Hand and Body Association in Crowded Environments for Human-Robot Interaction

Stephen McKeague, Jindong Liu and Guang-Zhong Yang[1]

***Abstract*— For mobile robot navigation in crowded environments, hand and body tracking to enable seamless human-robot interaction is a challenging problem. Many existing methods simplify the task with static camera assumptions, initial calibration stages, or ad hoc pose constraints, making them difficult to be applied to assistive robots used for healthcare applications. This paper introduces a method of hand-body association suitable for crowded environments, by incorporating depth cameras. A robust human hand and body detector, optimized for crowded environments, is first introduced. This is followed by a probabilistic framework for associating hands and bodies. Geodesic distances, based on depth information, are employed to isolate points local to a hand, regardless of their Euclidean proximity to points in other regions. This facilitates subsequent hand-body association based on a Bayesian framework with increased association robustness. The accuracy of the proposed method is evaluated using a range of parameters against an existing approach. A public dataset has been created to assess the method's practical value in crowded environments.**

I. Introduction

Human-Robot Interaction (HRI) is the study of natural and effective communication between robots and humans. With the prevalence of chronic diseases and improved survival rate due to advances in medicine, the use of assistive robots is expected to increase significantly in the next two decades. This general trend is also driven by a demographic shift associated with the aging population, and our increasing demand on improving the quality of life of the elderly and those with chronic diseases. HRI solutions provide an important means by which people can naturally command and control robots in an environment. Many complex HRI tasks, such as gesture recognition and attention detection, necessitate a robust understanding of the motion of hands and associated bodies in a given scene.

Crowded and dynamic environments pose particular problems for many existing methods of hand and body tracking. To simplify the task, body part detectors frequently make assumptions of body pose and occlusions [1]–[3]. In real-world environments, however, these constraints cannot be assumed. Most contemporary body tracking research has thus far focused on full-body pose estimation. To reduce the search space of this complex problem, many well-known methods make use of background modeling [4], or require a user to adopt a specific pose to initialize tracking [5], [6]. Naturally, these restrictions prevent such methods from being adopted into an effective HRI framework, particularly for the purpose of patient or elderly care.

The results from the recent ChaLearn Gesture Challenge [7] produced a surprising outcome: all top ranking methods made no explicit detection and tracking of humans or individual body parts. Clearly there is a need for the development of robust methods by which this can be accomplished. This paper proposes such a method.

The purpose of this work is to present a HRI framework for tracking hands and associated bodies in crowded environments. The framework operates on depth images and offers real-time execution. It can deal with rapid tracking initialization, differing clothing and skin tone, variable illumination conditions and multiple hypothesis considerations.

The proposed framework for hand-body association consists of three major components, each representing a novelty of the proposed method. Firstly, a hand-detector designed for crowded environments is presented. Secondly, a probabilistic method of filtering bodies from background noise within a cluttered scene is described. Finally, a Bayesian estimation framework is used to associate tracked bodies and hands. Hand detection accuracy is evaluated using a range of parameters, and validated against the performance of the shape context descriptor [8]. No publicly available data set could be found to replicate the crowded environments that we wished to use in assessing our method. As a result, the manually annotated files used to generate the results have been made available online.

II. Previous Work

The robust detection of human body parts has been the subject of extensive research in computer vision. Due to the availability of reliable depth images, the performance of these methods is steadily increasing. For example, Shotton and Sharp proposed a body part detector used by Microsoft's Xbox Kinect [4]. With this technique, individual pixels are classified as one of thirty one possible body parts using a random forest classifier. Classification is evaluated using the per-pixel depth differences of a subset of pixels, defined during offline training.

Ikemura and Fujiyoshi introduced a similar depth feature [9], which expresses the similarity in depth information over two regions. It requires normalized depth histograms to be computed over eight by eight pixel squares. The "relational depth similarity feature" is defined as the Bhattacharyya distance between the histograms of these two regions.

Plagemann et. al. developed an interest point detector for body parts that operates on depth images [1]. This "AGEX" (Accumulative Geodesic EXtrema) point is formulated by dividing an input depth image into connected surface meshes

[1] S. McKeague, J. Liu and G.Z. Yang are with the Department of Computing, Imperial College London {sjm05, jliu4, gzy} at imperial.ac.uk

to be analyzed individually. Keypoints are identified as those with the furthest geodesic distance from the centroid of the mesh.

Recently, Li and Kulic [2] presented a modification to the shape context descriptor for body part identification. A hierarchical algorithm was proposed for initially locating body endpoints. The descriptor, termed the "local shape context", encodes the distance from these endpoints to the nearest detected edge, in uniformly sampled radial directions.

Human detection is a long standing problem in computer vision, having been largely tackled using color images. However, there are many challenging aspects of crowded environments that necessitate distinct solutions to the problem.

To tackle the problem of crowd identification, Chen et. al. introduced a method employing object classification techniques [10]. Objects within a scene are segmented using a combination of background subtraction and temporal differencing of pixels. Objects are identified as crowds using a combination of temporal features classified by the AdaBoost algorithm [11], the object's self-similarity response, and analysis of its spatio-temporal energies.

