ORIENTED RADIAL DISTRIBUTION ON DEPTH DATA: APPLICATION TO THE DETECTION OF END-EFFECTORS

Xavier Suau *Javier Ruiz-Hidalgo* *Josep R. Casas*

Universitat Politècnica de Catalunya
{xavier.suau, j.ruiz, josep.ramon.casas}@upc.edu

ABSTRACT

End-effectors are considered to be the main topological extremities of a given 3D body. Even if the nature of such body is not restricted, this paper focuses on the human body case. Detection of human extremities is a key issue in the human motion capture domain, being needed to initialize and update the tracker. Therefore, the effectiveness of human motion capture systems usually depends on the reliability of the obtained end-effectors. The increasing accuracy, low cost and easy installation of depth cameras has opened the door to new strategies to overcome the body pose estimation problem. With the objective of detecting the head, hands and feet of a human body, we propose a new local feature computed from depth data, which gives an idea of its curvature and prominence. Such feature is weighted depending on recent detections, providing also a temporal dimension. Based on this feature, some end-effector candidate blobs are obtained and classified into head, hands and feet according to three probabilistic descriptors.

Index Terms— Motion capture, Motion analysis, Machine vision, Classification, Human computer interaction

1. INTRODUCTION

Depth cameras have raised from marginal research sensors to be a true alternative to multiview strategies in the field of human motion capture. These new cameras, which work in a range between 0.5 to 6 meters, provide a very fast and handy way of obtaining 3D information from a scene. More precisely, depth cameras provide a pixel-wise depth estimation of the recorded scene. The resulting data may be considered as a 3D sampling of the visible scene surface from the camera viewpoint. Since the acquired data is restricted to a single viewpoint, it is called *2.5D data* throughout this paper.

The increasing performance of depth sensors has forested human pose estimation from 2.5D data. Knoop *et. al* [1] propose a fitting of the 2.5D data with a 3D model by means of ICP (Iterative Closest Point). Grest *et. al* [2] use a non-linear least squares estimation based on silhouette edges, which is able to track extremities in adverse background conditions. Zhu *et. al* [3] propose a tracking algorithm which exploits temporal consistency over frames to estimate the pose of a constrained human model. Whilst the above three methods focus on upper-body pose, Plagemann *et. al* [4] present a fast method which localizes body end-effectors on 2.5D data at about 15 frames per second. Ganapathi *et. al* [5] extend the work in [4] and extract full body pose by filtering the 2.5D data, using the body parts locations. Shotton *et. al* [6] presented recently the pose recognition algorithm running in the Microsoft Kinect depth sensor, which takes advantage of a large and varied dataset to carry out a pixel-wise classification into different body parts.

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°248138.

This work has been partially supported by the Spanish Ministerio de Ciencia e Innovación, under project TEC2010-18094

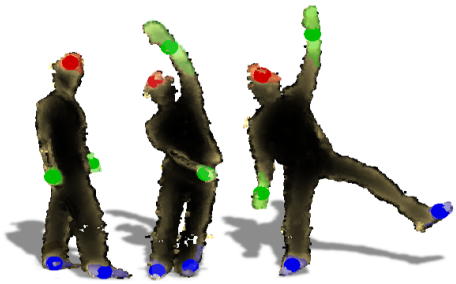

Fig. 1. Example of the proposed end-effector detection on three challenging situations. Pixels with dark values have a high *Oriented Radial Distribution* value, while light pixels resulted in low values. The obtained end-effectors are highlighted and labeled as head (red), hand (green) and foot (blue).

In this paper we propose a novel strategy to extract human end-effectors from 2.5D data (Figure 1). The method relies on a new feature, called *Oriented Radial Distribution* (*ORD*) of a pixel's neighborhood. Such feature presents high values on prominent zones of the scene surface, and low values on flat and interior zones. Therefore, it is opportune to use this feature to detect end-effectors on depth data. The most prominent end-effectors are classified into head, hand and foot according to some probabilistic descriptors which categorize their position, size and shape. The obtained end-effectors are used to reinforce the detection in further frames, providing a temporal dimension which increases the robustness of the algorithm.

