# COLLABORATIVE VOTING OF 3D FEATURES FOR ROBUST GESTURE ESTIMATION

*Daniel van Sabben, Javier Ruiz-Hidalgo, Xavier Suau Cuadros, Josep R. Casas*

Universitat Politècnica de Catalunya
Image Processing Group - Signal Theory and Communications
Barcelona, Spain

## ABSTRACT

Human body analysis raises special interest because it enables a wide range of interactive applications. In this paper we present a gesture estimator that discriminates body poses in depth images. A novel collaborative method is proposed to learn 3D features of the human body and, later, to estimate specific gestures. The collaborative estimation framework is inspired by decision forests, where each selected point (anchor point) contributes to the estimation by casting votes. The main idea is to detect a body part by accumulating the inference of other trained body parts. The collaborative voting encodes the global context of human pose, while 3D features represent local appearance. Body parts contributing to the detection are interpreted as a voting process. Experimental results for different 3D features prove the validity of the proposed algorithm.

## 1. INTRODUCTION

The detection of human features such as voice and gestures allows devices to respond in human detection applications. Over the past decade, new technologies have arisen to the point of enabling efficient human-machine interaction. This is the case of affordable depth sensors for computer vision. Leaving aside color images, depth data carry spatial information that may suit better geometrical measurements for space related detection. Additionally, advances in machine learning provide better computational models that adapt to training data. Improved data and classifiers allow for better detectors in estimation problems.

This paper focuses on the detection of particular configurations of the human body, providing relevant information as a strong indicator of human gestures. Detecting body pose and gesture leads to outstanding applications in motion capture, human-computer interaction, improved surveillance, body-language interpretation, activity classification, sports monitoring, etc. The main contribution of this paper is a novel collaborative voting framework for depth images where full body pose and position of the body skeleton are jointly estimated.

The structure of the paper is as follows: the next section gives an overview of the state of the art in body pose estimation over depth data. Section 3 presents an overview of the proposed algorithm, while section 4 explains in detail the collaborative voting framework. The 3D features analyzed are explained in section 5. Finally, results and conclusions are drawn in sections 6 and 7.

This research was developed in the framework of the BIGGRAPH project –TEC2013-43935-R– funded by the Ministerio de Economía y Competitividad and the European Regional Development Fund (ERDF)

## 2. STATE OF THE ART

Hough Forests for color images [1], as proposed by Gall *et al*. [2], are a successful example for the detection of body parts. A hierarchical perspective of body parts is proposed by Navaratnam *et al*. [3]. In the work presented by Eichner *et al*. [4], an articulated human body model is used to improve the segmentation of body parts. More recently, Dantone *et al*. [5] presented a double layered model for detecting body joints. Leibe06 *et al*. [6] presented an Implicit Shape Model that combines the recognition and segmentation of objects in a common probabilistic framework.

For depth information, the work of Shotton *et al*. [7] trains a Random Forest to detect body parts. Although it requires a large training dataset (i.e. +900K images), the use of synthetic data is an interesting strategy to easily enlarge the dataset [7, 8]. López *et al*. [9] propose to detect specific body gestures by means of an unbalanced Random Forest approach. Their approach is largely real-time and robust, allowing frame-wise tracking of these gestures over time.

Other depth-based methods define an energy function specifically for depth data, eventually leading to impressive results [10, 11]. In [11], a mixed Iterative Closest Point (ICP) which takes into account physical-spatial constrains is applied to modelled body parts. Schwarz *et al*. [12] robustly detect anatomical landmarks in the 3D data and fit a skeleton body model using constrained inverse kinematics. Grest *et al*. [13] use a non-linear least squares estimation based on silhouette edges able to track limbs in adverse background conditions. While many methods focus on upper-body pose, Plagemann *et al*. [14] present a fast method which localizes body parts on 2.5D data at about 15 frames per second.

