SELF-OCCLUSION AND 3D POSE ESTIMATION IN STILL IMAGES

Julio C. S. Jacques Junior, Leandro L. Dihl*, Cláudio R. Jung+, Soraia R. Musse**

* Pontifícia Universidade Católica do Rio Grande do Sul, Brazil
+ Institute of Informatics, Federal University of Rio Grande do Sul, Brazil

ABSTRACT

In this paper we propose a self-occlusion and 3D pose estimation model for human figures in still images based on a user-provided 2D skeleton. An initial segmentation model is used to capture labeled human body parts in a 2D image. Then, occluded body parts are detected when different body parts overlap, and are disambiguated by analyzing the energy of the corresponding contours around the intersection points. The estimated occlusion results feed the 3D pose estimation algorithm, which reconstructs a set of plausible 3D postures. Experimental results indicate that the proposed technique works well in non trivial images, effectively estimating the occluded body parts and reducing the number of possible 3D postures.

Index Terms— human body parts segmentation, self-occlusion estimation, 3D pose estimation.

1. INTRODUCTION

Estimating 3D articulated human pose from a single view is of great interest to numerous vision applications, including human-computer interaction, visual surveillance, activity recognition from images, etc. As related in [1], this problem remains very challenging for several reasons. First, recovering 3D human poses directly from 2D images is inherently ambiguous due to loss of depth information. In addition, the shape and appearance of articulated human body vary significantly due to factors such as clothing, lighting conditions, viewpoints, and poses.

As related in [1], human pose estimation algorithms can be categorized as discriminative (model-free, e.g. example-based and learning-based) and generative (model-based, e.g. tree structure based on prior knowledge). Exemplar-based approaches store a set of training samples along with their corresponding pose descriptors, and for a given test image, a similarity search is performed to find similar candidates. Learning-based approaches learn the direct mapping from image observations to pose space using training samples. Most discriminative pose estimation use silhouette images to perform pose estimation [2, 3, 4, 5]. However, it is important to notice that silhouettes are inherently ambiguous, as different 3D poses can have very similar silhouettes. Some generative pose estimation uses 2D body part positions estimated by detectors (e.g. [6]) for 3D pose recovery [7, 8, 9]. One drawback of this class of approaches is the inherent ambiguity of 2D images and the occurrence of self-occlusion in certain postures.

Self-occlusion in human pose is a classic problem in computer vision, and there are some approaches that tackle this problems. Sigal and Black [10] presented an occlusion-sensitive model to articulated pose estimation. The model uses local image likelihoods that approximate the global likelihood by accounting for occlusions and competing explanations of image evidence by multiple parts. The approach proposed by Huang and Yang [1] uses a regression model to learn the mapping from image feature space to pose space, but differs from [10] in that sparse representations are learned from examples with demonstrated ability to handle occlusions. Radwan et al. [11] used a Gaussian Process Regression models to learn the parameters of occluded body parts. Kim and Kim [12] detected whether a given body part is occluded or not by analyzing the eigenvalues of 3D time-of-flight image data gathered from the joint point of each body part. In [13], the authors proposed a self-occlusion state estimation method. In their approach, a Markov Random Field (MRF) is used to model the occlusion state that represents the pair wise depth order between two human body parts. A novel estimation method is proposed to infer a body pose and an occlusion state separately.

In this paper we assume that the 2D skeleton of the human figure in the picture is given, and our goal is to find individual body parts (with occlusions disambiguated) and to reconstruct the 3D pose. The algorithm proposed by Taylor [14] is widely used for this purpose, but due to the number of joints of the skeleton (20 joints, in our case), it can produce 2^{20} possible solutions. The main idea of the proposed approach is to reduce the number of pose candidates by using biomechanics constraints and the self-occlusion estimative to remove implausible solutions.

2. OUR MODEL

In this work we propose an approach to detect and disambiguate self-occlusion for human figures in 2D still images, with applications in 3D pose estimation. Figure 1 illustrates

Authors would like to thank Brazilian agency CNPq and FAPESP to the INCT-SEC, processes 573963/2008-8 and 08/57870-9.