Hydra is a mature system for the detection and tracking of multiple people within a group [12]. Individuals within a crowd are first segmented using background subtraction, followed by corner detection and region-based shape analysis. Detected heads are tracked with a dynamic template and a second-order motion model. An appearance model is constructed for each person to recover from lost tracking due to occlusion.

Many previous hand-body association techniques have been designed for problem-specific domains, and as such differ largely in their design.

To recognize pointing gestures during HRI, Nickel and Stiefelhagen proposed a method that analyzes the orientation of the associated person [13]. Hands and heads are detected using skin color segmentation and are localized using computational stereo. Hand-head tracking is dependent on skin detection probability, anatomic likelihood and movement since the previous frame. The optimal hand-head hypothesis is calculated by maximizing this probability over a number of previous frames.

Buehler et. al [14] developed a model for identifying the hand and arm poses of the sign language translator accompanying TV broadcasts. A template matching algorithm is firstly used to estimate the shape and position of the head and torso. The optimal hand-arm configuration is then found by minimizing a cost function using a sampling-based framework. This cost function depends on the color likelihood of different body parts, the edge response likelihood, the pose change between consecutive frames and anatomic likelihood.

III. Method

Generally, hand-body association consists of three major components. Firstly, a method of hand detection must be defined. Similarly, body detection must be performed in parallel. Finally, hands and bodies must be tracked and associated together.

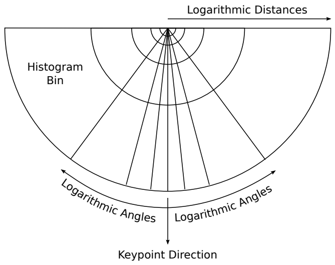

Fig. 1. Illustration of the proposed hand descriptor, based on log-polar sampling.

A. Hand Detection

The shape context is a successful image descriptor [8], [15]. The proposed descriptor, shown in Figure 1, is similarly a depth image histogram, optimized for hand detection in crowded environments.

Log-polar sampling, as used by the shape context, ensures that the highest concentration of sample points lies near the keypoint being analyzed [16]. In order to focus on local regions most indicative of a hand, the proposed descriptor extends log-polar sampling to angular, as well as distance, binning. This is achieved by assigning an angle of interest to each keypoint. Additionally, areas with the lowest concentration of points (those of opposite angle to the keypoint) can be ignored entirely.

Canny edge detection is initially performed to reduce the hand detection search space. The resulting edges are treated as potential hand keypoints in the remaining steps of the algorithm.

Using Euclidean space to analyze a keypoint's local points can lead to problems in crowded environments. As shown in Figure 2, for a keypoint on a person's fingertips, unconnected background and other people will frequently have a smaller Euclidean distance than points on the forearm. Thus, to ensure that distance from a keypoint corresponds to importance to the descriptor, distances are calculated in geodesic space.

The geodesic distance between two pixels in a depth image is the shortest cumulative distance between them, when traversing paths of neighboring pixels on the same mesh. Geodesic distances can be computed optimally using Dijkstra's algorithm [17]. Being a graph search algorithm, we can formulate the problem in terms of the input depth image as follows: Local pixels on the same mesh as the keypoint are represented by nodes in the graph. Two nodes are connected if they represent neighboring pixels.

Two parameters of this graph naturally arise. The first is a minimum Euclidean distance, above which neighboring pixels are considered belonging to different meshes, min_d. The other is a maximum geodesic distance from the keypoint at which pixels are considered local, max_d. min_d can be equated to the low hysteresis threshold of the Canny edge detector. max_d should be chosen through experimentation.

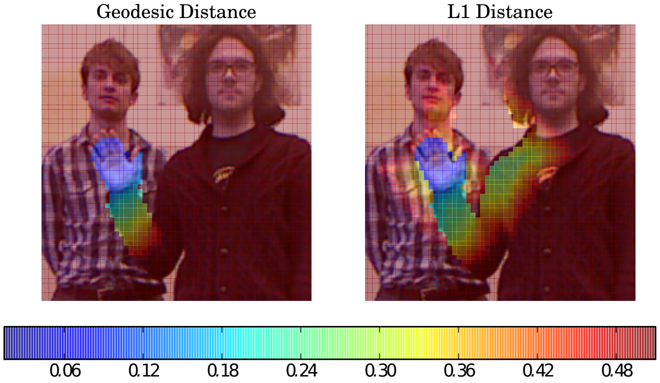

Fig. 2. Image showing the advantage of using geodesic distances to highlight important regions of a hand in crowded environments. Both geodesic and L1 distances within 0.5 m of the fingertips are shown.

Figure 2 shows how changing max_d effects the size of the local region of a hand.

Thus, for each keypoint, $\mathbf{k} = [x, y, z]^T$, in a depth image's corresponding point-cloud, $\mathcal{P}$, the geodesic distances of local points, $\mathcal{P}_k$, are calculated:

$$\mathcal{P}_k = \{\mathbf{x} | \mathbf{x} \in \mathcal{P}, g(\mathbf{x}, \mathbf{k}) < max_d\}, \tag{1}$$

where $g(\mathbf{x}, \mathbf{k})$ is the geodesic distance of point $\mathbf{x} = [x, y, z]^T$ from $\mathbf{k}$.