Closer to our proposal, Dantone *et al*. [15] proposed a human pose estimation system using two-layered random forests as joint regressors. Similarly, Baak *et al*. [16] proposed a solution where dataset samples are used to infer the current pose by looking for the best hypothesis that matches the current pose (based on a feature vector similarity). A generative method predicts the body pose and the final pose decision is determined by means of a voting process fusing both hypothesis components. Zeeshan Zia *et al*. [17] exploit 3D geometric class representations to recover object parts for recognition. Furthermore, a 3D mesh MoSIFT feature descriptor is used for the behavior analysis of hands and gesture in [18].

## 3. PROPOSED SCHEME

We propose a discriminative scheme for body pose estimation. First, a training phase processes point clouds and extracts a collection of templates which characterize local parts of the body. Detection is done by processing the input cloud with the same (common) training scheme and, afterwards, its output is passed onto the Collabora-

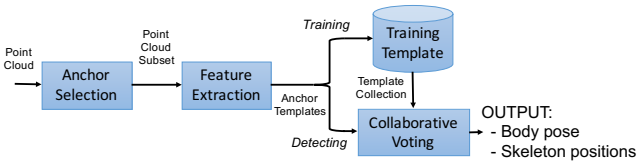


**Fig. 1**. General scheme. The functional block diagram used in both training and detection phases.

tive Voting algorithm which, using the previously trained templates, estimates a body pose.

The common scheme (Fig. 1) starts from a point cloud. In our case, the cloud is extracted from a Kinect depth camera, applying a plane clipping to remove background elements such as floor and walls. A point cloud subset is selected by random sampling as anchor points for further processing. After that, a 3D feature is computed from the anchor points neighborhood (Section 5 describes the features used). Finally, a data vector from every anchor is stored as training template.

In the detection phase, the common scheme first extracts the 3D features in a similar way, and a Collaborative Voting framework is then used to estimate the body pose together with the body joint positions.

## 4. COLLABORATIVE VOTING FRAMEWORK

The Collaborative Voting (CV) framework applies to multipart object detection. The *Voting* concept consists in inferring an object part location by an accumulation process, where each contribution can be counted as a vote, similarly to Hough accumulators. The word *Collaborative* comes from the idea that every object part cast votes to other object parts' locations, giving this sense of collaboration.

The proposed CV framework infers body joint locations in order to build a full body skeleton and a global pose ID identifying the current body gesture. Joints are found by accumulating votes from the cloud subset (anchors) contributing to joint locations (collaborative decision). Votes for each contribution are selected from templates trained based on anchors' local similarity.

In training, at the end of the common scheme, a data structure defined as training template is filled for every anchor point as shown in Fig. 2. A template is formed with: 3D anchor position, 1D feature histogram vector (see Section 5), difference vectors between body joints locations and anchor 3D position and a gesture ID number of the global pose. Note that the body joints locations are known from the groundtruth annotated in the training dataset.

In the detection phase, the same templates are filled for every anchor, but excluding the difference vectors and the pose gesture ID, which is the goal of the detection. A similarity measure is required for the detection process, as the algorithm has to select the most similar templates on the training template collection for a specific anchor. The distance between templates to find the more similar ones in the training template collection is as follows:

$$\mathcal{D}_{i,k} = w \frac{1}{\frac{1}{N}\sum_{j \in N} s_j} |s_i - s_k| + (1-w)\|y_i - y_k\|^2 \quad (1)$$

This distance is a blend of 2 factors: the normalized anchor position $s$ for all training samples $N$ and the squared distance between

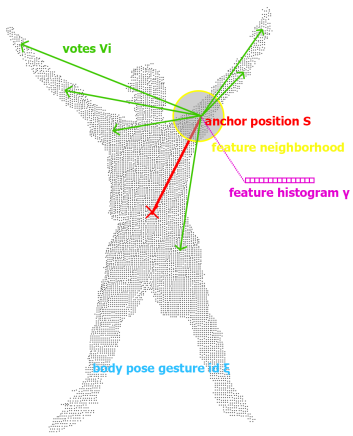


**Fig. 2**. Body anchor template components: 3D anchor position, neighboring points, votes to body joints, feature 1D histogram and body pose ID.