the main steps of our model. The first step is to associate a 2D skeleton model (Figure 1(a)) to the person in the picture (Figure 1(b), in cyan). This stage can be done manually or automatically, depending on the application. The second stage is to segment the person in the picture. For this purpose, we use the algorithm proposed by Jacques Junior et al. [15], which segments a person in a picture with semantic information based on an energy contour value for each body part, based on gradient information, coherence to the bone of the corresponding body part and anthropometric distances. Such model produces a closed contour, where each point is associated to a specific body part, as illustrated in Figure 1(b) using different colors. The third step of the model is to identify the intersections of body parts, in a higher level (e.g. the arm and the torso are intercepting – Figure 1(c)), characterizing the occlusions. The fourth stage is to analyze in a lower level each intersection candidate (the red dots shown in Figure 1(c)) to identify which body part is under occlusion. Disambiguation of occluded body parts is done by evaluating the behavior of the contour energy function in the neighborhood of the intersection points. The output of the self-occlusion model is a list of intersection pars and the self-occlusion information (the body part under occlusion, e.g. right arm, in Figure 1(d), is occluded by the left hand). This kind of information feed the 3D pose estimation algorithm, which is used to reconstruct the 3D pose (Figure 1(e)). The proposed model is described in the next sections.

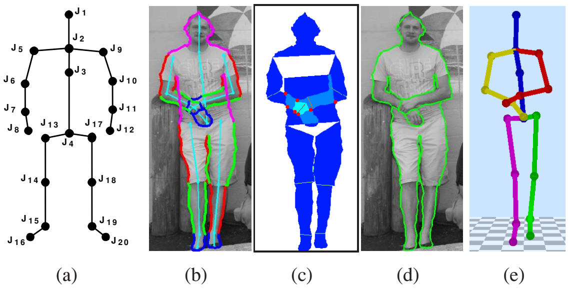


Fig. 1. Overview of the proposed model. (a) Adopted skeleton model. (b) Segmentation result. (c) Intersections between body parts and the intersection points (in red). (d) Illustration of the self-occlusion estimation result. (e) 3D estimated pose.

2.1. Finding and disambiguating occlusions

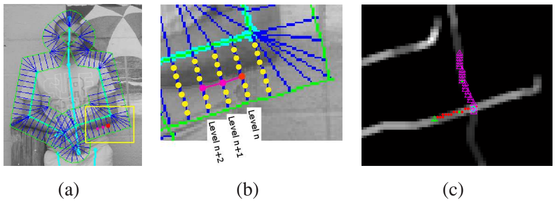
Given a 2D skeleton, the human segmentation model proposed by Jacques Junior et al. [15] explores edge information, orientation coherence and anthropometric estimated parameters to generate a weighted acyclic directed graph (WADG) around the 2D skeleton of the person. The graph is formed by levels orthogonal to the "bone" of the corresponding body part, and the contour with semantic information is a path with maximal cost containing exactly one vertex at each level. Figure 1(b) illustrates the segmented contour (with labeled body parts) based on the 2D skeleton provided in Figure 1(a).

To identify body part occlusions, we initially identify the 2D regions (blobs) associated to each body part based on the labeled contour information. Theoretically, an occlusion happens when the 2D projections of the corresponding regions overlap. However, due to inaccuracies when obtaining the contour of the person, several false intersections are detected, usually related to the intersection of just a few pixels. To reduce the number of false intersections, we use an area threshold. Thus, only body parts with intersection area higher than T_h are considered (illustrated in Figure 1(c) – both forearms and hands). The threshold T_h is defined as a fraction (set experimentally to 0.25) of the area of the smallest body part under analysis, considering pair wise blob comparisons.

Given an intersection region satisfying the area constraint, there are two possibilities: i) **Contours do not intersect**: in some cases, some intersection does not return a intersection point between the contours of the two body parts (e.g. the hand in front of the torso). In such case we use the hierarchy of the human body: for example, the hand has the same answer (occluded or not) of their previous adjacent body part (in this case, the forearm) and so on, until the pair of body parts being analyzed presents some contour intersection. ii) **Contours intersect**: there is at least one intersection point between the contours of the intersection regions. When it happens, we first compute the intersection points of body contours (e. g. the red dots shown in Figure 1(c)). For each intersection point, the energy of the contour is evaluated in a small neighborhood inside the occlusion region, for both body parts under analysis. More precisely, we consider the portion of the contour comprised in the $N = 2$ closest levels of the graph inside the occlusion area, as illustrated in Figures 2(a) and (b) for the left forearm, where the blue line segments denote the levels of the graph. Let consider an intersection point i and a body part m_i adjacent to i. Let $\mathcal{M}_i$ denote the neighborhood of the contour related to m_i that lies within the closest N graph levels in the occluded portion of m_i. The local contour strength of the m_i is given by

$$S(m_i) = \underset{\boldsymbol{p} \in \mathcal{M}_i}{\text{median}}\{E(\boldsymbol{p})\}, \qquad (1)$$

