S. PIÉRARD and M. VAN DROOGENBROECK. **On the human pose recovery based on a single view**. In *International conference on pattern recognition applications and methods (ICPRAM)*, Volume 2, pages 310-315, Vilamoura, Portugal, February 2012. http://orbi.ulg.ac.be/handle/2268/103450

ON THE HUMAN POSE RECOVERY BASED ON A SINGLE VIEW

Sébastien Piérard and Marc Van Droogenbroeck
INTELSIG Laboratory, Montefiore Institute, University of Liège, Belgium
Sebastien.Pierard@ulg.ac.be, M.VanDroogenbroeck@ulg.ac.be

Keywords: Human ; Silhouette ; Pose recovery ; Projection ambiguities

Abstract: Estimating the pose of the observed person is crucial for a large variety of applications including home entertainment, man-machine interaction, video surveillance, etc. Often, only a single side view is available, but authors claim that it is possible to derive the pose despite that humans evolve in a 3D environment. In addition, to decrease the sensitivity to color and texture, it is preferable to rely only on the silhouette to recover the pose. Under these conditions, we show that there is an intrinsic limitation: at least two poses correspond to the observed silhouette. We discuss this intrinsic limitation in details in this short paper. To our knowledge, this issue has been overlooked by authors in the past. We observe that this limitation has an impact on the way previous reported results should be interpreted, and it has clearly to be taken into account for designing new methods.

1 INTRODUCTION

The interpretation of video scenes is a crucial task for a large variety of applications including home entertainment, man-machine interaction, video surveillance, etc. As most scene of interest contain people, understanding their behavior is essential. This is a challenge, because of the wide range of poses and appearances human can take. In this paper, we discuss the feasibility to recover the pose from a single view. Moreover, we assume that the decision is based only on the observed person's silhouette.

In most places, ceilings are not high enough to place a camera above the scene and to observe a wide area. Moreover, in the context of home entertainment applications, most of existing applications (such as games) require to have a camera located on top or at the bottom of the screen. Therefore we focus on a side view in the following.

Our second assumption is that the decision is based on the silhouette. As a matter of fact, most authors consider that it is preferable to decrease the sensitivity to appearance, and therefore to rely on shapes instead of colors or textures. Silhouettes can be extracted reliably from videos by background subtraction techniques (*e.g.* ViBe (Barnich and Van Droogenbroeck, 2011)). Moreover, as stated by (Agarwal and Triggs, 2006), silhouettes "encode a great deal of useful information about 3D pose".

It is well established that pose recovery is a difficult problem since the relation between silhouettes and poses is multivalued. As stressed by Poppe *et al.* (Poppe and Poel, 2006), variations in morphologies and clothing result in a family of silhouettes corresponding to the same pose, and due to the self-occlusions and to the uncertainty on the location of the silhouette boundaries (limited visual accuracy, noise, clothing, etc) similar input silhouettes can correspond to a range of poses.

But, we have noticed that a more fundamental limitation exists. Even if we knew precisely the morphology of the person present in front of the camera, and if we were able to acquire noise-free silhouettes, at least two poses correspond to the observed silhouette. It does not matter if there are self-occlusions or not. This intrinsic limitation is the cornerstone of the discussions of this paper.

1.1 Various approaches for pose-recovery

Let us first briefly remind the various approaches that exist for estimating the pose. A few surveys covering the pose recovery have been published. For example, 125 references related to the pose-recovery, and published between the years 2000 and 2006, are given in (Moeslund et al., 2006). The methods that have been proposed in the literature to recover the pose can

S. PIÉRARD and M. VAN DROOGENBROECK. **On the human pose recovery based on a single view**. In *International conference on pattern recognition applications and methods (ICPRAM)*, Volume 2, pages 310-315, Vilamoura, Portugal, February 2012. http://orbi.ulg.ac.be/handle/2268/103450

be classified into three main families: *example-based*, *model-based*, and *learning-based*.