1D feature histograms $y$. A weighting parameter $w$ ranging from 0 to 1 is included to tune which of the 2 factors has more influence. Note that anchor position is normalized with the average anchor position of the training set samples $s_j$ to keep the same order of magnitude with the normalized 1D histogram distance.

The voting process consists in accumulating votes from each anchor towards the overall detection. First, for every anchor, a k-NN is applied (using distance $\mathcal{D}_{i,k}$) on the training templates collection to obtain the most similar templates. A vote is defined as a 3D position hypothesis where a specific body joint is located. In this context, votes are formed by adding the difference vectors (from the k-NN training templates) and the current anchor position being processed in detection. At this point, once all anchors have been processed, a collection of votes is retrieved. Next step is to accumulate all votes. For this purpose, a fixed 3D voxelization is proposed as the accumulator structure (votemap) for each body joint. The voting consists in incrementing the voxel value where a vote position falls into. The final estimated body joint location is the voxel position from its votemap with the maximum value. Note that with high voxelization resolution (small voxels) the voting process tends to increment isolated voxels and therefore leads to a non discriminative detection. To overcome this, a certain degree of smoothing is introduced by the influence of a Gaussian sphere around the votes. Fig. 3 shows an example of the Gaussian spheres voting for the hand position. Each vote increases the value of neighboring voxels around the vote position following a Gaussian decay distribution $I$, which is based on the distance between the voxel center $c$ and the vote position $v$. The variance parameter for this Gaussian decay is fixed to 12.5 $cm^2$:

$$I = e^{-\frac{1}{12.5}\|c-v\|^2} \quad (2)$$

The same idea is applied to detect the global pose ID but, instead of voxelizing, the accumulator is a 1D histogram with all possible body gestures as bins. In this case, a vote is an integer number which is the body gesture ID stored in the training template. Voting is done by increasing the histogram bin indicated by votes. Finally, the bin with maximum value is the detected global pose.

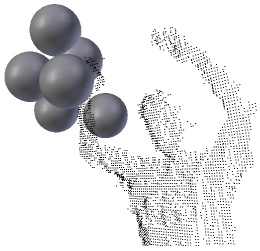

**Fig. 3**. Graphical representation of Gaussian sphere votes related to the right-hand joint.

## 5. 3D FEATURES

This section describes the 3D features used for the proposed collaborative voting framework.

We implemented the collaborative voting scheme with three different features: Curvature, Oriented Radial Distribution and Histogram of Oriented Normal Vectors. These 3D features represent better the shape of point clouds than isolated points. Formally, the features are functions of a single point $p$ in the input cloud $\mathcal{P}$, coupled with feature-specific parameters $x = [x_1, \ldots, x_n]$ as its domain, and valued either as a scalar or a fixed dimension vector $y = [y_1, \ldots, y_n]$.

$$y = f(N_{p,r}, x) \tag{3}$$

where $N_{p,r}$ are the points in the neighborhood where the feature is evaluated. These points are defined as points in $\mathcal{P}$ contained inside a sphere of radius $r$ and centered at $p$: $N_{p,r} = \{p_0 \in \mathcal{P} \,|\, \|p_0 - p\| < r\}$.

### 5.1. Curvature

Curvature $C$ is defined on the neighborhood $N_{p,r}$ as:

$$C = \frac{\lambda_0}{\lambda_0 + \lambda_1 + \lambda_2} \tag{4}$$

where $\lambda_i$ are the increasing eigenvalues of the covariance matrix of neighboring point locations, from Principal Component Analysis (PCA). In 3D, this measure indicates how much the smallest component differs from the other two. When $C$ equals zero, the points are contained in a plane. Curvature highlights the sparsity of non-planar point subsets.