The remaining of this paper is organized as follows. In Section 2 the oriented radial distribution feature is introduced. The classification step is described in Section 3, where the used descriptors are defined. Experimental results on accuracy and processing time are shown in Section 4. Finally, a concluding discussion and the direction of our future work are presented in Section 5.

2. ORIENTED RADIAL DISTRIBUTION

A 3D point cloud is obtained after mapping the depth pixels in the real world coordinate system. Such point cloud corresponds to a sampling of the visible scene surface from the camera viewpoint. With a simple foreground segmentation, using a depth threshold with

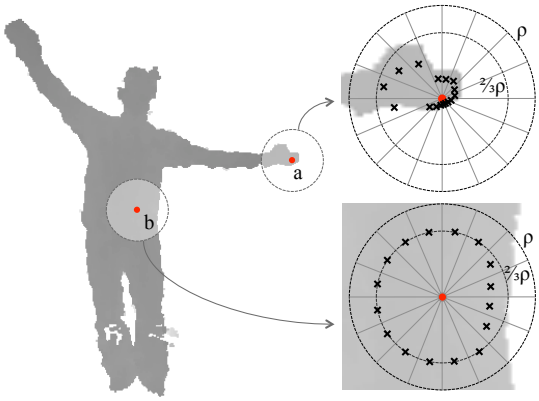

Fig. 2. Oriented Radial Distribution feature computation example. The point a belongs to an extreme, the distances between the dark crosses $\bar{\delta}_j$ and the baricenter of each zone $\frac{1}{\sqrt{2}}\rho$ present high values. On the other hand, point b belongs to a relatively uniform area, and most of the $|\bar{\delta}_j - \frac{1}{\sqrt{2}}\rho|$ distance values are very low.

respect to the empty scene (background), the subset Ω containing the foreground objects is obtained (human body, in this paper).

After simple inspection of Ω (Figure 2), one may notice that the head, hands and feet are relatively visible on the depth map. Indeed, these five end-effectors share the common characteristics of being extrema of Ω, and also of having a similar size. Therefore, being able to determine whether a pixel is located in an extremal zone of a given size is convenient to detect the end-effectors of a human body.

With this purpose, we propose a feature $\Theta : \{p, \Omega, \xi\} \longmapsto \mathbb{R}$ to measure the *Oriented Radial Distribution* of the neighborhood $\mathcal{N}_p^\rho$ around a point $p \in \Omega$ (an example is provided in Figure 2). More precisely, let the neighborhood $\mathcal{N}_p^\rho$ of p be all those points $\{p_i\} \in \Omega$ such that $|p - p_i| < \rho$. In other words, $\mathcal{N}_p^\rho$ contains all the points located in a ball of radius ρ centered at p. Therefore, the radius ρ is a parameter of the proposed feature. Such radius, and other parameters explained in this section, are noted as ξ.

The tangential plane $\mathcal{T}_p$ at p is roughly estimated by Principal Component Analysis (PCA) of the points surrounding p, the two principal axis of the PCA determining $\mathcal{T}_p$. Then, all the points $p_i \in \mathcal{N}_p^\rho$ are projected onto $\mathcal{T}_p$. Therefore, a disk $\mathcal{D}_p^\rho$ of radius ρ is obtained, containing all the projections of the points in the neighborhood of p. Projecting the neighborhood of every point in Ω onto its tangent plane is a key aspect of the proposed algorithm (Figure 3), making the feature Θ consistent over the whole point cloud Ω.

