple sources of training data can be used, hence, the trained models can also generalize to unconstrained images. Third, and most importantly, it enables the exact recovery of the absolute 3D pose up to a scaling factor, where the scale can be estimated additionally given the prior of the hand size. We also propose a novel CNN architecture to efficiently regress the 2.5D pose from images without the loss of spatial resolution. This is done by learning 2.5D heatmaps in a latent way by using a differentiable loss function. An overview of the proposed approach can be seen in Figure 12. Our proposed approach and the novel network architecture achieve the state-of-the-art accuracy for 2D and 3D hand pose estimation on several challenging datasets. Some qualitative results of can be seen in Figure 13.

Figure 13: Qualitative Results. The proposed approach can handle severe occlusions, complex hand articulations, and unconstrained images taken from the wild.
References


Marcin Eichner and Vittorio Ferrari. We are family: Joint pose estimation of multiple persons. In ECCV, 2010.


Sam Johnson and Mark Everingham. Learning Effective Human Pose Estimation from Inaccurate Annotation. In CVPR, 2011.