**Table 1**. Average distance error associated with articulations

| Joint | Error | Joint | Error |
|---|---|---|---|
| Head | 5.7cm | Neck | 4.9cm |
| Left Shoulder | 4.8cm | Right Shoulder | 5.2cm |
| Left Elbow | 12.1cm | Right Elbow | 10.9cm |
| Left Hand | 20.2cm | Right Hand | 17.1cm |
| Left Hip | 5.7cm | Right Hip | 6.6cm |
| Left Knee | 4.4cm | Right Knee | 5.8cm |
| Left Foot | 5.4cm | Right Foot | 7.8cm |

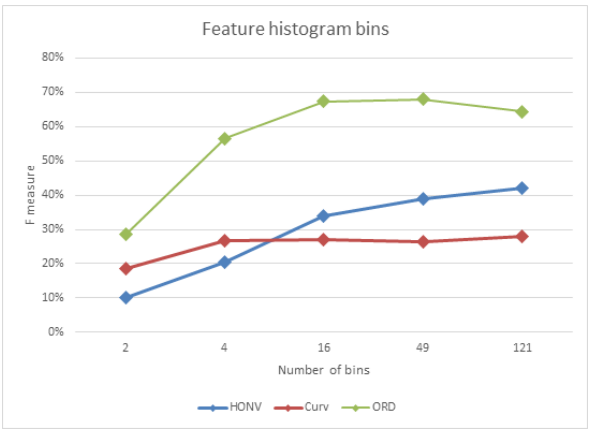


**Fig. 4**. Overall F-measure of HONV, Curvature and ORD with varying number of histogram bins evaluated on [19]

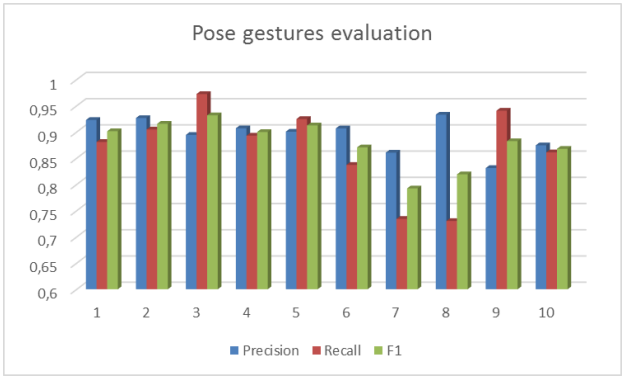


**Fig. 5**. Precision, Recall and F-measure for each pose gesture.

### 5.2. Oriented Radial Distribution

Oriented Radial Distribution (*ORD*), as proposed by Suau *et al*. [20], was designed to highlight prominent shapes in 3D space, i.e., point subsets considered as the part of a surface with non-homogeneous patterns. ORD is evaluated projecting the neighborhood onto its *tangent* plane and evaluating the homogeneity of its circular distribution. A specific filtering functionality is applied in low-density neighborhoods to avoid noise from depth sensors. ORD values for salient points in the surface are higher than those in planar regions.

### 5.3. Histogram of Oriented Normal Vectors

The Histogram of Oriented Normal Vectors (*HONV*) by Tang *et al*. [21] capture local geometric characteristics from the normal vectors of surfaces. The authors propose a coarse normal estimation based on the image gradient, with depth variation as magnitude. For point clouds in 3D space, we compute normals from PCA analysis in a smaller neighborhood, choosing the smallest component as the normal.