From these local points a 2D keypoint direction vector, $\mathbf{k}^d$, is calculated. This vector points towards the mean of the points in $\mathcal{P}_k$:

$$\mathbf{k}^d = \mathbf{P}\left(\frac{1}{N}\sum_{\mathbf{x}\in\mathcal{P}_k} \mathbf{x} - \mathbf{k}\right)^T,$$

where N is the number of points in $\mathcal{P}_k$ and $\mathbf{P} = [1, 1, 0]^T$.

For each local point, $\mathbf{x} \in \mathcal{P}_k$, a 2D direction vector from its keypoint, $\mathbf{k}$, is also calculated. The unnormalized vector is defined as:

$$\mathbf{x}^d = \mathbf{P}\left(\mathbf{x} - \mathbf{k}\right)^T.$$

As can be seen from Figure 1, points that lie further than $\frac{\pi}{2}$ from the keypoint direction are discarded. Additionally, $\hat{\mathbf{k}}^d$ lies at an angle of $\frac{\pi}{2}$ relative to the descriptor. For valid points, the angular difference, θ_{xk}, between the descriptor and $\hat{\mathbf{x}}^d$ is then calculated:

$$\theta_{xk} = \cos^{-1}\left(\hat{\mathbf{x}}^d \cdot \left(\mathbf{R}\hat{\mathbf{k}}^d\right)\right),$$

$$\mathbf{R} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Knowing both the geodesic distance, $g(\mathbf{x}, \mathbf{k})$, and angular difference, θ_{xk}, of all valid points, the keypoint's histogram can be constructed. Rather than assign appropriate bins by calculating the logarithm of these values, it is more efficient to calculate the static histogram bin boundaries.

The logarithmic distance boundary, b^d_n, of bin number n, out of a total of N^d, is defined as:

$$b^d_n = min_d\left(\frac{max_d}{min_d}\right)^{\frac{n}{N^d-1}}, \quad n = 0 \dots N^d - 1.$$

Again, max_d represents the maximum geodesic distance of a local pixel, whilst min_d represents the minimum Euclidean distance that separates different meshes.

Angular bin boundaries must be symmetric about the keypoint direction. The maximum boundary lies at $\frac{\pi}{2}$ either side of this direction. The logarithmic angular boundary, b^a_n, of bin number n out of $N^a - 1$ is then given as:

$$b^a_n = sign\left(n - \frac{N^a}{2} + 1\right)\frac{\pi}{2}^{\left|\frac{2n+2}{N^a}-1\right|} + \frac{\pi}{2}, \quad n = 0 \dots N^a - 1.$$

Note that only even values of N^a are considered.

Knowing the histogram bin thresholds, a local point, $\mathbf{x} \in \mathcal{P}_k$, is assigned a distance, x^i, and angular, x^j, bin number as follows:

$$x^i = \min_n\left(g(\mathbf{x}, \mathbf{k}) \leq b^d_n\right), \quad n = 0 \dots N^d - 1,$$

$$x^j = \min_n\left(\theta_{xk} \leq b^a_n\right), \quad n = 0 \dots N^a - 1.$$

The combined histogram descriptor is formed by concatenating all angular bin values in an anticlockwise direction, in order of increasing distance bin values. To ensure that the descriptor is invariant to the number of pixels used in its construction, each bin value, b_n, should be normalized to one. This ensures that the descriptor is scale-invariant.

$$b'_n = \frac{b_n}{\sum_{k=0}^{N^d N^a - 1} b_k}, \quad n = 0 \dots N^d N^a - 1.$$

Hand detection is performed using a Support Vector Machine (SVM), pre-trained on descriptors corresponding to hands. In addition to the standard training of positive and negative samples, a technique from [18] is adopted. An initial classifier is trained on a subset of the total positive and negative samples. This classifier is then used to detect hands in video sequences where none are present. The resulting false positives are considered "hard examples", and are added to the training set. The classifier is then retrained, resulting in improved performance.

B. Body Detection

1) Initial Clustering: The first step of the body detection algorithm is to segment the scene into spatially separated clusters. A necessary pre-requisite is the removal of the floor plane, which connects all subjects in the input image. With the depth camera mounted on a mobile robot, offline calibration of the floor plane normal, $\hat{\mathbf{n}}$, and a point in the floor plane, $\mathbf{x}_f$, can be performed.

All points, $\mathbf{x}$, in the input point-cloud, $\mathcal{P}$ are filtered if their Euclidean distance to the floor plane is within a small threshold, ϵ:

$$\mathcal{P}' = \{\mathbf{x} | \mathbf{x} \in \mathcal{P}, |(\mathbf{x} - \mathbf{x}_f) \cdot \hat{\mathbf{n}}| > \epsilon\}.$$

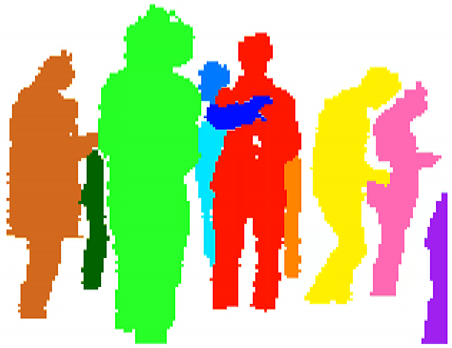

Fig. 3. Initial clustering output. Clusters are colored differently. Note the planar arm that has been clustered separately from the associated body.

