Model-Based Hand Gesture Tracking in ToF Image Sequences

Sigurjón Árni Guðmundsson[1,2], Jóhannes R. Sveinsson[1], Montse Pardàs[3],
Henrik Aanæs[2], and Rasmus Larsen[2]

[1] University of Iceland, Department of Electrical and Computer Engineering
{sag15,sveinsso}@hi.is
[2] Technical University of Denmark, DTU Informatics
{sag,haa,rl}@imm.dtu.dk
[3] UPC-Barcelona Tech, Department of Signal Theory and Communications
montse.pardas@upc.edu

Abstract. This paper presents a Time-of-Flight (ToF) camera based system for hand motion and gesture tracking. A 27 degree of freedom (DOF) hand model is constructed and fleshed out by ellipsoids. This allows the synthesis of range images of the model through projective geometry. The hand pose is then tracked with a particle filter by statistically measuring the hypothetical pose against the ToF input image; where the inside/outside alignment of the hand pixels and the depth differences serve as classifying metrics. The high DOF tracking problem for the particle filter is addressed by reducing the high dimensionality of the joint angle space to a low dimensional space via Principal Component Analysis (PCA). The basis vectors are learned from a few basic model configurations and the transformations between these poses. This results in a system capable of practical hand tracking in a restricted gesture configuration space.

1 Introduction

Recovering the complex motions and poses of a human hand from camera observations is one of the more challenging problems in computer vision. A hand gesture tracker has many uses in the modern computer applications, to name a few: Sign language recognition, gaming interfaces where the hand gestures are used as input, navigation by pointing and special computer interfaces where no physical touching is required such as for medical applications [1].

Numerous computer vision researchers have addressed the hand tracking problem. A good review is given in Erol et al. [2]. There approaches can be roughly divided into two categories appearance-based vs. model-based. Appearance-based methods strive at mapping image features to hand poses using e.g., clustering and fast search methods [3]. Model-based approaches use a deformable model, where the model's configuration space is searched for parameters that maximize the similarity between groups of features in the input image and the model. Particle Filters

F.J. Perales and R.B. Fisher (Eds.): AMDO 2010, LNCS 6169, pp. 118–127, 2010.

(PF) have been thoroughly applied to model-based hand and human body analysis due to their abilities in non-linear estimation. In particular, PF variants that deal with high degree of freedom (DOF) of the model configuration space, are of high interest. Methods such as annealed particle filtering [4], hierarchical methods [5] and manifold methods where lower dimensional pose spaces are learned from training pose data [6].

Most of the research mentioned in [2] is based on input from a single CCD camera approaches using features such as color, edges etc. Others use multiple cameras and include depth features into the tracking.

Time-of-Flight (ToF) sensors are camera like depth measuring devises built on an active illumination modulation principle [7]. ToF cameras offer real-time simultaneous amplitude images and range images (depth measurement in each pixel).

The Swissranger SR3000 [8] used in this paper is designed to be a cost-efficient and eye-safe range imaging solution. Basically, it is an amplitude modulated near infra-red light source and a specialized 176 × 144 two dimensional sensor built in a miniaturized package. ToF cameras have been found increasingly useful for solving various computer vision applications as is reviewed in [7].

Using ToF sensors for tracking purposes has many interesting benefits. They are free from some of the problems that are present in standard intensity images such as lighting changes with shadows and reflections, color similarity and clutter. Depth is a more natural foreground / background separator than intensity and color. On the other hand the current ToF cameras' main disadvantages is the low spatial resolution, the low quality intensity image and the depth accuracy, that may have systematic errors and errors that depend on the scene. Human body tracking using ToF cameras has been studied in a few papers(cf. [7]) and one where an articulated model of the upper body is fitted to the data [9]. Hand gestures have also been studied in [10,11,1], the first attempts to fit an static computer graphics model to the ToF data and the latter two strive at recognizing static hand gestures by analyzing the segmented hands in the ToF images.

Here, a novel approach to hand gesture tracking is presented using a model-based particle filter tracker with ToF-data. By representing the models pose as range images a simple and straight forward comparison to the ToF images can be made leading to robust tracking results. We will show that the search space can be limited to chosen poses by generating instances of poses of interest and reducing the dimensionality of the joint angle space to only a few dimensional pose space via Principle Component Analysis (PCA), yielding a fast and practical hand gesture tracker.