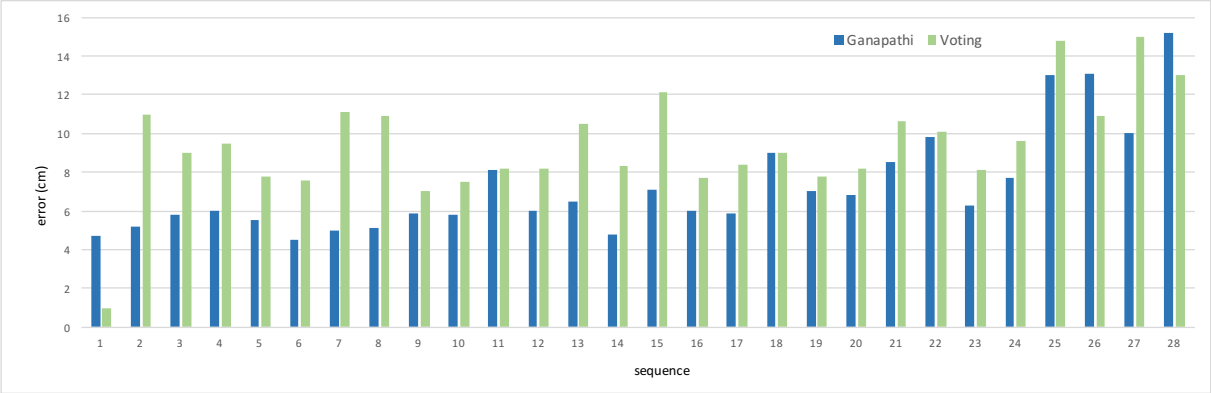


**Fig. 6**. Average detection error in centimeters.

## 6. RESULTS

### 6.1. Datasets and experiments

We evaluate the proposed collaborative voting framework using two different datasets: the *Stanford* dataset [22], as a baseline reference used to compare with the state-of-the-art, and the *UPC* dataset [19]. This second dataset features 12 subjects performing 10 different standstill body poses (or static gestures) of different complexity. The groundtruth consists of an associated label with the body pose and the positions from a frame-based articulated body model (14 joints), which makes it convenient to study the behavior of the 3D features, and to evaluate both the spatial detection accuracy, and the overall voting framework.

In both datasets, a leave-one-out cross validation strategy is used. In all experiments, the parameters are set to 20 frames per gesture, neighbors radius $r = 30cm$, Gaussian deviation of 25cm for creating the voting map and $K = 15$ nearest neighbors per anchor.

### 6.2. 3D Feature evaluation

For the dataset in [19], Fig. 4 assesses the behavior of the three 3D features varying the number of histogram bins. We set $w = 0$ in Equation 1 so only the feature information is used as the distance metric. In the case of HONV, for which the histogram is 2D, the number of bins is distributed equally in half to X bins and the rest to Y bins. The F-measure is used to evaluate the algorithm accumulating true/false and positives/negatives for the entire dataset.

ORD outperforms both Curvature and HONV in any configuration. All features tend to converge as the number of bins increase. In the ORD case, a late decay is observed due to over-fitting. Consequently, values around $7x7 = 49$ bins are suitable for evaluation, considering that ORD has not decayed and the other two features are reaching convergence.

### 6.3. Pose estimation evaluation

Fig. 5 assesses the general body pose estimation accuracy. Results are extracted considering the best configuration for the detector (i.e. 49 histogram bins in the ORD feature, $w = 0.5$ in Equation 1 and 50 random anchors in training). Fig. 5 shows the Precision, Recall and F-measure for each individual gesture. In general the system achieves a mean F-measure of 0.87 for the classification of the 10 body poses of the dataset.

Table 1 shows the average Euclidean distance between the estimated joint position of the proposed method and the groundtruth for the *UPC* dataset.

In Fig. 6, we also compare the proposed collaborative voting framework with the approach in [11] on the *Stanford* dataset. The work in [11] is more specific to the human body, and incorporates a 3D body model. On the contrary, our proposal does not exploit a body model, and would be generic enough for pose detection in other articulated objects or beings. Moreover, the Stanford dataset is composed of 28 different sequences, which makes it difficult to find training examples similar to the test ones, since we are using the complementary sequences as training for a given test sequence.

The work in [11] obtains an average error of 7.3 cm while our method obtains an average error of about 9.7 cm. Despite this inconvenient setup, the proposed Voting approach manages to select appropriate training templates, achieving good classification results with a generalized object parts detector.

## 7. CONCLUSIONS

After testing our proposed solution with the best configuration found we achieve a 0.87 F-score average on body gesture classification and an average articulation error of 9.7 cm on body skeleton estimation.

These results are competitive to current state-of-the-art techniques. Taking into account that the Collaborative Voting framework introduced in this paper exhibits a generic behavior and could be applied directly to other detection problems, we consider that the obtained results contribute positively in this scenario.

Furthermore, the proposed technique is able to estimate, at the same time, both the full body pose and the position of the body skeleton. As future work, we are investigating the use of the Collaborative Voting framework in other applications, such as hand detection and to incorporate other priors, such as color information, into the voting framework.