If the central point p is not located close to an extreme, it is likely to be surrounded by a regular amount of points in all directions. On the other hand, points located close to extremal zones only present neighbors in some directions. In order to measure whether point p is located close to an extremal zone of Ω or not, the content of $\mathcal{D}_p^\rho$ is analyzed. Indeed, $\mathcal{D}_p^\rho$ is divided into K equal zones as shown in Figure 2 ($K = 16$ in the example), where $K \in \xi$ is a parameter. These zones, noted Δ_j with $j = 1..K$, present two equal sides of length ρ and a third side of length $\frac{2\pi\rho}{K}$. The average distance $\bar{\delta}_j$ between the points in each zone Δ_j and the central point p is calculated, as shown in Figure 2. Therefore, a $\bar{\delta}_j$ value characterizes every zone Δ_j. Different situations may happen:

- Those Δ_j zones being completely filled with points will have a $\bar{\delta}_j$ value very close to $\frac{1}{\sqrt{2}}\rho$, which is the radius of baricenter of the zone (divides the zones into two equal areas).
- On the contrary, partially filled zones will result in a higher distance between $\bar{\delta}_j$ and $\frac{1}{\sqrt{2}}\rho$.

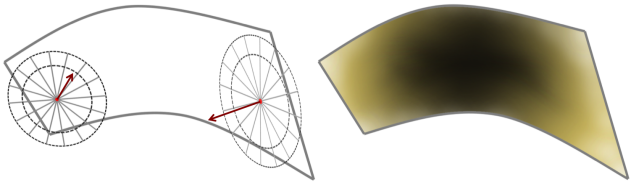

Fig. 3. Orientation of the measure disks $\mathcal{D}_p^\rho$ depending on the tangent plane to the depth surface at the measuring point (center of the disk). Two disks are illustrated in this example, and the resulting Θ feature on the right.

The *ORD* feature Θ is constructed as the average distance between the $\bar{\delta}_j$ and $\frac{1}{\sqrt{2}}\rho$, as shown in Equation (1). The obtained result is normalized with the maximal value $\frac{1}{\sqrt{2}}\rho$, so that $\Theta \in [0, 1]$. Only those zones Δ_j filled with more than 10 points are considered, resulting in a subset of K_f zones taken into account to compute Θ.

$$\Theta(p, \Omega, \xi) = \frac{1}{\frac{1}{\sqrt{2}}\rho K_f} \sum_{j=0}^{K_f} |\bar{\delta}_j - \frac{1}{\sqrt{2}}\rho| \quad with \quad \xi = \{\rho, K\} \quad (1)$$

2.1. Effect of the parameterization ξ

Two main parameters are involved in the computation of Θ, noted as $\xi = \{\rho, K\}$. The radius ρ determines the size of the neighborhood around p which will be analyzed. Indeed, the feature Θ, parameterized with a given ρ, will return its greatest values at the extrema presenting a radius similar to ρ. Thus, small radius return high values at thin extrema of Ω, while larger radius detect larger extrema. Therefore, the radius value may be used to *filter* some extrema while preserving others. For example, to find a human head, a radius of about $\rho = 15\,cm$ should be appropriate, which will filter other noisy and smaller extrema.

The second parameter in the calculation of Θ is the number of zones K into which the oriented disk $\mathcal{D}_p^\rho$ is divided. Low values of K (*i.e.* K=2) lead to noisy and non-robust detection of end-effectors, the spatial resolution of the Θ feature being too poor. On the other hand, high values of K (*i.e.* K=128) provide a too smooth transition between end-effectors and the non-desired zones. Reasonable trade-off values is between 8 and 16 divisions.

3. CLASSIFICATION OF END-EFFECTORS USING PROBABILISTIC DESCRIPTORS

A set of candidate end-effectors is obtained after the calculation of Θ on Ω. As shown in Figure 4, those points p over a threshold Θ_{min} (usually $\Theta_{min} \approx 0.15$) are labeled as candidate points. We look for those five zones with maximal Θ being large enough to be considered end-effectors. Therefore, all the very small blobs observed in Figure 4 (third column) are omitted, only keeping the largest ones, up to five candidate blobs. Remark that sometimes less than five blobs are obtained (*i.e.* in the case of a hidden hand).

