ESTIMATION OF HUMAN ORIENTATION BASED ON SILHOUETTES AND MACHINE LEARNING PRINCIPLES

Sébastien Piérard and Marc Van Droogenbroeck
INTELSIG Laboratory, Montefiore Institute, University of Liège, Liège, Belgium

Keywords: Human, Silhouette, Orientation, Machine learning, Regression.

Abstract: Estimating the orientation of the observed person is a crucial task for home entertainment, man-machine interaction, intelligent vehicles, etc. This is possible but complex with a single camera because it only provides one side view. To decrease the sensitivity to color and texture, we use the silhouette to infer the orientation. Under these conditions, we show that the only intrinsic limitation is to confuse the orientation θ with the supplementary angle (that is $180^\circ - \theta$), and that the shape descriptor must distinguish between mirrored images. In this paper, the orientation estimation is expressed and solved in the terms of a regression problem and supervised learning. In our experiments, we have tested and compared 18 shape descriptors; the best one achieves a mean error of 5.24°. However, because of the intrinsic limitation mentioned above, the range of orientations is limited to 180°. Our method is easy to implement and outperforms existing techniques.

1 INTRODUCTION

The real-time analysis and interpretation of video scenes are crucial tasks for a large variety of applications including gaming, home entertainment, man-machine interaction, video surveillance, etc. As most scenes of interest contain people, analyzing their behavior is essential. Understanding the behavior is a challenge because of the wide range of poses and appearances human can take. In this paper, we deal with the problem of determining the orientation of persons observed by a single camera.

To decrease the sensitivity to appearance, we propose to rely on shapes instead of colors or textures. The existence of several reliable algorithms, like techniques based on background subtraction, makes it tractable to detect silhouettes even in real-time (see (Barnich and Van Droogenbroeck, 2011) as an example). Therefore, our approach infers the orientation of a person from his silhouette (see Figure 1). Moreover, we consider the side view (instead of a top view), since it is not possible to place a camera above the observed person in most applications.

The purpose of this paper is twofold:

1. Ideally, one would want to determine an orientation angle comprised in $[0, 360^\circ[$ or equivalently in $[-180^\circ, 180^\circ[$, but it appears to be impossible to cover a range of 360°. We discuss this issue and show that the shape descriptor must distinguish between mirrored images (that is a *skew invariant* descriptor (Flusser, 2000; Hu, 1962)) to avoid a confusion between θ and $180^\circ + \theta$ angles. Moreover, we demonstrate that the working conditions mentioned above (*i.e.* a single side view silhouette) imply an intrinsic limitation: θ and $180^\circ - \theta$ orientations are equally likely in a side view silhouette. Therefore, we have to limit the angle range to $[-90^\circ, 90^\circ]$.

2. Secondly, we compare the results obtained with 18 different shape descriptors. Some of them outperform those previously reported in the literature. In addition, we have selected shape descriptors that are easy to implement, so that our method is faster than existing ones. However, because of the intrinsic limitation, we deal only with a 180° range of orientations. We explain how to solve the

65.2° -2.0° -71.5° 15.4° -47.4° -5.5°

Figure 1: Samples of our learning database. We want to derive the orientation of a person from his silhouette. This problem is solved as a regression problem in terms of supervised learning.

Piérard S. and Van Droogenbroeck M. (2012).
ESTIMATION OF HUMAN ORIENTATION BASED ON SILHOUETTES AND MACHINE LEARNING PRINCIPLES.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 51-60
DOI: 10.5220/0003729800510060

remaining underdetermination problem in order to cover a range of 360°.

The outline of this paper is as follows. Section 2 describes some applications of the estimation of the human orientation, and presents related work. Then, Section 3 explains our framework, and highlights the intrinsic limitation. In Section 4, we compare the results obtained with different sets of shape descriptors, and apply our method in the context of a practical application. Finally, Section 5 concludes this paper.

2 APPLICATIONS AND RELATED WORK

2.1 Applications of the Orientation Estimation

For home entertainment and man-machine interaction, it is useful to determine the configuration of the observed person. This configuration consists of parameters specific to the body shape (pose, morphology) and parameters related to the scene (position, orientation). The problem of determining the orientation is independent of the problem of pose estimation, but the knowledge of the orientation facilitates the determination of the pose parameters. For example, a pose-recovery method estimating the orientation in a first step has been proposed by Gond *et al.* (Gond et al., 2008).

There are many more applications to the estimation of the orientation of the person in front of the camera: estimating the visual focus of attention for marketing strategies and effective advertisement methods (Ozturk et al., 2009), clothes-shopping (Zhang et al., 2008), intelligent vehicles (Enzweiler and Gavrila, 2010), perceptual interfaces, etc.

2.2 Related Work

The different existing methods that estimate the orientation differ in several aspects: number of cameras and viewpoints, nature of the input (image, or segmentation mask), and nature of the output (discrete or continuous, *i.e.* classification or regression).

