Learn to Detect and Recognize Humans using Small Data Sets

Shichao Ou
Laboratory for Perceptual Robotics
Computer Science Department
University of Masschusetts Amherst
Amherst, Massachusetts
chao@cs.umass.edu

Rachel Lee
Computer Science Department
Swarthmore College, PA
rlee1@swarthmore.edu

Rod Grupen
Laboratory for Perceptual Robotics
Computer Science Department
University of Masschusetts Amherst
Amherst, Massachusetts
grupen@cs.umass.edu

ABSTRACT

Personal robotics is an area in which robot behavior is in service to few (or single) clients. This paper argues that the problems of human detection and recognition can be approached with simple yet efficient techniques that provide useful information to personal robots. By combining and taking advantage of coarse information such as motion, activities, shape, and color attributes, simple probabilistic inference algorithms can be applied to help a robot to become aware of nearby humans and their identities. Experimental results show that these simple models can be used to detect human presence robustly against a naturally clutterd and ambiguous background and perform well in a recognition test consisting of 10 subjects. Since this approach does not rely on the faces as crucial cue for detection or recognition, it can function under situations where conventional techniques would fail. Moreover, the simple model offers dramatic improvement in computation efficiency and can be used for robots to engage real-time interaction with human.

Keywords

human detection, human recognition, human centric robotics, learning

1. INTRODUCTION

Robotics research in recent years is beginning to shift towards “human-centric”—in constrast to static, highly constrained environments such as factory floors, are more dynamic domains in homes or offices where the robot is required to assist and collaborate with humans. Although a much more challenging goal, the interaction with humans also offers many opportunities to simplify problems: a human teacher/collaborator can provide on-line guidance and structure environments such that learning can be simplified interactively. This enables the robot to learn more complicated tasks that are previously difficult when human is out of the loop.

In previous work, Hart et al. have presented a hierarchical learning framework and shown that a bimaual robot can learn a complex policy for picking up objects as the trainer presents an increasingly challenging sequence of pick-up tasks [7, 6]. Through human-guided structured learning, the robot quickly learns a basic sequence of the pick-up behavior and then later adapts to cases of randomly placed objects, different scale objects, and even moving objects. However, in these experiments, guidance from the human came in an offline manner. The goal of this project is to extend the behavioral learning framework with elements that enables the our robot to become aware of nearby humans, to identify them, and subsequently, to react to their movements and instructions, or request human assistance. This interaction leads to more effective learning on more complicated tasks.

With this goal in mind, the approach of this paper differs from the conventional human detection and recognition problems in computer vision literature. This work considers:

- in-door environments—they are generally well-lit, and relatively static compared to out-door scenes. Handling extreme lighting conditions and distraction of moving background elements such as cars, bikes, birds or tree leaves are not the target of this study.
- much smaller datasets—in a normal household, a robot would only need to differentiate between two or three human subjects. Even in business environments it is reasonable to think a robot may only need to personalize its interaction behavior with no more than 20 people. The requirement for the recognition system on such a robot is significantly simpler than for instance a security face recognition system at an airport checkpoint.
- whole-body activity recognition and human detection in real-time—in cases where no facial features, it is still possible for robots to distinguish human collaborators from passers by and children from parents by

their distinctive appearance and behavior. Moreover, for the robot to maintain natural interaction with a human user, real-time performance is a necessary requirement. However, most state-of-the-art human detection and tracking algorithms are limited to offline processing [21, 9, 19, 14, 20].

Under these conditions, this paper focuses on how the combination of simple, multi-modal features, and activities can help robot detect and recognize humans. Section 2 reviews related work and contrasts the differences between them and the approach taken by this work. Section 3.1 presents the vision architecture for simultaneous feature extraction. Simple probabilistic human methods for both detection and recognition are presented in Section 3. Section 4 provides proof-of-concept experiments to demonstrate the feasibility of this approach in lab environments where the robot learns to differentiate 10 subjects.