Three descriptors are calculated for these candidate blobs, in order to classify them into $\gamma_i = \{head, hand, foot\}$, as in [4]. These descriptors, for a given blob $\mathcal{B}$ with centroid $\mathcal{B}^c$, are:

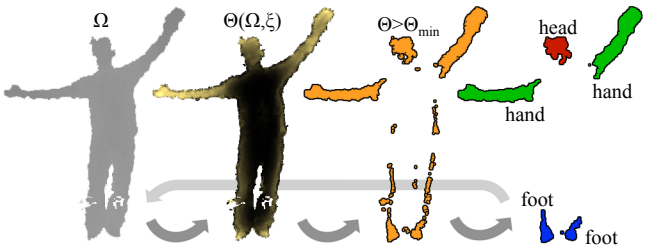

Fig. 4. Summary of the proposed method. From left to right: Segmented depth map Ω, *Oriented Radial Distribution* values Θ, candidate points $\Theta > \Theta_{min}$ and labeled end-effectors.

Y **- Position** The relative height (vertical y axis) with respect to the centroid Ω^c of Ω is calculated as: $Y = (\mathcal{B}^c_y - \Omega^c_y)\, cm$ (vertical coordinates of $\mathcal{B}^c$ and Ω^c).

S **- Size** The estimated size of $\mathcal{B}$, calculated from the apparent area of $\mathcal{B}$ on the depth image. A quadratic law Γ relating this apparent area and the physical one is obtained empirically for the Kinect sensor. More precisely, the conversion from a single pixel at depth z to a real world surface is $\Gamma_{pix}(z) \approx 1.12 \cdot 10^{-6} \cdot z^2 + 8.41 \cdot 10^{-5} \cdot z - 4.64 \cdot 10^{-3}$. Therefore, the size descriptor of a blob $\mathcal{B}$ containing N_B points is: $S = (N_{\mathcal{B}} \cdot \Gamma(\mathcal{B}^c_z))\, cm^2$.

A **- Shape** The shape descriptor is defined as the relation between the second α_2 and first α_1 eigenvalues of the PCA decomposition of $\mathcal{B}$. Thus, $A = \frac{\alpha_2}{\alpha_1}$, which gives an idea of the *roundness* of $\mathcal{B}$. Very elliptical shapes result in low A values, while very round $\mathcal{B}$ shapes result in A values near to 1.

These descriptors $\lambda_k = \{Y, S, A\}$ are analyzed over 1300 frames containing various human poses with annotated head, hand and foot parts. The statistical moments (mean μ and variance σ^2) of the obtained blobs are calculated for every extremity group $\gamma_i = \{head, hand, foot\}$ and for every descriptor (Table 1).

We propose to construct the probability density functions (PDF or f) which evaluate the probability of a given descriptor to belong to a given extremity group. Such PDF are considered gaussian, centered at $\mu^{\gamma_i}_{\lambda_k}$ with a standard deviation $\sigma^{\gamma_i}_{\lambda_k}$, as shown in Equation (2). Therefore, for each candidate blob $\mathcal{B}$ we may calculate the probability of belonging to a given group γ_i depending on the three descriptors. The combined probability of a blob belonging to a group γ_i is defined as the product of the separate probabilities of every descriptor of $\mathcal{B}$ belonging to γ_i (Equation (2)).

$$
\begin{aligned}
P(\mathcal{B} = \gamma_i) \quad &= P\Big((Y_B = \gamma_i) \wedge (S_B = \gamma_i) \wedge (A_B = \gamma_i)\Big) \\
&= f^{\gamma_i}_Y(\mathcal{B}) \cdot f^{\gamma_i}_S(\mathcal{B}) \cdot f^{\gamma_i}_A(\mathcal{B}) \\
\text{with PDF:} \quad f^{\gamma_i}_{\lambda_k}(\mathcal{B}) &= \frac{1}{\sigma^{\gamma_i}_{\lambda_k}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\lambda_k(\mathcal{B}) - \mu^{\gamma_i}_{\lambda_k}}{\sigma^{\gamma_i}_{\lambda_k}}\right)^2}
\end{aligned}
\tag{2}
$$

A decision about whether a blob belongs to any of the γ_i groups is taken, based on the obtained probabilities. Those candidate blobs with a probability of belonging to any of the groups γ_i smaller than $P(\mathcal{B} = \gamma_i) < 10^{-6}$ are not considered. The remaining candidate blobs are classified into $\{head, hand, foot\}$ depending on their probabilities, restricted to two feet, two hands and one head (Figure 4, farther right).