Several authors estimate the direction based on a top view (Ozturk et al., 2009; Zhang et al., 2008). As explained in Section 3.1, it is preferable to use a side view. In this case, methods based on the image instead of the segmentation mask have been proposed (Enzweiler and Gavrila, 2010; Gandhi and Trivedi, 2008; Nakajima et al., 2003; Shimizu and Poggio, 2004).

Some authors prefer to use the silhouette only to decrease the sensitivity to appearance. Lee *et al.* (Lee and Nevatia, 2007) apply a background subtraction method and fit an ellipse on the foreground blob. This ellipse is tracked, and a coarse estimate of the orientation is given on the basis of the direction of motion and the change of size. Therefore, their method requires a continually moving person. Agarwal *et al.* (Agarwal and Triggs, 2006) encode the silhouette with histogram-of-shape-contexts descriptors (Belongie et al., 2002), and evaluate three different regression methods.

Multiple silhouettes can be used to improve the orientation estimation. Peng *et al.* (Peng and Qian, 2008) use two orthogonal views. The silhouettes are extracted from both views, and processed simultaneously. The decomposition of a tensor is used to learn a 1D manifold. Then, a nonlinear least square technique provides an estimate of the orientation. Rybok *et al.* (Rybok et al., 2010) also demonstrate that using several silhouettes leads to better results. They use shape contexts to describe each silhouette separately and combine the single view results within a Bayesian filter framework. Gond *et al.* (Gond et al., 2008) used the 3D visual hull to recover the orientation. A voxel-based *Shape-From-Silhouettes* (SFS) method is used to recover the 3D visual hull.

As an alternative to the use of multiple cameras, we considered in (Piérard et al., 2011) the use of a range camera to estimate the orientation from 3D data. In that work, we addressed the orientation estimation in terms of regression and supervised learning. We were able to reach mean errors as low as those reported by state of the art methods (Gond et al., 2008; Peng and Qian, 2008), but in a much simpler way: complex methods such as camera calibration, shape from silhouettes, tensor decomposition, or manifold learning are not needed.

In this work, we focus on the possibility to estimate the orientation based on a single color camera. The only previous work (to our knowledge) that address this problem is due to of Agarwal *et al.* (Agarwal and Triggs, 2006). But, unlike these authors (who concentrate on regression methods), our work is focussed on shape descriptors. We show that it is possible to estimate the orientation from a single binary silhouette by methods as simple as those implemented in (Piérard et al., 2011). In addition, we study the theoretical conditions for the estimation of the orientation to be achievable, and demonstrate that there is an intrinsic limitation preventing working on 360°. This observation was missed by previous authors.

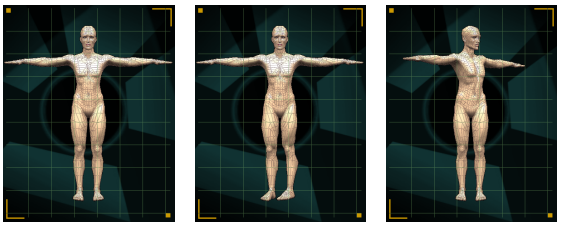


Figure 2: Defining the orientation of a person requires the choice of a body part. In this paper, we use the orientation of the pelvis. This figure depicts three examples of configurations corresponding to an orientation $\theta = 0°$. Note that the position of the feet, arms, and head are not taken into account to define the orientation.

3 OUR FRAMEWORK

In this paper, we consider a single camera that provides a side view. Moreover, to decrease the sensitivity to appearance, the only information used is that contained in silhouettes. In the following, we elaborate on our framework, define the notion of orientation for humans, and present an intrinsic limitation of orientation estimation techniques based on a single side view silhouette.

3.1 Motivations for a Side View

In most applications, it is preferable to observe the scene from a side view. Indeed most ceilings are not high enough to place a camera above the scene and to observe a wide area. The use of fisheye lenses raises a lot of difficulties as silhouettes then depend on the precise location of a person inside the field of view.

In the context of home entertainment applications, it would be possible to place a camera on the ceiling. However, most of existing applications (such as games) require to have a camera located on top or at the bottom of the screen. Therefore, if a top view is required, then it is mandatory to add a second camera, which is intractable.

3.2 Our Definition of the Orientation

There is not a unique definition of the orientation of a human. However, the orientation should not depend on the pose. Therefore, a practical way to define the orientation of a person is to choose a rigid part of the body. In this paper, we use the orientation of the pelvis. The orientation $\theta = 0°$ corresponds to the person facing the camera, with the major axis of the pelvis parallel to the image plane (see Figure 2). Another definition has been, for example, chosen by Gond *et al.* (Gond et al., 2008) who considered the torso to be the most stable body part. Indeed, these two definitions are almost equivalent and both correspond to the human intuition.