Example-based methods use a database of silhouettes stored together with the related pose parameters. They look in the database for the silhouette that is the closest to the input silhouette, and the predicted pose is the pose stored with the selected silhouette. To some extend, it is a nearest neighbor method. The three main challenges are: (1) how to handle a very large database (the high dimensionality space of poses has to be sampled sufficiently densely); (2) how to define a distance between two silhouettes that is robust to noise, to viewpoint variations, and to various appearances of humans (morphologies, clothing, hairstyles, ...); and (3) how to pre-compute a signature for each silhouette in the database in order to speed up the nearest neighbor search or to quickly discard a large part of the database as in (Shakhnarovich et al., 2003).

Model-based methods (also named *generative methods*) maximize the similarity between the observed silhouette and a synthetic rendered silhouette corresponding to the guessed configuration (pose and orientation). The two main issues here are (1) that the problem always has many local maxima of the likelihood, and (2) that it supposes that a good initialization configuration can be guessed.

Learning-based approaches (also named *discriminative methods*) try to learn, from samples, the function that associates the value of a pose parameter (*i.e.* kinematic angle) to the input silhouette. This function is known as being the "model". Usually, there is one model per pose parameter. The goal of machine learning is twofold. Firstly, it aims at building models that correctly generalize the learning samples. Secondly, it aims at pre-computing a model instead of using the whole set of learning samples at runtime, for efficiency reasons. This can also be seen as a smooth interpolation problem in a high dimensionality space, robust to irrelevant dimensions. The main problem with the learning-based approach is that the typical behavior of regression methods (such as the *ExtRa-Trees* (Geurts et al., 2006), etc) is to achieve a compromise between all the possible solutions (this fact has been underlined by (Agarwal and Triggs, 2006), and we have also observed this fact on several occasions by ourselves). Thus, the function that associates the value of a pose parameter to the input silhouette must be single-valued to avoid irrelevant estimates. Unfortunately, this is in general not the case.

In general, methods that assume a one-to-one mapping between silhouettes and poses are forced to arbitrarily or randomly choose one possible pose or to compromise. In this paper, we show that at least two poses correspond to the observed silhouette, and that these two poses are equally likely. Therefore, estimating the underlying probabilities does not help to choose the right pose. Also, note that the methods able to deal with multiple estimates do not solve completely the problem, as they sometimes still compromise between multiple poses (this has been observed in (Agarwal and Triggs, 2005)), even when the corresponding 3D shapes are very different.

1.2 Known ambiguities

The fact that several poses may correspond to a silhouette motivated the recent development of methods predicting a set of 3D poses, but the phenomenon responsible of the ambiguities was only partially understood. Before explaining the intrinsic limitation in Section 2, we give an overview of the current knowledge.

Firstly, the depth-direction of the limbs (forward/backward) is unobservable with silhouettes. Therefore, using the algorithm of Taylor (Taylor, 2000), there are 2^n 3D skeletons corresponding to a 2D skeleton (stick-figure) with n links. It is however easy to prune this large amount of possibilities since most of the solutions found by Taylor's method are physically impossible (*e.g.* self-intersections, knee or elbow bent the wrong way, etc).

Secondly, with silhouettes, an information related to the scene layers is missing. For example, with a frontal view, it may be impossible to know if the arm is behind or in front of the torso. This is due to occlusions, and therefore this problem does not exist systematically.

A more annoying problem is the following. The silhouettes of a person observed laterally are identical for mirror poses. Also, front and back views can lead to similar silhouettes. This means that the ambiguities may be related to the pose or to the orientation. With these observations, one might think that a temporal disambiguation is possible since ambiguities appear in two very particular orientations of the observed person. We show in this paper that this is *not* possible because there are much more ambiguities than these two ones. In fact, we derive a more general rule that establishes the link between the two sources of ambiguities (pose and orientation).