λ_k	$\mu^{head}_{\lambda_k}$	$\sigma^{head}_{\lambda_k}$	$\mu^{hand}_{\lambda_k}$	$\sigma^{hand}_{\lambda_k}$	$\mu^{foot}_{\lambda_k}$	$\sigma^{foot}_{\lambda_k}$
Y	62.18	7.48	29.43	29.06	-71.31	10.89
S	58.58	10.00	64.24	24.89	46.68	10.75
A	0.58	0.17	0.11	0.13	0.41	0.19

Table 1. Statistical moments of the descriptors

Temporal weighting of the Oriented Radial Distribution feature

The objective of this operation is to increase the robustness and consistence of the detection of end-effectors over time. The Θ values at time $t+1$ are weighted according to the location of the end-effectors at time t. The set of end-effectors locations is noted $\mathcal{L}$. More precisely, a distance factor $\tau \in [0,\, 0.25]$ is added to every $\Theta(p, \Omega, \xi)$ value, where τ decreases exponentially with the distance between the point p and $\mathcal{L}$, as shown in Equation (3).

$$
\tilde{\Theta}_{t+1}(p, \Omega, \xi) = \Theta_{t+1}(p, \Omega, \xi) + \tau_t \quad | \quad \tau_t = \frac{1}{4} e^{-\frac{|p - \mathcal{L}|}{10}} \tag{3}
$$

4. EXPERIMENTAL RESULTS

Experiments are obtained with the Microsoft Kinect sensor, and executed on an Intel Xeon CPU@3GHz. Color information is discarded, only using depth images. Different resolutions are tested: The Kinect original resolution of $640 \times 480\, px$, and also the downsampled version by a factor $N = 4$ ($160 \times 120\, px$ respectively).

A frame-rate of $9\, fps$ is achieved with the $N = 4$ solution, which is considered real-time. As expected, the full resolution version executes much slower, at about $0.05\, fps$.

The results presented hereafter correspond to the real-time version ($N = 4$) with a parameterization $\xi = \{\rho, K\} = \{15\, cm,\ 8\}$.

About 800 frames have been manually marked, obtaining a ground-truth sequence with head, hands and feet properly located. Such sequence contains challenging poses, as well as some frames with partial body capture (person slightly out of frame). A compilation with some poses may be consulted in Figures 1 and 5.

An end-effector is considered properly detected when the distance to the ground-truth is smaller than $30\, cm$, as proposed in [5].

The obtained *head* detection rate is of 97.7%, with an average error of $2.7\, cm$. As far as hands are concerned, none, one or two hands may appear marked in the ground-truth sequence. No distinction is considered between right and left hand, but between first (or only) hand and second hand (detection order). The first hand is detected in 90.31% of the cases, with an average error of $8.15\, cm$; while the second hand detection rate is of 76.31% with an average error of $9.2\, cm$. If an end-effector is detected as *hand*, and does not exist as ground-truth, a false positive is counted. About 8% of the overall *hand* detections are false positives in our experiments.

The detection rate of the first *foot* is of 86% with an average error of $11.2\, cm$. The second *foot* is detected in 71.6% of the cases, with an average error of $13.1\, cm$. The number of false positives is about 8.2% of de detections.

A confusion matrix is presented in Figure 6, built after the detected end-effectors, to give an overview of the precision of the classification. The values in brackets correspond to the full resolution version. Feet are the best-classified end-effectors, with only 1.07% of the detections being labeled as hands. The head is also well detected, being confused with hands about 1.12% of the times. Cross-confusion between feet and head are not observed in our experiments. Hands are more often confused, even if achieving a high

Fig. 5. Compilation of different poses in the experiment sequence. The proposed algorithm performs properly such challenging situations. Poses partially out of frame (farther left and right) are overcome. Remark that extremities are not detected when they are not prominent enough or when they are occluded (from left to right: 2nd, 7th and 8th poses).

	head	*hand*	*foot*
head	98.9% (95.8%)	1.12% (4.20%)	0% (0%)
hand	3.2% (5.1%)	96.3% (93.3%)	0.52% (1.61%)
foot	0%	1.07% (0%)	98.9% (100%)

Fig. 6. Classification confusion matrix for the real-time version, and the full resolution version (in brackets).

classification percentage of 96.3%, they are confused with the head (3.2%) and less often with feet (0.52%).