According to our definition, evaluating the orientation of the pelvis is sufficient to estimate the orientation of the observed person. But, evaluating the orientation of the pelvis is not a trivial task. As a matter of fact, one would first have to locate the pelvis in the image, and then to estimate its orientation from a small number of pixels. One of our main concerns is thus to know which body parts can be used as clues. Unfortunately, this is still an open question. Therefore, we decided to implement and to test several silhouette descriptors, some of them being global, and others focusing on the area around the centroid (see Section 4.2). Indeed, we assume that the pelvis is located in this area.

3.3 Regression Method

The machine learning method we have selected for regression is the *ExtRaTrees* (Geurts et al., 2006). It is a fast method, which does not require to optimize parameters (we do not have to setup a kernel, nor to define a distance), and that intrinsically avoids overfitting.

3.4 Intrinsic limitation of Estimating the Orientation from a Single Silhouette

In this paper, we assume that the rotation axis of the observed person is parallel to the image plane (*i.e.* we see a side view) and the projection is nearly orthographic. In other words, the perspective effects should be negligible which is an acceptable hypothesis when the person stands far enough from the camera.

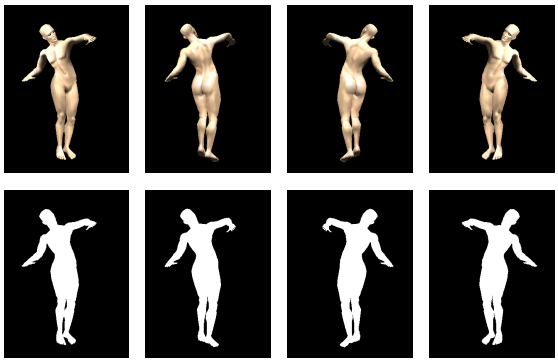
This section explains that under these assumptions there is an intrinsic limitation of estimating the orientation from a single silhouette. However, it is not our purpose to prove it rigorously. Instead, we prefer to give an intuitive graphical explanation, and to validate it with experimental results.

3.4.1 Graphical Explanation

Let us consider two mirror poses p_1 and p_2 as the ones depicted in Figure 3. They have the same probability density to be observed. If no prior information on the orientation is available, θ follows a uniform probability density function. Thus, the four cases depicted in Figure 4 have the same probability density to be observed. Hence, there is a 50% or 75% probability to be wrong depending on whether or not the silhouette

pose p_1 pose p_2

Figure 3: The poses p_1 and p_2 are mirror poses. They have the same probability density to be observed.



(p_1,θ) $(p_2,180°-\theta)$ $(p_1,180°+\theta)$ $(p_2,-\theta)$

Figure 4: Four configurations leading to similar silhouettes. These configurations have the same probability density to be observed. Note that two poses are considered here but that silhouettes are unaware of the notion of pose.

descriptor is *skew invariant* (that is whether it can distinguish between mirrored images (Hu, 1962) or not).

As shown in Figure 4, the configurations (p_1,θ) and $(p_2,-\theta)$ give rise to the same silhouettes under reflection. Moreover, if the person turns with an angle of 180°, the observed silhouette is approximately the same, under reflection (the small differences are due to perspective effects). Note that p_1 and p_2 have not be chosen to be a particular case: they are neither symmetrical nor planar. Therefore, the previous observations are valid for all poses.

Peng *et al.* (Peng and Qian, 2008) claimed that a 180° ambiguity is inherent. There are indeed some configurations (pose and orientation) for which it is impossible to discriminate between the θ and $\theta+180°$ orientations, but these configurations are statistically rare. As shown in Figure 4, it is sufficient to use a silhouette descriptor sensitive to reflections to be able to discern the angles θ and $\theta+180°$ in most of the cases. However, even if we use a *skew invariant* silhouette descriptor, there still remains an ambiguity: the configurations (p_1,θ) and $(p_2,180°-\theta)$ give rise to the same silhouette. Thus, the intrinsic limitation of estimating the orientation from a single side view silhouette is not to make a mistake of 180°, but to confuse the orientation θ with the supplementary angle $180°-\theta$ (see Figure 5). It is therefore impossible to

Figure 5: The intrinsic limitation is to confuse the orientations θ and $180°-\theta$.

estimate the orientation or the direction from a single camera, and so we should limit ourselves to orientations $\theta \in [-90°, 90°]$.

In practice, there are always perspective effects. In (Piérard et al., 2011), we have proven that when the camera is very close to the observed person, the perspective effects cannot be considered as negligible anymore. In this case, these perspective effects tend to overcome the intrinsic limitation, but not enough to reach acceptable results. Moreover, the perspective effects only impact on small details, which can be ruined by noise. This confirms that the intrinsic limitation is also valid for pinhole cameras.