To segment the filtered point-cloud, $\mathcal{P}'$, into spatially separated clusters, a connected components algorithm is used [19]. The results of this processing step can be seen in Figure 3.

Note that for many human gestures, such as the one pictured, planar arms can lie over 0.3 m from the associated body, whilst foreshortened hands can be over 0.5 m. This will cause hands and arms to be clustered separately, unless the minimum clustering distance is set to a reliably large value. However, in the crowded environments that the proposed HRI framework is designed for, this will lead to incorrect clustering of adjacent bodies. It is this fact that necessitates the introduction of the following Bayesian association algorithm, and selection of an appropriately small minimum clustering distance.

2) Body Detection: It can be seen from Figure 3 that upper bodies have a smaller width and height variance than random background clusters. Principal component analysis of a cluster provides a sound method of extracting this shape information. A probabilistic filtering method can then be applied to the results in order to detect potential bodies.

Firstly, N points local to the top of a body, $\mathcal{P}_b$, are isolated using equation 1. The covariance matrix, $\mathbf{C}$, of these points, $\mathbf{x} = [x, y, z]^T$, is defined as:

$$\mathbf{C} = \frac{1}{N}\sum_{k=0}^{N}(\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^T,$$

where $\bar{\mathbf{x}}$ is the mean of the data points.

The cluster's first principal component, denoted as $\mathbf{v}_0$, will give a posture invariant vector running from head to toe. Conversely, the cluster's second principal component, $\mathbf{v}_1$, will run horizontally along the upper body:

$$\mathbf{v}^T\mathbf{C}\mathbf{v} = \lambda.$$

The eigenvalue corresponding to a given principal component gives a measure of the variance of points along its direction. The first eigenvalue, λ_0, thus gives a measure of upper body length. Conversely, the second eigenvalue, λ_1, gives a measure of width.

f_0 shall be used to denote the cluster feature governed by λ_0, whilst f_1 shall denote the cluster feature governed by λ_1. The probability that these features indicate the presence of a body, b, can be naturally modeled by Gaussian distributions:

$$P(f_0 \mid b) \sim \mathcal{N}(f_0 \mid \mu_0,\, \sigma_0^2),$$

$$P(f_1 \mid b) \sim \mathcal{N}(f_1 \mid \mu_1,\, \sigma_1^2).$$

Because of the orthogonality property of principal components, f_0 and f_1 are independent variables:

$$\mathbf{v0} \cdot \mathbf{v1} = 0.$$

The Mahalanobis distance of a cluster's feature vector, $d(f_0, f_1)$, can be used as a measure of similarity to the average body, defined during offline parameter fitting. The square of this distance will be chi-square distributed, with two degrees of freedom. Selecting the 0.95 quantile of this distribution allows us to reasonably filter all feature vectors with a larger d^2 value as not characterizing bodies;

$$d(f_0, f_1) = \sqrt{\sum_{k=0}^{1}\frac{(f_k - \mu_k)}{\sigma_k^2}}.$$

Finally, a noise-invariant reference point, $\mathbf{x}^r$, is defined for every detected body. It is calculated from the mean of the points in $\mathcal{P}_b$:

$$\mathbf{x}^r = \frac{1}{N}\sum_{\mathbf{x}\in\mathcal{P}_b}\mathbf{x}.$$

C. Hand-Body Association

Temporal tracking of detected hands and bodies is achieved using multiple independent Kalman filters. The task remains, however, of associating one with the other. Recursive Bayesian estimation can be used to provide a solution to this problem.

Naturally, if a hand has been detected in a cluster which is also a known body, a high probability can be assigned to their association. However, separate clustering of hands and bodies occurs when the pixels connecting the arm and body are completely occluded. This can only occur with a select number of poses, such as in Figure 3. A natural likelihood for hand-body association in these situations can be formulated from analysis of the displacement between hands and bodies.

Still ensuring generality, a separately clustered hand, h, will be at one of G different positions. As such, association likelihood can be most accurately represented using a Gaussian mixture model. The probability of h, being at position g, is dependent on the displacement, $\mathbf{d} = \{x, y, z\}$, from its associated body:

$$P(g|\mathbf{d}) = \pi_g \mathcal{N}\left(\mathbf{d} \mid \mu_g, \mathbf{\Sigma}_g^2\right).$$

The likelihood of h being associated with a body, b_i, is thus given by the weighted sum of the individual probabilities:

$$P(h \mid b_i) = \sum_{g\in G} P\left(g|\mathbf{d}_i\right),$$

where $\mathbf{d}_i$ is the distance between h and b_i.