where $\boldsymbol{p}$ is a pixel on the contour, and $E(\cdot)$ is the energy map used to compute the silhouette as defined in [15]. The occluding region is expected to present a larger energy strength S, since the image edges of the occluding region tend to be visible, and hence E tends to be larger. Thus, if m_i and n_i are two intersecting body parts in the neighborhood of intersection point i, we decide (locally) that m_i is the occluding region if $S(m_i) > S(n_i)$, and n_i is the occluding region otherwise. For instance, Figure 2(c) illustrates the energy maps E related to the forearm-torso intersection shown in Figure 2(b). The local contour strength related to the forearm is $S(m_i) = 8.8897$, whereas the local contour strength related to the torso is $S(n_i) = 4.2874$. Hence, the forearm is (locally) occluding the torso in the neighborhood of the intersection point, which is the correct decision.



(a) (b) (c)

Fig. 2. (a) Example of occlusion and the WADG used to obtain the contour. (b) zoom of (a), highlighting the intersection point between the left forearm and the torso (shown in Figure 1(c)). (c) Analyzed contour points for the left forearm (in red) and the torso (in magenta).

The procedure described so far works for disambiguating occlusions locally around one contour intersection point. In general, overlapping body parts may have more than one contour intersection point, and a final decision is made based on the local decisions at each intersection point. Let i_1, i_2,...,i_k denote k intersection points related to a pair of occluding body parts m_i and n_i, and let $D(i_k)$ be the local decision at point i_k (either m_i or n_i). The final decision is given by:

- ***Most voted wins***: if a given body part gets more than $k/2$ votes, it is selected as the occluding region.
- ***Largest strength wins:*** if there is a tie in the voting, we decided based on the largest local strength, i.e., m_i is selected if $\max_{j=1,\ldots k}\{S(m_{i_j})\} > \max_{j=1,\ldots k}\{S(n_{i_j})\}$.

Finally, conflicts are analyzed: if an estimated occlusion generates a conflict, for example, the left hand is in front of the right arm and at same time it is behind the right forearm, we choose the answer based on the maximum local strength of each body part under conflict (in this case, the hand will be in front of the both right arm and forearm or behind them).

2.2. 3D Pose Recovery

Our solution to obtain the 3D postures is based on Taylor's work [14], which presents a method for recovering information about articulated objects from a single image. The similarities with our work are that both methods assume a scaled orthographic projection and both use geometrical information as constraints. According to Taylor, if we have a line segment of known length l in the image under scaled orthographic projection (in our case the line segment is a bone in the skeleton), the corresponding 3D end points (x_1, y_1, z_1) and (x_2, y_2, z_2), which represent joints in the corresponding bone, are projected to image coordinates (u_1, v_1) and (u_2, v_2), respectively. If the scale factor s of the projection model is known, it would be a simple matter to compute the relative depth of the two joints, denoted by $\Delta z = z_1 - z_2$, using the following equation [14]:

$$\Delta z^2 = l^2 - \frac{(u_1 - u_2)^2 + (v_1 - v_2)^2}{s^2}. \qquad (2)$$

Given the depth z_{J_1} of the first joint J_1 in the hierarchical skeleton, we compute the depth variation Δz for the "bone" connecting J_1 and J_2 (and hence the depth z_{J_2}) and so on, until the extremities of the limbs are reached. However, since the sign of Δz can not be determined (i.e., we can have $z_1 > z_2$ or $z_2 > z_1$), such formulation generates a set of ambiguous 3D postures for each 2D skeleton. More precisely, if the skeleton has twenty joints, there at most 2^{20} possible postures in the worst scenario ($\Delta z_{J_i} \neq 0$ for all joints i). Fortunately, biomechanics constraints can be used to reduce ambiguity and/or eliminate impossible/implausible postures for human beings.

2.2.1. *Biomechanical Constraints for 3D Pose Estimation*

The biomechanical constraints consider the relationship of bones linked through joints, and also the angles of rotation of the bones. The constraints preserve the distances between any two joints regardless of the movement of a human body, avoiding angles in the estimated posture that are not possible for a real human being. Our assumptions and Biomechanical constraints are presented next.

i) $z_{J_1} = 0$. Assumption 1: The top of the head is always in the picture plane (i.e. defined as $z = 0$).

ii) $sign(\Delta z_{J_4}) = sign(\Delta z_{J_3}) = sign(\Delta z_{J_2})$. Constraint 1: The goal is to avoid having sequential joints alternated positively and negatively. To this end, once Δz_{J_2} is determined, Δz_{J_3} and Δz_{J_4} are set accordingly.