## 8. REFERENCES

[1] M. Van den Bergh, E. Koller-Meier, and L. Van Gool, "Real-time body pose recognition using 2d or 3d haarlets," *Interna-*

*tional Journal of Computer Vision*, vol. 83, no. 1, pp. 72–84, 2009.

[2] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2188–2202, 2011.

[3] R. Navaratnam, A. Thayananthan, P. Torr, and R. Cipolla, "Hierarchical part-based human body pose estimation," in *British Machine Vision Conference*, 2005, vol. 1, pp. 479–488.

[4] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2d articulated human pose estimation and retrieval in (almost) unconstrained still images," *International Journal of Computer Vision*, vol. 99, no. 2, pp. 190–214, 2012.

[5] M. Dantone, J. Gall, C. Leistner, and L. Van Gool, "Human pose estimation using using body parts dependent joint regressors," in *Computer Vision and Pattern Recognition*, 2013, vol. 1, pp. 3041–3048.

[6] B. Leibe, A. Leonardis, and B. Schiele, "An implicit shape model for combined object categorization and segmentation," in *Towards Category-Level Object Recognition*, pp. 496–510. Springer, 2006.

[7] J. Shotton, R. B. Girshick, A. W. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake, "Efficient human pose estimation from single depth images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2821–2840, 2013.

[8] C. Keskin, F. Kiraç, Y. E. Kara, and L. Akarun, "Hand pose estimation and hand shape classification using multi-layered randomized decision forests," in *European Conference in Computer Vision*, 2012, pp. 852–863.

[9] A. López-Méndez and J. R. Casas, "Can our TV robustly understand human gestures? real-time gesture localization on range data," in *European Conference on Visual Media Production*, 2012, pp. 18–25.

[10] M. Siddiqui and G. Medioni, "Human pose estimation from a single view point, real-time range sensor," in *Computer Vision and Pattern Recognition Workshops*, 2010, pp. 1–8.

[11] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real-time human pose tracking from range data," in *European Conference in Computer Vision*, 2012, pp. 738–751.

[12] L. A. Schwarz, A. Mkhitaryan, D. Mateus, and N. Navab, "Human skeleton tracking from depth data using geodesic distances and optical flow," *Image and Vision Computing*, vol. 30, no. 3, pp. 217–226, 2012.

[13] D. Grest, V. Krüger, and R. Koc, "Single view motion tracking by depth and silhouette information," *Lecture Notes in Computer Science*, vol. 4522, pp. 719–729, 2007.

[14] C. Plagemann, V. Ganapathi, D. Koller, and S. Thrun, "Real-time identification and localization of body parts from depth images," in *IEEE International Conference on Robotics and Automation*, 2010, pp. 3108–3113.

[15] M. Dantone, J. Gall, C. Leistner, and L.V. Gool, "Human pose estimation using body parts dependent joint regressors," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3041–3048.

[16] A. Baak, M. Müller, G. Bharaj, H. Seidel, and C. Theobalt, "A data-driven approach for real-time full body pose reconstruction from a depth camera," in *Consumer Depth Cameras for Computer Vision*, pp. 71–98. Springer London, 2013.

[17] M. Zeeshan Zia, M. Stark, B. Schiele, and K. Schindler, "Detailed 3d representations for object recognition and modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2608–2623, 2013.

[18] Y. Ming, "Human activity recognition based on 3d mesh mosift feature descriptor," in *International Conference in Social Computing*, 2013, pp. 959–962.

[19] D. Van Sabben, A. Gil, and J. Ruiz-Hidalgo, "Human body pose dataset, https://imatge.upc.edu/web/resources/body-pose-dataset," January 2016.

[20] X. Suau, J. Ruiz-Hidalgo, and J. R. Casas, "Oriented radial distribution on depth data: application to the detection of end-effectors," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 789–792.

[21] S. Tang, X. Wang, X. Lv, T. X. Han, J. Keller, Z. He, M. Skubic, and S. Lao, "Histogram of oriented normal vectors for object recognition with a depth sensor," in *Asian Conference in Computer Vision*, 2013, pp. 525–538.

[22] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 755–762.