3.4.2 Observations for a 360° Estimation

To understand the implications of the intrinsic limitation, it is interesting to observe what happens when we try, trivially, to estimate the orientation in a 360° range. In our preliminary tests, we tried to estimate the orientation $\theta \in [-180°, 180°[$. As in Agarwal *et al.* (Agarwal and Triggs, 2006), we did two regressions to maintain continuity –one regression to estimate $\sin(\theta)$ and the other regression to estimate $\cos(\theta)$–, and to recover θ from these values in a simple post-processing step. We found that a lot of silhouette descriptors lead to acceptable estimators of $\sin(\theta)$, but that it is impossible to estimate $\cos(\theta)$. This is illustrated in Figure 6. The reason is the following. The regression method tries to compromise between all possible solutions. Because the configurations (p_1,θ) and $(p_2,180°-\theta)$ lead to similar silhouettes, $\widehat{\sin}$ is a compromise between $\sin(\theta)$ and $\sin(180°-\theta)$, and $\widehat{\cos}$ is a compromise between $\cos(\theta)$ and $\cos(180°-\theta)$. As $\sin(180°-\theta)=\sin(\theta)$, the sine can be estimated without any problem. However, $\cos(\theta)$ and $\cos(180°-\theta)$ have opposite values, and therefore the estimated cosine may take any value between $-\cos(\theta)$ and $\cos(\theta)$.

It can be noted that, if two orthogonal views are available, one can process each silhouette separately and estimate $\sin(\theta)$ with one camera, and $\cos(\theta)$ with the other one. It is therefore not surprising that Peng *et al.* (Peng and Qian, 2008) achieve full orientation estimation based on two orthogonal cameras. Note however that Peng *et al.* use a much more complex

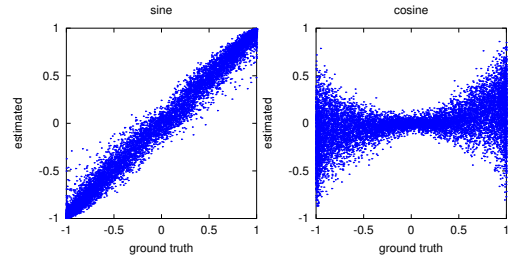


Figure 6: The typical behavior that can be observed when trying to estimate the sine and cosine of the orientation with a supervised learning method (*eg ExtRaTrees*) when the learning set contains silhouettes corresponding to orientations in the range $[-180°, 180°[$. These graphs illustrate the relationship between the real value and the estimated value of the sine and cosine. As we can see, estimating the sine is not a problem, while the estimate of the cosine is unusable. This is due to the inherent limitation of estimating orientation from a single silhouette. The fact that we observe a butterfly-like cloud shape instead of the two lines $\widehat{\cos} = \pm\cos$ is due to the compromise done by *ExtRaTrees*. The set of attributes used for regression is $R(16,400)$ (see Section 4.2.5).

approach taking into account simultaneously the two silhouettes.

4 EXPERIMENTS

4.1 Data

We found it impractical to use real data for learning the orientation estimator. Hand-labeling silhouettes with the orientation ground-truth is an error prone procedure. An alternative is to use motion capture to get the ground-truth. However, it is easy to forget a whole set of interesting poses, leading to insufficiently diversified databases. Moreover, using a motion capture system (and thus sequences) has the drawback to statistically link the orientation with the pose.

In order to produce synthetic data, we used the avatar provided with the open source software *MakeHuman* (The MakeHuman team, 2007) (version 0.9). The virtual camera looks towards the avatar, and is placed approximately one meter above the ground. For each shooting, a realistic pose is chosen (Piérard and Van Droogenbroeck, 2009), and the orientation is drawn randomly within $[-90°, 90°]$. We created two different sets of $20,000$ human silhouettes: one set with a high pose variability and the other one with silhouettes closer to the ones of a walker. They correspond to the sets $\mathcal{B}$ and $\mathcal{C}$ of (Piérard and Van Droogenbroeck, 2009) and are shown in Figure 7.

Figure 7: Examples of human synthetic silhouettes (our data sets), with a weakly constrained set of poses (upper row) and a strongly constrained set of poses (lower row).

Each of these sets has been equally and randomly divided into two parts: a learning set and a test set.

4.2 Silhouette Description

In order to use machine learning algorithms, silhouettes have to be summarized in a fixed amount of information called *attributes*.

The attributes suited for our needs have to satisfy invariance to small rotations, to uniform scaling, and to translations. This gives us the guarantee that the results will be the same even if the camera used is slightly tilted, or if the precise location of the observed person is unknown. The most common way to achieve this is to apply a normalization in a pre-processing step: input silhouettes are translated, rescaled, and rotated before computing their attributes. To achieve this, we use the centroid for translation, a size measure (the square root of the silhouette area) for scaling, and the direction of the first principal component (PCA) for rotation. As we expect people to appear almost vertically in images, we can safely choose the orientation of the silhouette from the direction of the first principal component.

Once the pre-processing step described hereinbefore has been applied, we compute the attributes on the normalized silhouette. One could imagine taking the raw pixels themselves as attributes, but this strategy is not optimal. The first reason is that it gives rise to a huge amount of attributes, which is difficult to manage with most machine learning methods. And the second reason is that (as it is highlighted by our results) machine learning methods have –in general– difficulties to exploit information given under that form. Therefore, we need to describe the silhouettes.