2 The System

Fig. 1 shows an overview of the proposed system. The grey arrow path, inside the particle filter blue box, shows the path of the particles and the black arrows that come into the box are the input data and parameters, and finally the arrow out of the box is the output: the hand pose estimation for frame t. The input

ToF frame t is preprocessed and sent to the PF. In PF the posterior probability distribution is approximated by a weighted particle set, where the particles are instances of the state space. In our case, the state vector describes instances of a hand model, i.e. vector descriptions of the position in space (translation and rotation) and the joint angles of the hand (the pose). Additionally, each particle bears a weight that is updated according to how well the particle matches the input. These weights are then used to make the weighted average of the particles generating the estimate. In the remainder of this section the input and preprocessing are described, then the hand model, followed by the particle filter and pose estimation.

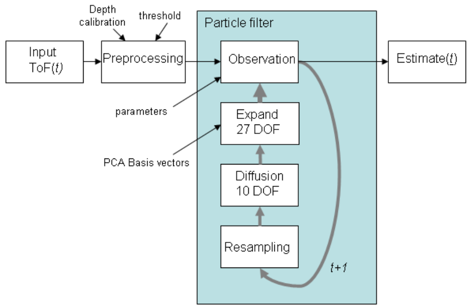

Fig. 1. Overview of the system pipeline.

2.1 Input and Preprocessing

ToF cameras have systematic depth errors that can be resolved to an extent by depth calibration [12]. Here we use the multi-camera ToF-CCD rig calibration method and tool described by Schiller et al. [12]. The tool finds an optimal higher order polynomial to compensate for the depth error and also provides the camera calibration parameters needed later for the ToF camera's projection matrix. After undistorting the input frame the hand is segmented by thresholding. Here we simply find the closest pixel, assume that it belongs to the hand and throw everything that is farther than 20 cm away from this closest pixel. This works in this ”man in front of a camera scenario”, but can easily be replaced by, e.g., a fast foreground segmentation algorithm [13] or hand detection algorithm [1] for different scenarios.

2.2 The Hand Model

The hand is modelled as a kinematic chain skeleton model similar to many other studies in hand and human body analysis [2]. The hand position and pose is described by a 27 dimensional vector where 6 dimensions are the global position (translations and rotations along and around the X, Y and Z axes) and then 5 dimensions for the thumb and 4 dimensions for the joint angles of the other

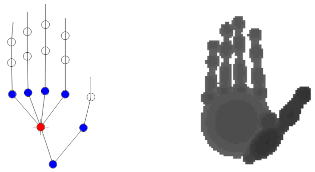

Fig. 2. The Hand model. *Left:* The kinematic chain model. Red signifies 6 DOF movement, blue 2 DOF angle movement and white 1 DOF. *Right:* A synthesized range image of a hand model instance fleshed out with ellipsoids. Darker pixels are closer to the camera.

fingers. Fig. 2 illustrates the joint DOF. Each joint 3D position in a kinematic chain is found by exponential maps and twists, i.e., simple multiplications of rotational matrices and translations as has been described in various robotics and human analysis literature, cf. [14,15].

Synthesizing Range Images of the Model Poses. A quadric is a 4 × 4 matrix $\mathbf{Q}$ which describes a surface in 3D so that all points $\mathbf{X}$ on this surface fulfill:

$$\mathbf{X}^T\mathbf{Q}\mathbf{X} = 0. \tag{1}$$

The points inside the normalized quadric give a negative result and positive on the outside. A conic is a 3 × 3 matrix $\mathbf{C}$ that has the same properties in the plane as quadrics have in 3D. A projective camera or pinhole camera model is $\mathbf{P} = \mathbf{K}[\mathbf{I}|\mathbf{0}]$, where $\mathbf{K}$ is the calibration matrix of the camera. $\mathbf{P}$ maps a 3D point $\mathbf{X} = [X, Y, Z, 1]^T$ to the pixel positions $\mathbf{x} = [x, y, 1]^T$ in the image plane by $\mathbf{x} = \mathbf{P}\mathbf{X}$.

The projective camera furthermore maps quadrics to conics in the image plane. This can be shown using the duality property of quadrics in projective geometry [16]. If the dual quadric of $\mathbf{Q}$ is mapped to $\mathbf{Q}^*$ and the dual conic of $\mathbf{C}$ is mapped to $\mathbf{C}^*$, then: $\mathbf{C}^* = \mathbf{P}\mathbf{Q}^*\mathbf{P}^T$.
The mapping of the dual conic to the conic is straight forward.