iii) $\Delta z_{J_5} \geq 0$ and $\Delta z_{J_9} \geq 0$. Assumption 2: We firstly assume that Δz_{J_5} and Δz_{J_9} are positive, but they can be manually defined by the user. Constraint 2: In order to avoid anthropometric inconsistencies, we include a biomechanical constraint to deal with angles between the two shoulders, defined in two vectors: $\vec{v} = J_5 - J_2$ and $\vec{u} = J_9 - J_2$. If the inner angle between $\vec{v}$ and $\vec{u}$ is smaller than 165° (according to [16]), then the sign of Δz_{P9} is opposite to Δz_{P5}.

iv) $\Delta z_{J_8} \geq 0$, $\Delta z_{J_{12}} \geq 0$, $\Delta z_{J_{16}} \geq 0 and \Delta z_{J_{20}} \geq 0$. Assumption 3: The hand and foot extremities are always positively displaced.

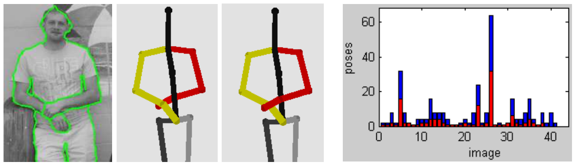
v) $\Delta z_{J_{13}} \geq 0$ and $\Delta z_{J_{17}} \geq 0$. Assumption 4: Similar to Assumption 2. Constraint 3: Analog to Constraint 2, but using the vectors between points J_{13}, J_4 and J_{17}, setting $\vec{v} = J_{13} - J_4$ and $\vec{u} = J_{17} - J_4$, and testing if the inner angle is smaller than 180° [16]. If it is true, the sign of $\Delta z_{J_{17}}$ is opposite to $\Delta z_{J_{13}}$.

The proposed constraints indeed reduce the number of ambiguities to a maximum 2^9 possible poses[1]. An additional rule we created in order to reduce the possible postures is concerned with impossible leg poses. This restriction is only applied in cases where the angle between the thigh and lower leg is larger than 180° [16], and it prevents poses with "broken knees". Indeed, in most cases, the number of possible postures is smaller than 512 due to the intrinsic nature of usual pictures, i.e. presenting many joints with $\Delta z = 0$.

[1] List of joints that can generate ambiguities if $\Delta z_{joint} \neq 0$: J_2, J_6, J_7, J_{10}, J_{11}, J_{14}, J_{15}, J_{18} and J_{19}.

2.2.2. *Ambiguity minimization using self-occlusion detection*

To further reduce the number of candidate 3D poses related to a 2D skeleton, we validate each 3D posture after applying the biomechanical constraints with the self-occlusion results. More precisely, we compute a simple edge intersection between the 2D projections of the 3D bones involved in a detected self-occlusion. Based on the depths (z coordinates) of the intersection point computed for both bones, we can verify if the 3D posture is in agreement with the self-occlusion result. If not, the posture is removed. An example is shown in Figure 3(a-c). Figure 3(a) illustrates the input image with detected body parts and occlusion detection. Figure 3(b) illustrates one of 3 incorrect 3D poses obtained by using only Taylor's approach and biomechanical constraints (the left forearm is behind the torso). The addition of the self-occlusion information discards incorrect 3D poses, and only one 3D pose (correct) remains, illustrated in Figure 3(c).



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

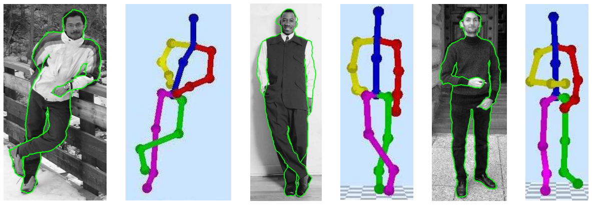
Fig. 3. (a) Detected self-occlusion. Possible poses imposing only biomechanical constraints (b) and also self-occlusions (c). (d) Number of 3D poses per image, using only biomechanics constraints (blue) and including self-occlusion (red).

3. EXPERIMENTAL RESULTS

In this section we illustrate some results of the proposed model[2]. The experimental results were generated using a dataset of 41 images containing self-occlusion. To quantitative analyze the self-occlusion estimation model, we generated ground truth data manually, where the information of self-occlusion is annotated (for each image) in a higher level (which pairs of body parts are conflicting) and also in a lower level (which body part is in front of which).

The total number of occlusions in this dataset observed by the user was 106. The proposed occlusion detection approach identified correctly 89.6% of them, of which 83.16% were correctly disambiguated. The detection procedure also produced around 10% of false positives, i.e. detection results that were not corroborated by ground truth data. It is important to point out that all the computation was done using grayscale images (as in [15]), and the use of color images could improve the segmentation results as well as the self-occlusion estimation. One limitation of the self-occlusion estimation model arises when the limbs are not approximately on the image plane (which affects the anthropometrical estimates in the projected image [15] and prevents a simple identification of the blob associated to each body part).