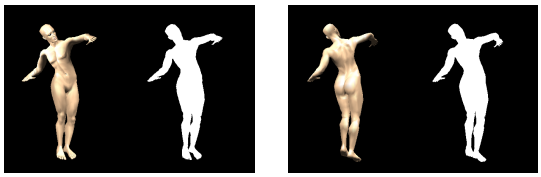
2 THE INTRINSIC LIMITATION

In this paper, we assume that the rotation axis of the observed person is parallel to the image plane. This means that the camera is looking horizontally

S. PIÉRARD and M. VAN DROOGENBROECK. **On the human pose recovery based on a single view**. In *International conference on pattern recognition applications and methods (ICPRAM)*, Volume 2, pages 310-315, Vilamoura, Portugal, February 2012. http://orbi.ulg.ac.be/handle/2268/103450

pose p_1 pose p_2

Figure 1: The poses p_1 and p_2 are mirror poses.



(p_1, θ) $(p_2, 180° - \theta)$

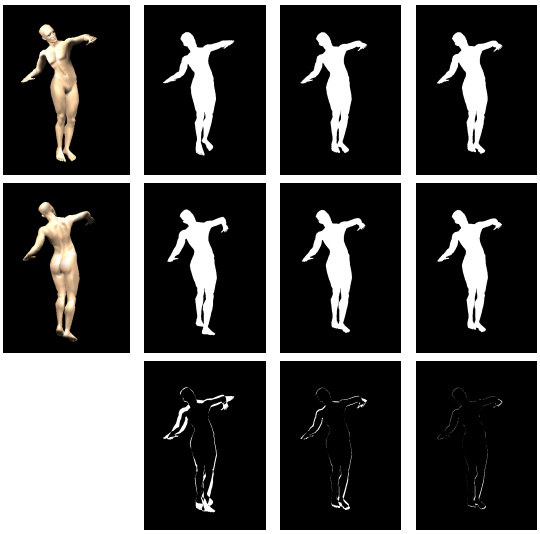
Figure 2: Two configurations leading approximately to the same silhouettes. In this figure, $\theta = 30°$.

(*i.e.* we see a side view). In the following, we adopt the following convention (without any loss of generality): the orientation $\theta = 0°$ corresponds to the person facing the camera (so, a translation of the person in space implies a change of orientation).

Under this assumption there is an intrinsic limitation of estimating the pose from a single silhouette. In most of the cases, two different poses correspond to the same silhouette. This remains true even when the precise morphology of the person present in front of the camera is known, when the silhouette is noise-free, and when there are no occlusions.

Let us consider two mirror poses p_1 and p_2 as the ones depicted in Figure 1. Obviously, they have the same probability density to be observed in any application. Note that p_1 and p_2 are neither symmetrical nor planar, and therefore, our observations are valid for all poses. As shown in Figure 2, the configurations (p_1, θ) and $(p_2, 180° - \theta)$ give rise to approximately the same silhouettes.

A closer look at silhouettes of Figure 2 shows that there are small differences between the two silhouettes. These are due to perspective effects and become negligible for large distances between the camera and the observed person. When the distance becomes infinite, the pinhole camera becomes an orthographic camera. For this limit case, the two silhouettes are, strictly speaking, identical. In practice however and despite the presence of differences, it is not possible to robustly differentiate the two silhouettes even when the camera is very close to the observed person. This is illustrated in Figure 3. Perspective effects lead to small details located on the boundaries of the silhouettes. But in a practical application, noise alters the boundaries of the observed silhouettes. Therefore, the information relative to the pose that is potentially present because of perspective effects is unusable when there is noise.



3 m 10 m 40 m

Figure 3: Magnitude of the perspective effects. For different distances between the camera and the observed person, this figure shows the silhouette corresponding to the configuration $(p_1, 30°)$ on the first row, and the silhouette corresponding to the configuration $(p_2, 150°)$ on the second row. The difference between the two silhouettes is depicted on the third row.

In summary, at least two poses must always be taken into account for each silhouette observed. This intrinsic limitation is not due to noise, nor to occlusions (*e.g.* there are no significant occlusions in Figure 2). It should also be stressed that the two poses are equally likely. The only rare cases in which only one pose has to be considered arise when the pose of the observed person is itself symmetrical.