Here the skeletal hand model is fleshed out by ellipsoids which are quadrics that are thus mapped to ellipses on the ToF cameras image plane. The ellipsoids are constructed so that the axes of the ellipsoid is seen as a covariance matrix $\mathbf{V}$ with the main axis length l_1 and thickness l_2 and l_3 on the other axes, the covariance matrix is thus:

$$\mathbf{V} = \begin{bmatrix} l_1^2 & 0 & 0 \\ 0 & l_2^2 & 0 \\ 0 & 0 & l_3^2 \end{bmatrix}. \tag{2}$$

The kinematic model provides the 3D endpoint positions of the ellipsoid which give the center-point position and the 3D rotation of the ellipsoid. The quadric is thus rotated by a 3 × 3 rotation matrix $\mathbf{R}$ and translated from the origin by the 4 × 4 translation matrix $\mathbf{M}$. The ellipsoid is constructed as:

$$\mathbf{Q} = \mathbf{M}^T \begin{bmatrix} \left(\mathbf{R}^T\mathbf{V}\mathbf{R}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \mathbf{M}. \tag{3}$$

For a range image representation of the hand model configuration, the depth from the camera for each of the pixels inside the ellipses $\mathbf{C}$ need to be found. This is done by stepping along the ray towards the pixel $\mathbf{x}$. Solving the projection equation for X and Y for each Z along the ray:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \mathbf{K}^{-1}(\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \cdot Z). \tag{4}$$

These points $\mathbf{X}$ for each step on the ray are tested with the quadric equation (1) thus finding the zero crossing which is at the desired surface depth Z.

An initial point on the ray is found by using the ellipsoids center-point and main axis-length so that the surface depth is found in only few steps. Occlusion (overlapping ellipsoids) is simply handled by saving the smallest Z.

2.3 The Particle Filter

Particle filters, or often called CONDENSATION in visual tracking [17], are sequential Monte Carlo methods based on *particle* representations of probability densities. They are powerful in solving estimation and tracking problems where the variables are non-linear and non-Gaussian. In this paper the Sample Importance Resampling (SIR) approach is followed as, e.g., is described in [18]. The PF can be summarized into 3 steps: *1. Resampling* of particles, *2. Observation* of particles (weighting function) and *3. Diffusion* of particles (propagation in search space).

The resampling step is done to avoid degeneracy of the particles and here the standard procedure is followed as described in, e.g. [18]. The observation and diffusion steps are described in the following sections.

Observation: The purpose of the observation step is to find the observation likelihood of the particles: $p(X|Z_k)$, i.e., the probability of the observation X given the k^{th} particle Z_k. In our case X is the depth image after preprocessing. The observation step is usually the most expensive in the PF. A full Bayesian solution is often difficult to model so that all the aspects of the data are taken into account. Often the likelihood is replaced by more intuitive weighting function $w(X, Z_k)$. Here, the function is modelled by general statistical metrics: correct, false, missed pixel detections and an F-measure (cf. [19]). The pixels obtained with the projection of the hand model with the parameters corresponding to a given particle Z_k are here referred to as Z_k pixels.

Correct pixels: The number of Z_k pixels for which $e^{-\gamma|d_X - d_{Z_k}|} > \alpha$, where d_X and d_{Z_k} are the pixel depth values for input X and particle Z_k. The threshold α

and γ are chosen so that the pixel is classified correct if the distance is smaller than 2 cm.

False pixels: The number of Z_k pixels for which $e^{-\gamma|d_X - d_{Z_k}|} \leq \alpha$.

Missed pixels: The number of input image pixels that are in a neighbourhood region of Z_k pixels. The neighbourhood region, as shown in Fig. 3, is defined by the binary distance transform of Z_k; $DT(Z_k)$. The size of the region is controlled so that it is in proportion to the fingers length. The arm does not fall into the neighbourhood region as it is removed from the region by projecting an ellipsoid onto the wrist in $DT(Z_k)$, the wrist position is given by the kinematic model.

Fig. 3 illustrates the measurement principle and the three classes for one particle instance.