Slightly worse percentages are obtained with the full resolution version, due to the use of the $N = 4$ statistical moments in the experiment. However, the confusion matrix is still acceptable in both cases. The average errors of the full resolution version are $2.58\,cm$ (head), $4.13\,cm$ (first hand), $4.53\,cm$ (*second hand*), $5.81\,cm$ (first foot), $6.90\,cm$ (second foot). Therefore, the main advantage of the full resolution is a better average precision, which is about twice better for hands and feet.

The proposed method compares to the work in [4] and [5], since their image resolution (SR4000 TOF camera, $176 \times 144\,px$) and objective (detection of human end-effectors) are similar as ours. In [5], a tracking error between $10-20\,cm$ is achieved when extremities are visible. Our proposal obtains a detection error of about $10\,cm$, with an insignificant error on head detection. While in [5] a model-based approach is proposed (which delivers a human pose every frame) our proposal only delivers robust end-effector locations when they are visible. However, our proposal executes at $9\,fps$ on a regular CPU core, while the proposal in [5] achieves a frame-rate of $4-6\,fps$ using a specific GPU implementation. Given the differences of speed of both technologies (GPU implementations are considerably faster), the strategy proposed in this paper performs much faster.

In [4], head, hands and feet are classified without confusion about 98%, 82% and 79% of the times, respectively. These percentages are very similar to the ones obtained in this paper, except for the feet, that are much better classified in our proposal (98.9%).

5. CONCLUSION

A novel feature on depth data has been proposed in this paper, which is used to locate candidate blobs to end-effectors. A fast classification strategy which exploits the statistics of local and global descriptors of these blobs is also proposed.

Experimental results show that the proposed method performs slightly better in terms of classification than a recent reference method. Furthermore, similar average accuracy errors are obtained with a lower computational time. Such difference in computational load is emphasized, considering that the reference method in [5] is implemented on a GPU, while our proposal runs on a standard CPU.

We consider adding other descriptors to enhance the classification accuracy in the future. As well, temporal consistence is to be increased with a body model, which should be updated with a smart combination of the obtained end-effector locations with other independent features from the literature.

6. REFERENCES

[1] S. Knoop, S. Vacek, and R. Dillmann, "Sensor fusion for 3D human body tracking with an articulated 3D body model," in *International Conference on Robotics and Automation*. 2006, number May, pp. 1686–1691, IEEE.

[2] Daniel Grest, Volker Krüger, and Reinhard Koch, "Single view motion tracking by depth and silhouette information," *Lecture Notes in Computer Science*, vol. 4522, pp. 719–729, 2007.

[3] Y. Zhu, Behzad Dariush, and K. Fujimura, "Controlled human pose estimation from depth image streams," in *International Conference on Computer Vision and Pattern Recognition Workshops*. June 2008, pp. 1–8, IEEE.

[4] Christian Plagemann, Varun Ganapathi, Daphne Koller, and Sebastian Thrun, "Real-time identification and localization of body parts from depth images," in *Intl. Conference in Robotics and Automation (ICRA), 2010*. 2010, pp. 3108–3113, IEEE.

[5] Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun, "Real Time Motion Capture Using a Single Time-Of-Flight Camera," in *International Conference on Computer Vision and Pattern Recognition*. 2010, pp. 755–762, IEEE.

[6] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake, "Real-Time Human Pose Recognition in Parts from Single Depth Images," in *Computer Vision and Pattern Recognition*, Colorado Springs, 2011, pp. 1297–1304, IEEE.