3 DISCUSSION

3.1 Practical implications

The first practical implication of the intrinsic limitation for pose recovery systems is that once a pose has been estimated, both the pose itself and its mirrored version have to be considered. This is true for example-based, model-based, and learning-based methods.

Another implication of the intrinsic limitation concerns the learning set used in learning-based approaches. As we have already noticed in Section 1.1, there should be only one pose corresponding to a silhouette. Otherwise, the regression methods would

S. PIÉRARD and M. VAN DROOGENBROECK. **On the human pose recovery based on a single view**. In *International conference on pattern recognition applications and methods (ICPRAM)*, Volume 2, pages 310-315, Vilamoura, Portugal, February 2012. http://orbi.ulg.ac.be/handle/2268/103450

produce irrelevant estimates of the pose parameters. It is therefore necessary to discard half of the configurations in the learning set (Section 3.3 provides some indications on how to do this correctly).

3.2 Overcoming the intrinsic limitation

It is clear that we must rely on as few as possible information related to the appearance (colors and textures may be at best considered as weak cues because of their high variability and their sensitivity to lighting conditions), and that it is advised to base the decisions on geometric information. However, the silhouettes are not the only geometric information that can be obtained from a single camera.

With a color camera, one can also extract contours, but in a less reliable way than silhouettes. Unfortunately, the internal contours do not always provide enough information to overcome the intrinsic limitation. For example, if we consider the situation depicted in Figure 2, the indetermination persists because no contour can be detected inside the silhouettes. This is because there is no significant occlusion in that case.

With a range camera, it is possible to acquire silhouettes annotated with depth. Nowadays there exist cheap range cameras (for example *Microsoft's kinect*). Therefore, range cameras are a viable alternative to color cameras. It is not surprising that it is possible to estimate human poses from silhouettes annotated with depth (Shotton et al., 2011; Girshick et al., 2011) since the intrinsic limitation presented in this paper does not apply with that kind of data.

3.3 Does it help to know the orientation?

If we take again a quick look at Figure 2, we see that among the two configurations leading to the same silhouettes, the orientation of one pose is comprised in the $[-90°, 90°]$ angle range, and the other one in the $[90°, 270°]$ range. Naively, one may think that the knowledge of the orientation helps to choose the right configuration, and thus the right pose. But this is only true if the orientation is not close to $-90°$ or $90°$, because in those cases there exists still two possibilities and the knowledge of the orientation does not help to choose the right one (see Figure 4). In conclusion, knowing the orientation may help to overcome the intrinsic limitation in most of the cases, but not always.

The orientation and the pose are two independent notions. We can evaluate them simultaneously, or independently. An example of pose recovery method

$(p_1, 90°)$ $(p_2, 90°)$

Figure 4: Even when we know the orientation of the person, the intrinsic limitation still implies that there may be two different poses giving rive to the same silhouette. This happens when the orientation is close to $90°$ or $-90°$.

estimating the orientation in a first step has been proposed by Gond *et al.* (Gond et al., 2008). Another interesting procedure is the one adopted by Peng *et al.* (Peng and Qian, 2008). They start by choosing an initial pose estimate and an initial orientation estimate. Then, alternatively, they improve the estimated pose considering that the orientation is known, and they improve the estimated orientation considering that the pose is known, until convergence. Both the methods proposed by Gond *et al.* and Peng *et al.* tend to demonstrate that, as expected, knowing the orientation helps to recover the pose.

A simple, but fast and effective, method to estimate the orientation has been proposed in (Piérard et al., 2011). In that paper, we explain that a silhouette annotated with depth is necessary to estimate the orientation. Because there is the same kind of intrinsic limitation for the orientation estimation than for the pose recovery, it is impossible to estimate the orientation of the person in front of the camera by deriving it from a binary silhouette, but possible with a silhouette annotated with depth.

3.4 A few notes about the previous methods described in the literature

In this subsection, we would like to discuss results on pose recovery that were reported in the literature.