As for the 3D pose estimation problem, validation was performed by visual inspection, since we do not know exactly which 3D posture relates to a given 2D image. Correct poses were manually annotated in the dataset and used to numerically compare results and inform percentages of success. Figure 3(d) illustrates the number of plausible poses detected using only biomechanical constraints (blue bars) and using both biomechanical constraints and self-occlusion results (red bars). The average number of postures in these two scenarios are 7.52 and 3.04, respectively. Considering the full approach (biomechanics + occlusions), the correct pose was within the set of selected poses in 73.17% of the cases. The exclusion of the correct pose happens due to errors in the occlusion detection/disambiguation approach, usually associated to bad segmentation results. Also, in a few cases (e.g. images 39, 40 and 41 – Figure 3(d)), Taylor's work [14] conflicts with the self-occlusion detection (probably related to errors in the occlusion handling or perspective problems), resulting in an empty set of possible postures.



(a) (b) (c) (d) (e) (f)

Fig. 4. Self-occlusion and 3D pose estimation results.

4. CONCLUSION

This paper presented a new approach for self-occlusion detection and disambiguation of human figures in still images. The proposed approach is based on the detection of individual body parts and their intersection in the image domain, exploring also the expected edge-based energy along the contour of each body part. The paper also presented an application of the self-occlusion method to the problem of 3D pose estimation, including also biomechanical constraints. Experimental results indicate that self-occlusions can be successfully detected in 89.6% of the cases, from which 83.16% are correctly disambiguated. Also, we have shown that the proposed 3D pose estimation model presents a significantly smaller number of plausible postures compared to [14], retrieving the correct pose in 73.17% of the cases. Future work will concentrate on exploring color information to measure the energy of the contour, as well as in the segmentation approach.

[2] See `www.cpca.pucrs.br/icip13` for more results.

5. REFERENCES

[1] Jia-Bin Huang and Ming-Hsuan Yang, "Estimating human pose from occluded images," in *9th Asian Conference on Computer Vision*, 2010, pp. 48–60.

[2] A. Agarwal and B. Triggs, "Recovering 3d human pose from monocular images," *Pattern Analysis and Machine Intelligence*, vol. 28, pp. 44 –58, 2006.

[3] Ahmed Elgammal and Chan-Su Lee, "Inferring 3d body pose from silhouettes using activity manifold learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, Washington, USA, 2004, pp. 681–688.

[4] G. Mori and J. Malik, "Recovering 3d human body configurations using shape contexts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 7, pp. 1052 –1062, july 2006.

[5] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P.H.S. Torr, "Randomized trees for human pose detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1 –8.

[6] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1014–1021, 2009.

[7] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3d pose estimation and tracking by detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 623–630.

[8] Yi Yang and Deva Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1385–1392.

[9] E. Simo-Serra, A. Ramisa, G. Alenya, C. Torras, and F. Moreno-Noguer, "Single image 3d human pose estimation from noisy observations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2673 –2680.

[10] Leonid Sigal and Michael J. Black, "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, Washington, USA, 2006, vol. 2, pp. 2041–2048.

[11] Ibrahim Radwan, Abhinav Dhall, Jyoti Joshi, and Roland Goecke, "Regression based pose estimation with automatic occlusion detection and rectification," in *IEEE International Conference on Multimedia and Expo*, Washington, USA, 2012, pp. 121–127.

[12] Daehwan Kim and Daijin Kim, "Self-occlusion handling for human body motion tracking from 3d tof image sequence," in *Proceedings of the 1st International Workshop on 3D Video Processing*, New York, USA, 2010, pp. 57–62, ACM.

[13] Nam-Gyu Cho, Alan Yuille, and Seong-Whan Lee, "Self-occlusion robust 3d human pose tracking from monocular image sequence," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2012, pp. 254 –257.

[14] Camillo J. Taylor, "Reconstruction of articulated objects from point correspondences in a single uncalibrated image," *Computer Vision and Image Understanding*, vol. 80, pp. 349–363, 2000.

[15] J. C. S. Jacques Junior, C. R. Jung, and S. R. Musse, "Skeleton-based human segmentation in still images," in *IEEE International Conference on Image Processing*, Orlando, USA, 2012, pp. 1–4.

[16] B.M. Nigg and W. Herzog, *Biomechanics of the Musculo-skeletal System*, John Wiley and Sons, inc, 1994.