A wide variety of shape descriptors has been proposed for several decades (Loncaric, 1998; Zhang and Lu, 2004), but most of them have been designed to be insensitive to similarity transformations (*i.e.* uniformly scaling, rotation, translation, and reflection). As a consequence, they are not *skew invariant*, but we have explained in Section 3.4 that it is important to use a *skew invariant* shape descriptor! Therefore,

raw $\times 1$ raw $\times 2$ raw $\times 3$ raw $\times 4$

Figure 8: Four variants of the raw descriptors.

there are a lot of available descriptors not suited to meet our needs, or that would require modifications.

We have compared the results obtained by several *skew invariant* shape descriptors that are fast to compute and easy to implement. Our goal is to determine which shape descriptors contain the information related to the orientation of the observed person and that are suitable for machine learning methods. In all cases, the pre-processing step described hereinbefore was applied. Several families of *skew invariant* descriptors are detailed hereafter.

4.2.1 Raw Descriptors

Our raw descriptors are quite simple. The idea is to let the learning algorithm decide by itself which shape characteristics are most appropriate. Therefore, the attributes are the raw pixel values of a 80×80 pixels image centered on the gravity center of the silhouette. This leads to 6400 binary attributes. Depending on the size of the region which is captured around the center, several variants are considered (see Figure 8). This allows us to focus on the region around the pelvis.

4.2.2 Descriptor based on the Principle of a Histogram

We have tried to merge the pixels of our small square images into larger rectangular regions. To achieve this, the image is sliced horizontally and vertically; each slice width increases with the distance to the centroid to focus on the region around the pelvis. We count the number of pixels of the silhouette that fit into each rectangular box. There are 20 horizontal slices and 20 vertical slices leading to 400 attributes. Figure 9 shows the borders of the slices.

4.2.3 Moments

We have also implemented several statistical moments. First, we tried the 7 moments introduced by Hu (Hu, 1962), which have been selected to be rotation invariant. Flusser (Flusser, 2000) demonstrated that Hu's system of moment invariants is dependent and incomplete, and proposed a better set of 11 rotation invariant moments. Therefore we also tried this set. Finally, we tried the 12 central moments (which are not rotation invariant) of order two, three, and four.

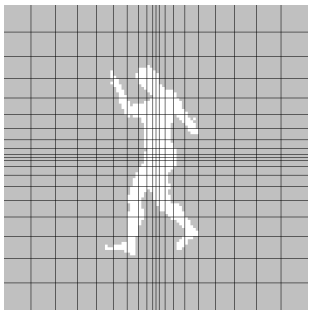


Figure 9: Borders of the rectangular boxes considered in our "histogram"-like descriptor.

Note that only a few moments are *skew invariant* (Flusser, 2000; Hu, 1962). Unfortunately, we have no way to encourage the learning algorithm to use mostly the *skew invariant* descriptors. Future work will consider a weighting mechanism to adapt the significance of pixels according to their position relative to the centroid.

4.2.4 Fourier Descriptors

We have selected two popular types of Fourier descriptors: those computed from a signal related to curvature (Zahn and Roskies, 1972), and those derived from the direct use of complex coordinates. Usually, the spectrum is not used as such to define attributes. Attributes invariant to rotation, translation, scaling, and to the choice of the initial contour point are extracted from the complex values of the spectrum. However, this methodology leads to descriptors which are not *skew invariant*. Therefore, we keep all the spectrum information to define attributes. After all, a normalization has already been performed in our pre-processing step, and all we have to do is to systematically start describing the contour at, for example, the top most boundary point. The attributes are the real and imaginary parts of the 41 lowest frequencies.

4.2.5 Descriptors based on the Radon Transform

We have used a subset of the values calculated by a radon transform as attributes. Radon transform consists in integrating the silhouettes over straight lines. $R(x,y)$ denotes such a subset, where x is the number of line directions, and y is the number of line positions for a given direction.

4.2.6 Descriptors based on the Shape Context

Shape contexts have been introduced by Belongie *et*

*a*l. (Belongie et al., 2002) as a mean of describing a pixel by the location of the surrounding contours. A shape context is a log-polar histogram. In our implementation, we have a sole shape context centered at the gravity center which is only populated by the external contour. We denote $SC(x,y)$ a shape context with x radial bins and y sectors. Belongie *et al.* (Belongie et al., 2002) use $SC(5,12)$, but other configurations have been chosen by other authors. Therefore, we have tested several configurations.

4.3 First Experiment: Choosing a Shape Descriptor

In this section, we report the results obtained with the descriptors that have been previously mentioned. We are not interested in combining different descriptors, because (i) the resulting method would be unnecessarily time-consuming, (ii) it would be difficult to interpret the results, and (iii) the number of combinations would be too large to consider them all in our experiments.