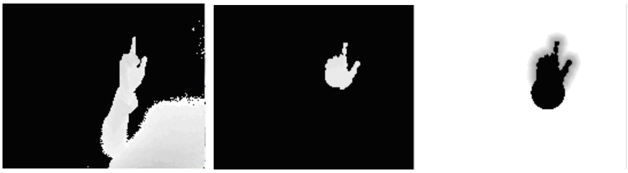

Fig. 3. The measurement of one particle hypothesis. *Left to right:* The preprocessed input X, The particle Z_k range image, Z_k's neighbourhood region in grey and classified image with correct, false and missed pixels indicated with red, green and blue.

The particles performance is measured for precision and recall. The precision is given by: $w_{prec}(X, Z_k) = \frac{\text{correct}}{\text{correct+missed}}$, and measures the exactness of the fit, while the recall; $w_{rec}(X, Z_k) = \frac{\text{correct}}{\text{correct+false}}$, measures the completeness. The final weight is then the F-measure:

$$w(X, Z_k) = \frac{(1 + \beta^2) \cdot w_{prec}(X, Z_k) \cdot w_{rec}(X, Z_k)}{\beta^2 \cdot w_{prec}(X, Z_k) + w_{rec}(X, Z_k)}. \tag{5}$$

Here β controls the balance between w_{prec} and w_{rec}, and is chosen ad-hoc to be 1.5 thus giving the recall more weight. It was seen in most typical scenes, that the missed detections were usually much fewer than the false ones and therefore needed extra penalization.

Diffusion in Subspace: It has been shown in [20] that the required number of particles for standard PF probability density estimation increases exponentially with the variable dimensionality. Standard PF can thus not handle 27 DOF hand tracking effectively. Here a reduction of dimensionality approach is followed, where a low dimensional pose space is learned from pose data. In a ”proof of concept”-experiment, synthetic data is used: The model joint angle dimensions are set to three basic poses: flat palm (or ”high 5”), fist, and pointing index finger (or ”gun”), also the basic transformations between these poses with some

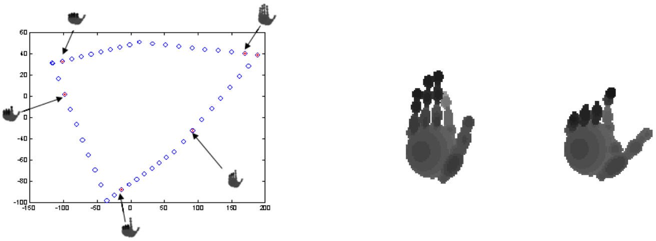

Fig. 4. *Left:* The 2 first PCs of the learning data. The 3 first PC describe 97.4% of the variance and the 6th PC added 0.1% but captured some important thumb motions. *Middle and right:* Extreme boundary poses using the maximum (middle) and minimum (right) value in all PCA basis directions. These are unlikely poses but show in a way that the poses within the subspace boundaries are not that far off.

additional thumb configurations; in total 79 poses. A PCA model was trained on these 79 points in the 21 dimensional joint angle space. Fig. 4 shows the first 2 PCs of the training data. Here, 4 PC basis vectors were used describing 97.5% of the covariance in the data.

The maximum and minimum values of the training points in each of the four dimensions, bound the pose space. Within these boundaries the particles propagate randomly according to a Gaussian density. The points, that are far off the path the training points lie on, can generate unnatural hand poses. Fig. 4 shows two of the extreme corners of the 4 dimensional hypercube. One of these poses is an unlikely pose (index finger bending backwards) but not that far from possible poses or from the poses that were used for the training.

After the propagation in the low dimensional space the particles are expanded via the 4 PC basis vectors to the full 21 dimensions where they are synthesized for the observation step.

3 Hand Tracking Results

Experiments were performed where the hand was tracked through poses in the predefined pose space. The initialization is done by a rough ”manual” positioning, and 300 particles were used in all experiments.

The results in Fig. 5 show that the tracker does a good job at catching the pose changes and out of image plane rotations, although the estimation lags somewhat. Also note that the index finger is slightly bent when it should be straight: This might be caused by the fact that this position is close to the boundary and thus the diffusion of the particles is truncated, which gives this tendency to move slightly from the boundary. Furthermore, Fig. 5 illustrates how the tracker recovers from self occlusion in large motion situations. The thumb is occluded and then it reappears and the PF detects it in the next frame. In the end of the sequence; part of the hand goes out of the frame, here the PF recovers correctly.