The posterior probability for hand-body association can then be given by:

$$P(b_i \mid h) = \frac{P(h \mid b_i)P(b_i)}{\sum_{k=0}^{I} P(h \mid b_k)P(b_k)}, \tag{2}$$

where I is the number of detected bodies.

For selecting an appropriate prior probability, $P(b_i)$, in equation 2, it would be pertinent to incorporate temporal dynamics into the proposed system. As hand and body tracking has been achieved using Kalman filtering, it is logical to also model hand-body associations as a Markov process. Thus, the probability of b_i being associated with h at time t, given the associated probability at time $t-1$, is conditionally independent of all prior associations:

$$P(b_i^t \mid b_i^{t-1}, b_i^{t-2}, \ldots, b_i^0) = P(b_i^t \mid b_i^{t-1}).$$

Additionally, the likelihood of h being associated with b_i depends only on the current association, and is conditionally independent of all prior associations:

$$P(h^t \mid b_i^t, b_i^{t-1}, \ldots, b_i^0) = P(h^t \mid b_i^t).$$

Equation 2 can then be rewritten, incorporating the new temporal prior:

$$P(b_i^t \mid h^t) = \frac{P(h^t \mid b_i^t)P(b_i^t \mid h^{t-1})}{\sum_{k=0}^{I^t} P(h^t \mid b_k^t)P(b_k^t \mid h^{t-1})}, \tag{3}$$

where I^t is the number of detected bodies at time t.

The posterior probability for each hand-body association at time t is set to the prior probability at time $t+1$. Thus, the optimal associated body for a hand at any given time is equal to the *maximum a posteriori* solution of equation 3.

IV. Results

A. Hand Detection

1) Discussion: Microsoft's Kinect was used as the input depth camera for all experiments, downsampled to a resolution of 160 by 120 pixels. max_d, the maximum geodesic distance of pixels local to a keypoint, was set to 0.5 m. This maximizes the descriptiveness of the forearm whilst reducing the variability of the bend at the elbow. Run-time performance of the hand detector averages at just over 15 frames per second, with SVM classification alone accounting for over 70% of execution time.

To evaluate hand detection parameter choices, a series of relatively noiseless video sequences were recorded, some containing people and others of random background. A total of 1400 hands were manually labeled from the appropriate videos. Due to the high cardinality of possible background samples, 11500 were selected randomly from background videos. The combined samples were split into training and test sets with a 70:30 ratio. This resulted in a test set of 420 positive and 3450 negative samples.

The SVM classification accuracy of the proposed descriptor is shown in Figure 4. Results are shown for a range of angle and distance bins. As can be seen, the descriptor exhibits good performance, even for low numbers of distance and angle bins. Increasing the number of angle bins has a large effect on accuracy. Increasing the number of distance bins, however, does not produce such monotonic behavior. This indicates the importance that the angular binning scheme has on the descriptor's performance.

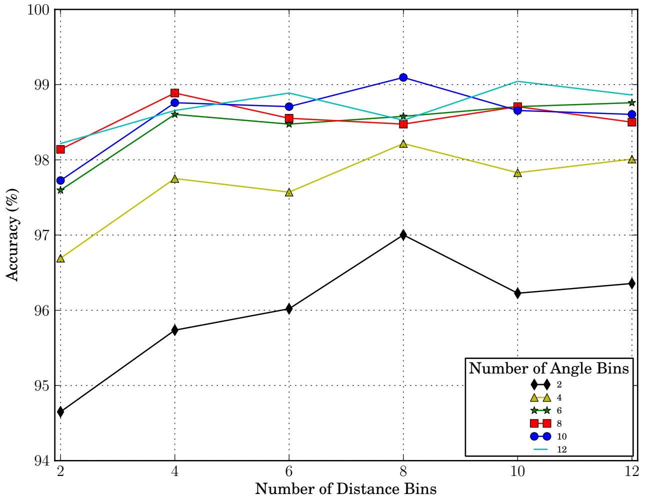

Fig. 4. Graph showing the classification accuracy of the proposed descriptor. Accuracy clearly increases with the number of angle bins used.

TABLE I

Confusion matrix of the proposed descriptor. Columns represent the predicted class. Rows represent the actual class.

	Hand	Background
Hand	405	15
Background	16	3433

Optimal numbers of distance and angle parameters bins were chosen in order to further analyze the maximum performance of the proposed descriptor. Each entry of the confusion matrix, shown in Table I, details the corresponding classifier's responses to the hand and background samples from the test set.

2) Validation: To more fully validate its performance, the sensitivity of the proposed descriptor was compared against that of the original shape context, using a depth image as input. Sensitivity is defined as:

$$\text{sensitivity} = \frac{TP}{TP+FN},$$

where TP denotes the number of true positives, and FN denotes the number of false negatives. The results, displayed in Figure 5, show a clear improvement of the proposed descriptor for all combinations of distance and angle bins.