Poppe *et al.* (Poppe and Poel, 2006) compared three shape descriptors to recover the pose from a unique silhouette in an example-based approach, with a nearly horizontal camera (the elevation was chosen to be $10°$ or $20°$). From their results, it seems possible to retrieve the pose from a unique silhouette. But, they only used silhouettes corresponding to orientations in the range $[-80°, 80°]$. The success of their method is thus coherent with the intrinsic limitation described in this paper. Since we know that the orientation is between $-90°$ and $90°$, these bounds being excluded, it is possible to select the good pose out of the two possibilities.

Another pose recognition system based on a single

silhouette has been proposed by Agarwal *et al.* (Agarwal and Triggs, 2006). They announced mean angular errors of about 5° on synthetic (noise-free) data. This seems to be good results, but the problem is that they estimate the pose based on a sole silhouette, for orientations in a 360° range, and we have proved in Section 2 that this is impossible. Because their data (poses and orientation) are taken from real human motion capture sequences, three hypotheses about their learning set and their test set could explain their unusually optimistic results: (1) that the orientation is not uniformly distributed over 360°, (2) that the orientation is statistically linked to the orientation, and most likely (3) that their method takes into account the small details due to perspective effects[1]. Nevertheless, we suppose that their method should work as expected if the allowed orientations are restricted to the $]-90°, 90°[$ range.

In the description of their model-based method, Sminchisescu and Telea (Sminchisescu and Telea, 2002) explain that extracting pose from silhouettes is an under-constrained problem, and suggest to use temporal disambiguation. Similarly, Howe (Howe, 2004) predicted a set of 3D poses for each silhouette, and tried to select the right one using temporal coherence. This led to tracking failures. Following the explanation given in this paper, it is now clear that at least two motions may correspond to a sequence of silhouettes. Therefore, the temporal information does not suffice to resolve ambiguities.

In their work, Elgammal *et al.* (Elgammal and Lee, 2009) assume that the motion is known (*e.g.* walking, running, performing a golf swing or kicking). They estimate the 3D pose and the viewpoint from a single silhouette (or edges). This is done by learning both the visual manifold (*e.g.* the manifold related to the viewpoint) and the kinematic manifold (*e.g.* the manifold related to the pose), and using a particle filter for tracking. The joint manifold is topologically equivalent to a torus and each point on this torus corresponds to a pose and a viewpoint. This methodology allows to recover the multiple poses and viewpoints corresponding to the input silhouette since the particles may converge to multiple areas on the torus. The illustration in their paper clearly shows that the particles may converge towards two areas on the torus. Note that given a motion, every pose does not necessarily admit a symmetrical one, especially when the learning set is only populated for right-handed persons (walking is an exception).

In summary, it is very important to consider the intrinsic limitation when building learning and testing sets. Also, it is useful to consider this limitation to better understand the results reported in the literature until now and to correctly understand the limitations of the corresponding methods.

[1]We showed in (Piérard et al., 2011) how it is easy, even unintentionally, to learn the small details due to perspective effects when working with synthetic silhouettes.

3.5 On the evaluation methods

In recent years, a few methods predicting a set of 3D poses ("modes") from the silhouette have emerged (Rosales and Sclaroff, 2001; Howe, 2004; Sminchisescu et al., 2005; Agarwal and Triggs, 2005; Elgammal and Lee, 2009). In theory, these methods are better suited for the monocular human pose recovery because they are able to properly handle the ambiguities. However, when evaluating such methods, authors fall back on a deterministic method (Sminchisescu et al., 2005; Rosales and Sclaroff, 2001), for example by only measuring the error with respect to the most probable mode. Doing this, they loose the potential of their methods (the correct solution is not always the one predicted as the most probable). To avoid pessimistic results, the authors sometimes limit the ambiguity by using a learning set specific to the test sequence (Sminchisescu et al., 2005), or by using test sequences of symmetrical poses (*e.g.* the *jumping in the air while rotating* sequence in (Elgammal and Lee, 2009)). Unfortunately, this gives rise to highly optimistic results. A better solution would be, for example, to consider the most probable mode and its symmetrical pose, and to report the minimum error obtained with these two poses.