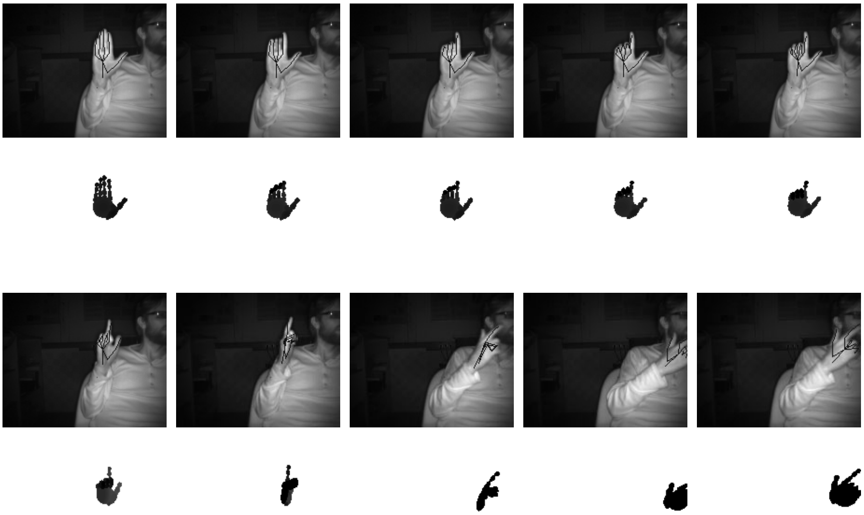

Fig. 5. 10 frames of tracking a "high 5" gesture transformed to "gun" with rotation, translation, self- and object-occlusion. *Top and 3^{rd} row:* ToF amplitude images with superimposed skeleton estimation. *2^{nd} and bottom row:* Corresponding range image of the estimated model. The transformation is successful although the model estimation is slightly lagging, i.e., the fingers should be more bent in the 2^{nd} and 3^{rd} frame . In the 8^{th} frame the PF has not recovered the thumb, but it reappears in the next frame. Here the PF had no problems when the hand partially exited the cameras field of view.

Fig. 6 shows that out of pose space gestures give of course false estimations. The weight measurement however gives a strong indication of a poor match so these cases can be classified as lost or out of limits. A recovering system can then be triggered where the PF is helped back on track by a larger number of particles and wider diffusion variance. More video examples are available on this projects homepage[1].

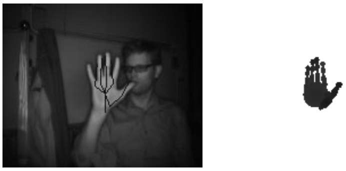

Fig. 6. Out of pose space gestures are incorrectly estimated. *Left:* ToF amplitude image with superimposed skeleton estimation. *Right:* Corresponding range image.

Currently the system runs at about 2 seconds per frame on a standard laptop PC (Core Duo 1.66 GHz, 1 Gb RAM). The implementation is done by using C++

[1] `http://www.hi.is/~sag15/handtracking.html`

libraries and has not been optimized for higher performance. We are confident that the performance can be enhanced greatly with, e.g., faster implementations of the range image construction and particle weighting. Then, the real-time goal should be achievable on newer hardware.

4 Conclusion

This paper presented a novel hand tracking system that is capable of accurately capturing the hand pose in a restricted pose space. A ToF real-time range imaging device was used so that the surface of the hand and a kinematic model were matched using 3D features in a quick and simple manner.

The main obstacle of hand pose tracking is the high DOF problem. The proposed PCA approach is simple but restricted by design. Not surprisingly the low dimensional PCA model nearly perfectly described the simple synthesized pose data used here. However, unrestricted hand motion is extremely complex, and the proposed method with manually synthesized hand pose configurations with linear transformations is not prone for success. Several researchers have used data-gloves for hand-motion capture and trained models on this data. Some have used PCA on such data ([21]), while others ([3]) have shown how natural hand motions lie on low dimensional non-linear manifolds. Then, a methodology similar to what is proposed in [22], might be used: First, learn the manifold using Locally Linear Embedding, or other manifold learning method, and then map back to the original dimensionality using, e.g. a kernel method. Such a method can be incorporated directly into the framework described here; replacing the PC basis with the kernel basis.

On the other hand, for many applications non-restricted hand motion is not required. E.g., applications like human computer interfaces, navigation, and games; where the simplicity of this system can be an asset. The results presented here show that this tracker is robust to difficult scenarios, self occlusion and complex global motions, and can therefore suit perfectly for such an application. In the near future our research will include expanding this approach to a more multi-pose gesture tracking for practical interfacing purposes.