The optimal distance and angle bin combination was chosen for the shape context, and used to construct the results shown in Table II. The accuracy, sensitivity and specificity of both descriptors is detailed, where specificity is defined as:

$$\text{specificity} = \frac{TN}{TN+FP}.$$

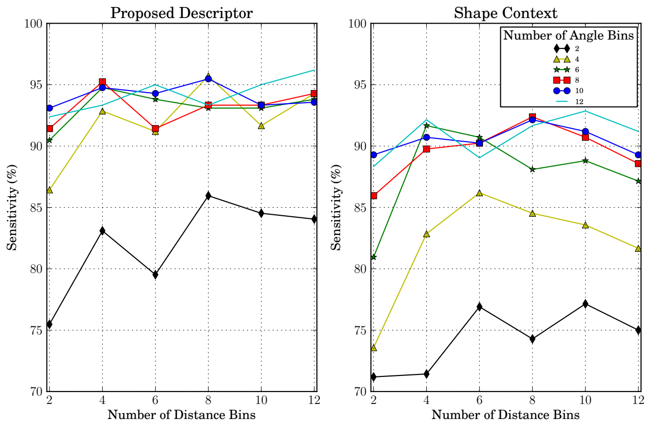

Fig. 5. Graph comparing the sensitivity of the proposed descriptor and the shape context. The proposed descriptor outperforms the shape context for every bin combination.

TABLE II

TABLE SHOWING THE IMPROVED PERFORMANCE OF THE PROPOSED DESCRIPTOR OVER THE SHAPE CONTEXT.

	Accuracy	Sensitivity	Specificity
Proposed Descriptor	99.2%	96.4%	99.5%
Shape Context	98.5%	91.4%	99.3%

TN denotes the number of true negatives and FP denotes the number of false positives.

B. Hand-Body Association

In testing hand-body association, four gestures were defined: a wave to grab the robots attention, a subtle push to have the robot leave, a subtle follow me motion, and a raised hand indicating that the robot should stop. Multiple people, from a range of backgrounds, were used to perform these gestures. Images of the gestures can be seen in Figure 7.

To test the temporal behavior of the hand-body association algorithm, an experiment was constructed in which two people were placed side by side. One of the subjects was asked to move their hand in a pushing gesture. The tracked hand was then moved in front of the other subject, so as the hand location would be more naturally associated with the wrong person. This process was repeated for a period of twenty-five seconds, with the tracked hand alternating in front of the two subjects. The effects on per-frame likelihood, and posterior probability were recorded. The results can be seen in Figure 6.

A higher likelihood represents a higher hand-body association probability in the current time instant. In Figure 6, time periods where the hand is in front of the wrong subject are thus obvious. The posterior probability incorporates association probabilities from previous time instants. As can be seen, this gives a more stable association during transient periods of incorrectly lower likelihood. When this time period is too long, the posterior probability will decrease appropriately. This behavior can be seen at the twenty second mark, and allows the framework to recover from incorrect associations.

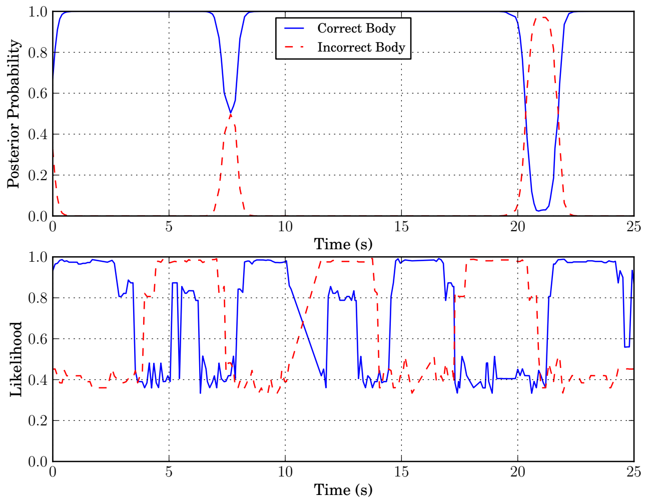

Fig. 6. Plot showing the effect of differing likelihood on posterior probability. The posterior probability provides stable association results during transient periods of incorrectly lower likelihood.

In order to evaluate the complete hand-body association framework, three different crowded scenarios, were recorded. Each scenario takes place in a different location with different lighting conditions. As can be seen from the images shown in Figure 7, these scenarios represent a challenging, real-world environment for HRI. Each scenario is over a minute long, and consists of various people within the crowd gesturing towards the robot. People and gestures were manually annotated to facilitate automatic results generation. The datasets have been made publicly available[1]. Figure 7 shows hand-body association results from each of the three scenes. Tracked bodies are displayed in red, tracked hands are displayed in blue, and associations are denoted with a white line.

Hand-body association accuracy for the scenarios is shown in Table III. These results were obtained by backprojecting the estimated 3D hand and body positions, and comparing them with ground truth, image-space, annotations made on a 640 by 480 pixel image. Ignoring a detection's z-component, a hand is considered correctly detected if it lies within 0.1 m of the ground truth, whilst a body is considered correctly detected if it is within 0.3 m. Using the well-known tri-phase gesture model, only the stroke phase [20] (the unique component) of a gesture was analyzed during results generation. The association accuracy of the shape context was generated using the presented hand-body association algorithm, with hands detected using the shape context rather than the proposed hand detector. To minimize the effect of crowded background noise on the shape context, only edge points within 0.3 m of a keypoint were analyzed.