Because of the intrinsic limitation, our learning set and test set have been populated only with silhouettes corresponding to orientations in the range $[-90°, 90°]$. The results obtained with real data depend on the conditions under which data is acquired, and the background subtraction algorithm chosen. We prefer to draw conclusions that are not biased by the conditions in which the data acquisition is performed. Therefore, we ran our first experiment on synthetic data.

It has been showed in (Piérard et al., 2011) that the perspective effects may be useful to overcome the intrinsic limitation. Therefore, we hoped to obtain better results with a pinhole camera than with the orthographic camera considered in the theoretical considerations of Section 3.4. We have thus led our experiments with a pinhole camera located at several distances from the avatar. The vertical opening angle of the camera has been adjusted accordingly to keep a silhouette of the same size. The selected distances are 3 m (with a vertical field of view of 50°), 20 m (with a vertical field of view of 8°), and ∞ (with a vertical field of view of 0°, *i.e.* an orthographic camera).

The mean error results are provided in Table 1 for both the sets of weakly and strongly constrained poses. The mean error is defined as $\mathrm{E}[|\theta - \hat{\theta}|]$, where $\mathrm{E}[\cdot]$ denotes the mathematical expectation (the same error measure has been used in (Agarwal and Triggs, 2006) and (Piérard et al., 2011)). Four conclusions can be drawn from these results:

1. Taking the raw pixels themselves as attributes is not an optimal strategy. Using a carefully chosen shape descriptor may improve the results. Among the 18 *skew invariant* shape descriptors that we have considered, three families of descriptors perform very well: the Radon transform, our descriptor based on the principle of a histogram, and a shape context located at the gravity center.
2. The diversity of the poses in the learning set has a negative impact on the result. This observation corroborates those of (Piérard et al., 2011).
3. The distance between the camera and the avatar has only a slight impact on the results (the general trend is that perspective effects slightly alter the results). So, for the estimation of the orientation from a single binary silhouette in the range $[-90°,90°]$, the camera can be placed at any distance from the person. But, of course, the learning set has to be taken accordingly.
4. With synthetic silhouettes, it is possible to obtain very accurate estimations of the orientation. We do not think that it is possible to do much better, because our results are already much more accurate than the estimates a human expert could provide. Indeed, according to (Zhang et al., 2008), the uncertainty on the orientation estimation given by a human expert is approximately about 15°.

Our results are difficult to compare with those reported for techniques based on a classification method, such as the one proposed by (Rybok et al., 2010), instead of a regression mechanism. Therefore, we limit our comparison to results expressed in terms of an error angle. However, one should keep in mind that a perfect comparison is impossible because the set of poses used has never been reported by previous authors. The following results are reported in the literature. It should be noted that, like for our experiments, these results were obtained for learning sets and test sets populated with synthetic silhouettes.

- Gond *et al.* (Gond et al., 2008) obtained a mean error of 7.57° using several points of view.
- Peng *et al.* (Peng and Qian, 2008) reported 9.56° when two orthogonal views are used.
- Agarwal *et al.* (Agarwal and Triggs, 2006) obtained a mean error of 17° from monocular images (binary silhouettes). But the problem is that they estimate the orientation on 360° based on a sole silhouette, and we have explained in Section 3.4 that this is impossible. Because their data (poses and orientation) are taken from real human motion capture sequences, three hypotheses could explain their results: (1) that the orientation is not uniformly distributed over 360°, (2) that the orientation is statistically linked to the pose, and (3) that

Table 1: The mean error obtained with 18 shape descriptors to estimate the orientation.

		weakly constrained poses			strongly constrained poses		
		pinhole at 3 m	pinhole at 20 m	ortho-graphic	pinhole at 3 m	pinhole at 20 m	ortho-graphic
silhouette descriptor	$R(16,400)$	8.45°	7.37°	7.18°	5.24°	4.87°	4.88°
	$R(8,400)$	8.56°	7.60°	7.39°	5.28°	4.99°	4.92°
	"histogram"-like descriptor	8.57°	7.44°	7.31°	5.74°	5.45°	5.42°
	$R(4,400)$	10.51°	9.36°	9.17°	5.66°	5.32°	5.28°
	$SC(8,12)$	10.96°	9.55°	9.22°	6.72°	6.22°	6.22°
	$SC(5,12)$	12.55°	10.62°	10.23°	7.49°	7.01°	6.99°
	$SC(8,8)$	13.60°	11.52°	10.96°	7.41°	7.12°	7.05°
	raw $\times 2$	14.23°	12.92°	12.49°	8.84°	8.54°	8.29°
	raw $\times 4$	14.40°	12.40°	12.41°	10.02°	9.45°	9.02°
	raw $\times 3$	14.47°	12.35°	12.33°	9.51°	9.11°	8.60°
	raw $\times 1$	15.02°	12.90°	12.95°	9.14°	8.95°	8.94°
	$SC(5,8)$	16.37°	13.54°	13.30°	7.83°	7.45°	7.52°
	curvature Fourier descriptors	22.55°	23.02°	23.14°	13.10°	12.72°	12.94°
	complex Fourier descriptors	24.92°	25.50°	24.44°	12.31°	12.44°	12.39°
	$R(2,400)$	29.04°	27.37°	26.84°	13.74°	13.02°	12.83°
	central moments	35.13°	31.91°	30.69°	20.16°	18.93°	18.40°
	Flusser moments	44.88°	44.96°	45.01°	43.73°	44.34°	44.40°
	Hu moments	45.50°	45.34°	45.19°	45.02°	44.73°	45.04°

their method takes small details due to perspective effects into account (see (Piérard et al., 2011)).