4 CONCLUSION

The pose recovery from a single silhouette acquired by a camera placed horizontally is an underdetermined problem: at least two poses may correspond to an input silhouette. Surprisingly, it seems that this intrinsic limitation has never been discussed completely by previous authors (to our knowledge).

From our point of view, it is necessary either to assume that the orientation is comprised between $-90°$ and 90° or equivalently between 90° and 270° (the bounds being excluded, this is important), or to use a range sensor in order to annotate the silhouettes with depth, or to use a method able to predict a set of 3D poses instead of a unique pose.

Note that the human pose recovery has been often compared to the hand pose recovery in the literature. The intrinsic limitation presented in this paper originates from the symmetry of the human body. A hand

does not have such a symmetry, and therefore the human pose recovery is much complex than the hand pose recovery.

The intrinsic limitation presented in this paper changes the way we should interpret some of the previous reported results, and has to be taken into account for designing new methods. It has an impact at many levels: on the choice of the sensors best suited for the pose recovery, on the choice of the approach to follow (model-based, example-based, or learning-based), as well as on the choice of the attributes to use for learning-based methods.

REFERENCES

Agarwal, A. and Triggs, B. (2005). Monocular human motion capture with a mixture of regressors. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, volume 3, San Diego, USA.

Agarwal, A. and Triggs, B. (2006). Recovering 3D human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):44–58.

Barnich, O. and Van Droogenbroeck, M. (2011). ViBe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image Processing*, 20(6):1709–1724.

Elgammal, A. and Lee, C. (2009). Tracking people on a torus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3):520–538.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Girshick, R., Shotton, J., Kohli, P., Criminisi, A., and Fitzgibbon, A. (2011). Efficient regression of general-activity human poses from depth images. In *International Conference on Computer Vision (ICCV)*, Barcelona, Spain.

Gond, L., Sayd, P., Chateau, T., and Dhome, M. (2008). A 3D shape descriptor for human pose recovery. In Perales, F. and Fisher, R., editors, *Articulated Motion and Deformable Objects*, volume 5098 of *Lecture Notes in Computer Science*, pages 370–379. Springer.

Howe, N. (2004). Silhouette lookup for automatic pose tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, volume 1, pages 15–22, Washington, USA.

Moeslund, T., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104:90–126.

Peng, B. and Qian, G. (2008). Binocular dance pose recognition and body orientation estimation via multilinear analysis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Anchorage, USA.

Piérard, S., Leroy, D., Hansen, J.-F., and Van Droogenbroeck, M. (2011). Estimation of human orientation in images captured with a range camera. In *Advances Concepts for Intelligent Vision Systems (ACIVS)*, volume 6915 of *Lecture Notes in Computer Science*, pages 519–530. Springer.

Poppe, R. and Poel, M. (2006). Comparison of silhouette shape descriptors for example-based human pose recovery. In *International Conference on Automatic Face and Gesture Recognition*, pages 541–546, Southampton, UK.

Rosales, R. and Sclaroff, S. (2001). Learning body pose via specialized maps. In *Proceedings of Neural Information Processing Systems*, pages 1263–1270, Vancouver, Canada.

Shakhnarovich, G., Viola, P., and Darrell, T. (2003). Fast pose estimation with parameter-sensitive hashing. In *International Conference on Computer Vision (ICCV)*, volume 2, pages 750–757, Nice, France.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs.

Sminchisescu, C., Kanaujia, A., Li, Z., and Metaxas, D. (2005). Discriminative density propagation for 3d human motion estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 390–397, San Diego, USA.

Sminchisescu, C. and Telea, A. (2002). Human pose estimation from silhouettes - a consistent approach using distance level sets. In *International Conference for Computer Graphics, Visualization and Computer Vision*, volume 10, pages 413–420, Plzeň, Czech Republic.

Taylor, C. (2000). Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Computer Vision and Image Understanding*, 80(3):349–363.