References

1. Soutschek, S., Penne, J., Hornegger, J., Kornhuber, J.: 3-D Gesture-Based Scene Navigation in Medical Imaging Applications Using Time-Of-Flight Cameras. In: Proc. Conference on Computer Vision and Pattern Recognition Workshops (2008)
2. Erol, A., Bebis, G., Nicolescu, M., Boyle, R.D., Twombly, X.: Vision-based hand pose estimation: A review. Computer Vision and Image Understanding 108 (2007)
3. Stenger, B., Thayananthan, A., Torr, P.H.S., Cipolla, R.: Model-Based Hand Tracking Using a Hierarchical Bayesian Filter. Trans. on Pattern Analysis and Machine Intelligence 28(9), 1372–1384 (2006)
4. Deutscher, J., Reid, I.: Articulated body motion capture by stochastic search. International Journal of Computer Vision 61(2), 185–205 (2005)

5. Canton-Ferrer, C., Casas, J.R., Pàrdas, M.: Exploiting structural hierarchy in articulated objects towards robust motion capture. In: Perales, F.J., Fisher, R.B. (eds.) AMDO 2008. LNCS, vol. 5098, pp. 82–91. Springer, Heidelberg (2008)
6. Kato, M., Chen, Y.W., Xu, G.: Articulated hand motion tracking using ICA-based analysis and particle filtering. Journal of multimedia 1(3) (2003)
7. Kolb, A., Barth, E., Koch, R., Larsen, R.: ToF-Sensors: New Dimensions for Realism and Interactivity. Computer Graphics Forum (2009)
8. MESA Imaging AG, `http://www.mesa-imaging.ch/`
9. Zhu, Y., Dariush, B., Fujimura, K.: Controlled human pose estimation from depth image streams. In: Proc. Conference on Computer Vision and Pattern Recognition Workshops (2008)
10. Breuer, P., Eckes, C., Müller, S.: Hand gesture recognition with a novel ir time-of-flight range camera-a pilot study. In: Gagalowicz, A., Philips, W. (eds.) MIRAGE 2007. LNCS, vol. 4418, pp. 247–260. Springer, Heidelberg (2007)
11. Kollorz, E., Hornegger, J.: Gesture recognition with a time-of-flight camera. Int. J. on Intell. Systems and Techn. and App. (IJISTA), Issue on Dynamic 3D Imaging (2007)
12. Schiller, I., Beder, C., Koch, R.: Calibration of a PMD-Camera Using a Planar Calibration Pattern Together with a Multi-Camera Setup. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, pp. 297–302 (2008)
13. Guðmundsson, S.A., Larsen, R., Aanæs, H., Pardás, M., Casas, J.R.: ToF imaging in smart room environments towards improved people tracking. In: Proc. Conference on Computer Vision and Pattern Recognition Workshops (2008)
14. Bregler, C., Malik, J.: Tracking people with twists and exponential maps. In: Proc. Conference on Computer Vision and Pattern Recognition, pp. 8–15 (1998)
15. Mikic, I.: Human body model acquisition and tracking using multi-camera voxel data. PhD Thesis, University of California, San Diego (2002)
16. Kanatani, K.: Statistical optimization for geometric computation: theory and practice. Elsevier Science Ltd., Amsterdam (1996)
17. Isard, M., Blake, A.: CONDENSATION - Conditional density propagation for visual tracking. International Journal of Computer Vision 29(1), 5–28 (1998)
18. Arulampalam, M., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. IEEE Transactions on Signal Processing 50(2), 174–188 (2002)
19. van Rijsbergen, C.J.: Information Retrieval. Butterworths, London (1975)
20. MacCormick, J., Isard, M.: Partitioned sampling articulated objects and interface quality hand tracking. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 3–19. Springer, Heidelberg (2000)
21. Wu, Y., Lin, J., Huang, T.S.: Capturing natural hand articulation. In: Proc. Intl Conf. Computer Vision, ICCV, pp. 426–432 (2001)
22. Jaeggli, T., Koller-Meier, E., Gool, L.V.: Multi-activity tracking in lle body pose space. In: ICCV workshop: 2nd Workshop on Human Motion - Understanding, Modeling, Capture and Animation, pp. 42–57 (2007)