The results reported by Gond *et al.* and Peng *et al.* are of the same order of magnitude as ours, but our method is much simpler. However, because we use only one point of view, we are limited to a 180° range whereas the results reported by them relate to a 360° range estimation. But we think that our method could also be used to estimate a full range orientation in an effective way. Indeed, the orientation estimations obtained independently from two orthogonal views could be fused during a simple post-processing step. Whether the use of two views allows one to decrease the mean error or just to resolve the inherent ambiguity is currently an open question. As already explained in (Piérard et al., 2011), another possible solution to the underdetermination is to use a range camera.

4.4 Second Experiment: Observations for a Practical Application

In order to evaluate our method for real world applications (which motivates our work), real images have to be considered instead of the synthetic data used in our first experiment. In this second experiment, the model used to estimate the orientation of the observed person is still learned from synthetic data, but the test set contains real silhouettes.

4.4.1 The Application

We applied our method to a real application driven by a color camera. The estimated orientation has been applied in real time to an avatar, and projected on a screen in front of the user. This allowed a qualitative assessment (see Figure 10). The acquisition of ground-truth data for a quantitative evaluation would require to use motion capture, which is out of the scope of this paper.

A state of the art background subtraction method named "ViBe" (Barnich and Van Droogenbroeck, 2011) has been used to extract the silhouettes of the person in front of the camera. Such a method provides clean silhouettes, with precise contours. Also, the selected background subtraction method intrinsically ensures a spatial coherence. A morphological opening was applied to remove isolated pixels in the foreground mask. No shadow detection method has been implemented, but this should be done in a real application in order to suppress shadows from the foreground if needed.

4.4.2 The Learning Set

The fundamental questions that arise are related to the contents of the learning set. What poses should be included in the learning set: strongly constrained poses, weakly constrained poses, or a mixture of both? Which morphology (or morphologies) must be given to the avatar to build the learning set? Is it nec-

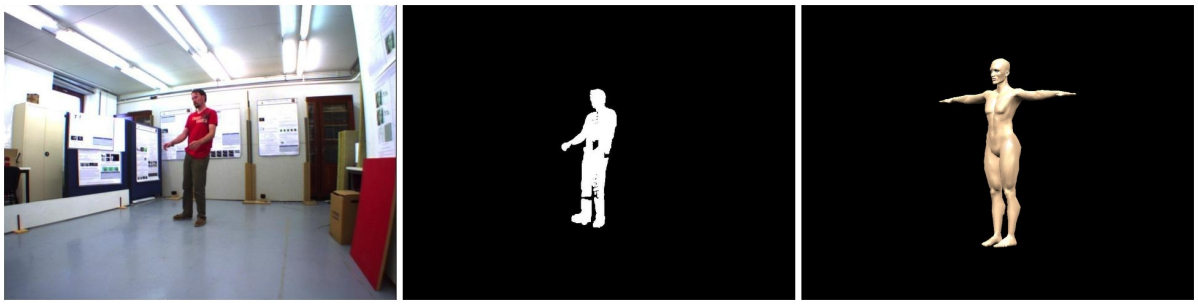


Figure 10: A screen capture of the application used to assess quantitatively our method on real data. From left to right: the input image, the result of the background subtraction, and the estimated orientation applied to an avatar. The full video is available at http://www.ulg.ac.be/telecom/orientation/.

essary to use an avatar with hair and clothes, or can we do something useful with the avatar of *MakeHuman*? Of course, the content of the learning set should reflect the situations that may be encountered in the target application.

In this experiment, we built the learning database as follows. We excluded loose-fitting clothing, thus a nude avatar such as *MakeHuman* can be used. Moreover, we populated the learning sets with 90% of strongly constrained poses and 10% of weakly constrained poses. And, finally, we used 8 different morphologies of avatars to be able to handle the morphology of the person.

4.5 The Results

The results of our method applied to a real sequence are available at http://www.ulg.ac.be/telecom/orientation/. According to our first experiment, the following three shape descriptors have been evaluated: $R(16,400)$, our "histogram"-like descriptor, and $SC(8,12)$.

The differences between synthetic data and real data are important: (i) the avatar we used to build our learning sets does not have any clothes and is hairless, (ii) the synthetic silhouettes are free of noise. However, it appears that it is possible to learn models able to estimate the orientation of the performer.