With a minimum association accuracy of 74.9% in such a challenging environment, these results validate the success of the proposed algorithm. This is further reinforced by

[1] `http://www.imperial.ac.uk/hamlyn/eo/gesturedataset`

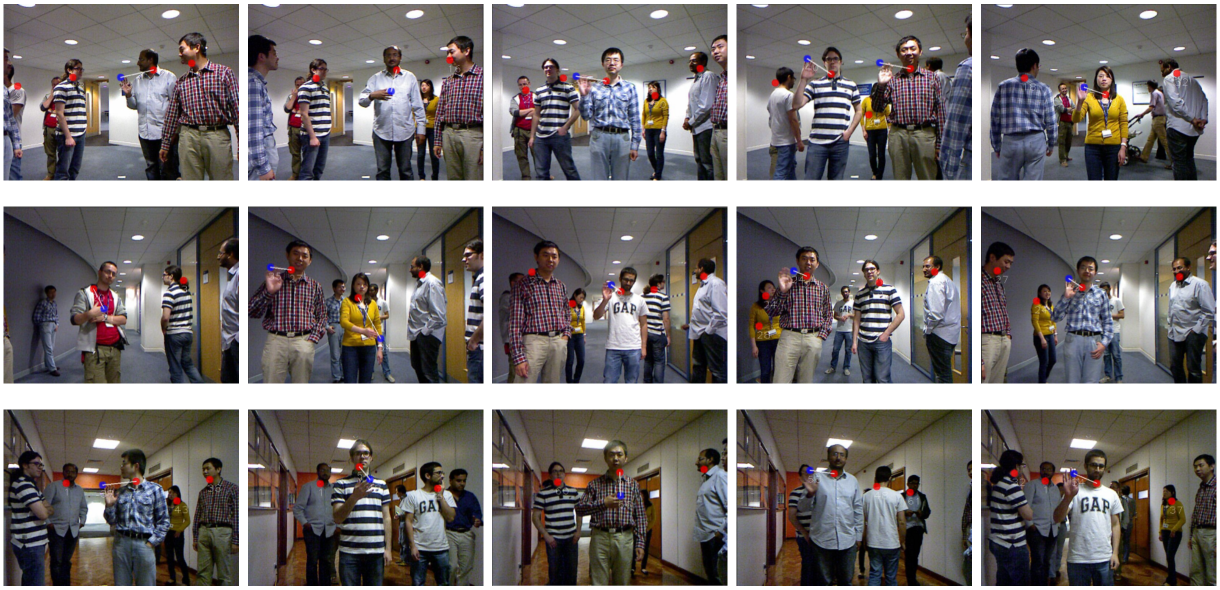

Fig. 7. Example results showing the accuracy of the proposed hand-body association method in three different crowded scenes. Each scene has different lighting conditions, and a range of gesturing people from various backgrounds. Tracked bodies are displayed in red. Tracked hands are displayed in blue. Associations are shown with a connected white line.

TABLE III

TABLE SHOWING HAND-BODY ASSOCIATION RESULTS FOR THE THREE SCENARIOS EVALUATED.

	1	**2**	**3**
Av. Gesture Length (s)	2.3	2.0	2.1
Body Detection Accuracy (%)	88.4	86.8	92.5
PROPOSED DESCRIPTOR			
Hand Detection Accuracy (%)	78.4	76.4	83.0
Association Accuracy (%)	86.4	74.9	90.9
False Positive Associations per Frame	0.11	0.13	0.05
SHAPE CONTEXT			
Hand Detection Accuracy (%)	39.7	40.0	35.7
Association Accuracy (%)	76.2	46.8	56.3
False Positive Associations Per Frame	0.42	0.32	0.23

the 90.9% association accuracy in the third scenario. Hand detection performance using the shape context is notably worse. This can be explained due to the varied gestures used in the crowded test environments that were not present during classifier training. The shape context does not generalize well to samples that it has not been trained on. However, the use of geodesic distances in the proposed descriptor allows it to perform well outside the noiseless environments that it was trained on, and generalize to new hand postures. Additionally, the number of false positives produced by the proposed descriptor is almost four times less on average than the shape context.

Association accuracy for the shape context is much higher than its hand detection accuracy. This can be explained as a result of the subtleties of the gestures. If a Kalman filter is instantiated at the start of a subtle gesture, it can lie within 0.1 m of the ground truth for much of its duration, even without further detections to update its position. Although decreasing the ground truth threshold below 0.1 m can alleviate this effect, this results in incorrectly reduced detection accuracy.

Despite these results, a source of inaccuracy in the proposed method can be identified, being that the hand detector is susceptible to self-occlusions. When a foreshortened hand is presented that the detector was not trained on, detection frequently fails. Shorter gestures are usually less pronounced, and thus exhibit this behavior more often. Decreasing max_d can alleviate this problem, at the expense of increased false positives.