The model learned with the descriptor based on the Radon transform is efficient, and outperforms the models learned with the other descriptors (for example the one based on the shape context). This is not surprising since our first experiment selected the descriptor based on the Radon transform as the most suitable descriptor to use with machine learning methods such as the *ExtRaTrees*. Also, we expect surface-based descriptors (such as the Radon transform) to be more robust to noise than boundary-based descriptors (such as the shape context) because, for binary silhouettes, the noise alters contours significantly.

Unlike what we have done in (Piérard et al., 2011), we found that (in this case) it is not necessary to apply a temporal filtering to the orientation signal to avoid the oscillations of the avatar. This is probably because the real silhouettes were noisy in (Piérard et al., 2011) and that they are relatively clean in this work (the difference is due to the different kind of the sensors used to acquire the silhouettes).

5 CONCLUSIONS

Estimating the orientation of the observed person is a crucial task for a large variety of applications including home entertainment, man-machine interaction, and intelligent vehicles. In most applications, only a sole side view of the scene is available. To decrease the sensitivity to appearance (color, texture, ...), we consider the silhouette only to determine the orientation of a person. Under these conditions, we studied the limitations of the system, and found that the only intrinsic limitation is to confuse the orientation θ with $180° - \theta$; poses are different but silhouettes are unaware of poses. Therefore, the orientation is limited to the $[-90°, 90°]$ range. Furthermore, we have demonstrated that the shape descriptor must distinguish between mirrored images.

We addressed the orientation estimation in terms of regression and supervised learning with the *ExtRaTrees* method. To obtain attributes, we have implemented and tested 18 shape descriptors. We were able to reach low mean error, as low as 8.45° or 5.24° depending on the set of poses considered. Our results are of the same order of magnitude as those previously reported in the literature, but our method is faster and easier to implement.

If a full range orientation estimation is required, two solutions could be considered. A depth camera can be used. As an alternative, two orientation estimations (*eg* the sine and the cosine) could be obtained

independently from two orthogonal views, and fused during a simple post-processing step.

ACKNOWLEDGMENTS

S. Piérard has a grant funded by the FRIA. We are grateful to Jean-Frédéric Hansen and Damien Leroy for sharing their ideas.

REFERENCES

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.

Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522.

Enzweiler, M. and Gavrila, D. (2010). Integrated pedestrian classification and orientation estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 982–989, San Francisco, USA.

Flusser, J. (2000). On the independence of rotation moment invariants. *Pattern Recognition*, 33(9):1405–1410.

Gandhi, T. and Trivedi, M. (2008). Image based estimation of pedestrian orientation for improving path prediction. In *IEEE Intelligent Vehicles Symposium*, Eindhoven, The Netherlands.

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

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.

Hu, M. (1962). Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8(2):179–187.

Lee, M. and Nevatia, R. (2007). Body part detection for human pose estimation and tracking. In *IEEE Workshop on Motion and Video Computing (WMVC)*, Austin, USA.

Loncaric, S. (1998). A survey of shape analysis techniques. *Pattern Recognition*, 31(8):983–1001.

Nakajima, C., Pontil, M., Heisele, B., and Poggio, T. (2003). Full-body person recognition system. *Pattern Recognition*, 36(9):1997–2006. Kernel and Subspace Methods for Computer Vision.

Ozturk, O., Yamasaki, T., and Aizawa, K. (2009). Tracking of humans and estimation of body/head orientation from top-view single camera for visual focus of attention analysis. In *International Conference on Computer Vision (ICCV)*, pages 1020–1027, Kyoto, Japan.

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.

Piérard, S. and Van Droogenbroeck, M. (2009). A technique for building databases of annotated and realistic human silhouettes based on an avatar. In *Workshop on Circuits, Systems and Signal Processing (ProRISC)*, pages 243–246, Veldhoven, The Netherlands.

Rybok, L., Voit, M., Ekenel, H., and Stiefelhagen, R. (2010). Multi-view based estimation of human upper-body orientation. In *IEEE International Conference on Pattern Recognition (ICPR)*, pages 1558–1561, Istanbul, Turkey.

Shimizu, H. and Poggio, T. (2004). Direction estimation of pedestrian from multiple still images. In *IEEE Intelligent Vehicles Symposium*, pages 596–600, Parma, Italy.

The MakeHuman team (2007). The MakeHuman website. http://www.makehuman.org.

Zahn, C. and Roskies, R. (1972). Fourier descriptors for plane closed curves. *IEEE Transactions on Computers*, 21(3):269–281.

Zhang, D. and Lu, G. (2004). Review of shape representation and description techniques. *Pattern Recognition*, 37(1):1–19.

Zhang, W., Matsumoto, T., Liu, J., Chu, M., and Begole, B. (2008). An intelligent fitting room using multi-camera perception. In *International conference on Intelligent User Interfaces (IUI)*, pages 60–69, Gran Canaria, Spain. ACM.