V. CONCLUSIONS

In this paper, we have introduced a framework for hand and body association in crowded environments, for human-robot interaction. Three main novelties were presented. A hand detector, optimized for crowded environments, was described. Detailed results were presented and its performance was validated against the shape context descriptor. A method of body detection was presented, along with a probabilistic algorithm for associating the results with detected hands. Quantitative analysis of the framework was performed in a number of crowded environments, where the robustness and generality of the method was again validated against the shape context. Additionally, the datasets for this work have been made publicly available.

The most obvious application of this work is to form part of a gesture recognition system. Gesture recognition is a vital component of HRI and as such requires a robust underlying hand-body association framework. Human attention detection is another natural application of this work. Without such a system, a robot that detects and associates gestures simultaneously from multiple people will have no way of prioritizing detected commands.

References

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

[2] Z. Li and D. Kulic, "Local shape context based real-time endpoint body part detection and identification from depth images," *Computer and Robot Vision, Canadian Conference*, vol. 0, pp. 219–226, 2011.

[3] G. Mori and J. Malik, "Estimating human body configurations using shape context matching," in *Proceedings of the 7th European Conference on Computer Vision-Part III*, ser. ECCV '02. London, UK, UK: Springer-Verlag, 2002, pp. 666–680. [Online]. Available: http://dl.acm.org/citation.cfm?id=645317.649329

[4] J. Shotton and T. Sharp, "Real-time human pose recognition in parts from single depth images," *IEEE Conference on Computer Vision and Pattern Recognition (2008)*, vol. 2, no. 3, pp. 1297–1304, 2011. [Online]. Available: http://www.stat.osu.edu/~dmsl/BodyPartRecognition.pdf

[5] Y. Zhu and K. Fujimura, "Constrained optimization for human pose estimation from depth sequences," in *Proceedings of the 8th Asian conference on Computer vision - Volume Part I*, ser. ACCV'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 408–418. [Online]. Available: http://portal.acm.org/citation.cfm?id=1775614.1775663

[6] D. Grest, J. Woetzel, and R. Koch, "Nonlinear body pose estimation from depth images," in *DAGM-Symposium'05*, 2005, pp. 285–292.

[7] I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hamner, and H. Escalante, "Chalearn gesture challenge: Design and first results," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, june 2012, pp. 1 –6.

[8] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 509–522, 2001.

[9] S. Ikemura and H. Fujiyoshi, "Real-time human detection using relational depth similarity features," in *Proceedings of the 10th Asian conference on Computer vision - Volume Part IV*, ser. ACCV'10. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 25–38. [Online]. Available: http://dl.acm.org/citation.cfm?id=1966111.1966115

[10] D.-Y. Chen, K. Cannons, H.-R. Tyan, S.-W. Shih, and H.-Y. Liao, "Spatiotemporal motion analysis for the detection and classification of moving targets," *Multimedia, IEEE Transactions on*, vol. 10, no. 8, pp. 1578 –1591, dec. 2008.

[11] Y. Freund and R. Schapire, "A short introduction to boosting," *Japonese Society for Artificial Intelligence*, vol. 14, no. 5, pp. 771–780, 1999.

[12] I. Haritaoglu, D. Harwood, and L. Davis, "Hydra: multiple people detection and tracking using silhouettes," in *Image Analysis and Processing, 1999. Proceedings. International Conference on*, 1999, pp. 280 –285.

[13] K. Nickel and R. Stiefelhagen, "Visual recognition of pointing gestures for human-robot interaction," *Image Vision Comput.*, vol. 25, pp. 1875–1884, December 2007. [Online]. Available: http://dl.acm.org/citation.cfm?id=1287837.1287944

[14] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman, "Upper body detection and tracking in extended signing sequences," *International Journal of Computer Vision*, vol. 95, no. 2, pp. 180–197, 2011. [Online]. Available: http://www.springerlink.com/index/10.1007/s11263-011-0480-9

[15] E. Kalogerakis, A. Hertzmann, and K. Singh, "Learning 3d mesh segmentation and labeling," in *ACM SIGGRAPH 2010 papers*, ser. SIGGRAPH '10. New York, NY, USA: ACM, 2010, pp. 102:1–102:12. [Online]. Available: http://doi.acm.org/10.1145/1833349.1778839

[16] D. Young, U. of Sussex. School of Cognitive, and C. Sciences, *Straight lines and circles in the log-polar image*, ser. Cognitive science research papers. School of Cognitive and Computing Sciences, University of Sussex, 2000. [Online]. Available: http://books.google.co.uk/books?id=daV_GwAACAAJ

[17] M. C. et. al, "Priority queues and dijkstras algorithm," The University of Texas at Austin, Department of Computer Sciences, Tech. Rep., 2007.

[18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, june 2005, pp. 886 –893 vol. 1.

[19] R. Jain, R. Kasturi, and B. G. Schunck, *Machine vision*. New York, NY, USA: McGraw-Hill, Inc., 1995.

[20] S. Mitra and T. Acharya, "Gesture recognition: A survey," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 37, no. 3, pp. 311 –324, may 2007.