2. RELATED WORK

Finding humans, tracking human motions and ultimately identifying them in natural settings are the holy-grail of computer vision. Many have worked on different aspects of these problems for decades. Literature on this subject is too exhaustively enumerate. In this section, a few examples are described to illustrate the current prevalent approaches for human detection and recognition:

Human detection is a difficult problem because humans are dynamic in appearance and motion. Occlusion, variations in pose, clothing, and articulated motion all contribute to the challenge. Currently, the most effective approaches for human detection and tracking are part-based methods, where the human is modeled as an assemblage of parts with kinematic relationships between features can be modeled. Earlier work in this line of research uses 3D kinematic models [8, 5, 13]. However, for these methods, stereo correspondence is an issue and also 3D models have many parameters and degrees of freedom that introduce computational complexity.

As a simpler alternative, there have been approaches where the human body is modeled as a tree of 2D parts [21, 9, 19] where a generative probabilistic model of humans is learned using labeled training data. Inference (using Nonparametric Belief Propagation) is performed on the graph structure for the detection of humans and estimation of human poses. For implementation simplicity, some researchers do not rely on a complex generative graphic model approach. Instead they define a number of constraints using prior knowledge about the human body and then either use brute force search [14] or dynamic programming to solve the assignment problem [20].

For human recognition, the majority of the effort has been put on face recognition as a face is a distinctive part of the human body. As described in several recent surveys [22, 10], face recognition methods can be categorized into *holistic* methods (Eigenfaces [11], Fisherfaces [1] and LDA [3] etc..), *feature-based* methods (e.g. pure geometry[2], HMM [15]) and *hybrid* methods (e.g. modular eigenfaces [18] and hybrid LFA [17]). However, these methods are engineered specifically for the task of face recognition and do not lend insights to handling cases when faces are not visible and how to rely on other cues for human recognition.

The common issue with all of the above mentioned methods is that the design goal is to have them function in the most general settings, e.g. a single image, and under arbitrary lighting conditions. As a result, only the most reliable features are used, such as edges, corners, or texture features. Texture features are computationally intensive to extract. Simple features such as edges however are ubiquitous in the environment and often large number of such features are detected (Figure 2). With a large N, the computational complexity has restricted these methods to off-line video processing. Our approach is hierarchical: first simple but coarse features are used to reduce N, and allow the system to quickly focus on candidate regions. Then, if the coarse features are not sufficient to make an distinction, then texture features can be computed on the candidate regions instead of the entire image to improve efficiency.

Secondly, due to the standard evaluation method, e.g. for face recognition algorithms, classification using a single image selected the database, effective elements that humans often take advantage of, such as motion, activity, posture or even clothing habits, are not taken into account in these algorithms.

3. PROPOSED APPROACH

The following are the key points of the proposed approach:

1. A set of simple features is used to both simplify the feature extraction process and reduce the number of features (N) extracted in each frame. Motion is used extensively in early stages of learning.

2. A simplified version of the constellation of feature approach by Fergus et al. [4] is proposed in this work for human detection.

3. The detection algorithm functions as a pre-processing step for the recognition algorithm to reduce the number of outlier features used for learning and classification.

4. A simple belief network is proposed for human recognition. Multi-sensor features extracted from previous steps, including feature blob position, scale, and color, and activity pattern (likely-to-visit locations encoded in Cartesian coordinates) are taken into account for subject classification.

Details of each of these steps are described in the following sections.

3.1 Feature Extraction Architecture

As shown in Figure 1, the robot perceives the world through a broad range of features extracted from visual, proprioceptive, and force signals.

Each channel of sensory signal feedback is passed through a signal processing pipeline (Figure 1 left) where raw sensory input is filtered using a feature mask, e.g. hue values within a certain range. For visual inputs, connected components are used to segment contiguous regions that share a feature. A Kalman filter is used to provide optimal, least squares position ($\widetilde{u}$) and scale (Σ_u) estimates of the feature as well as its first order dynamics ($f'(u,t)$) in the presence of noise. Thus, a summary of discrete events and the first order dynamics of each type of feature in space and time is delivered as a perceptual basis for the subsequent object/human modeling and behavioral learning.

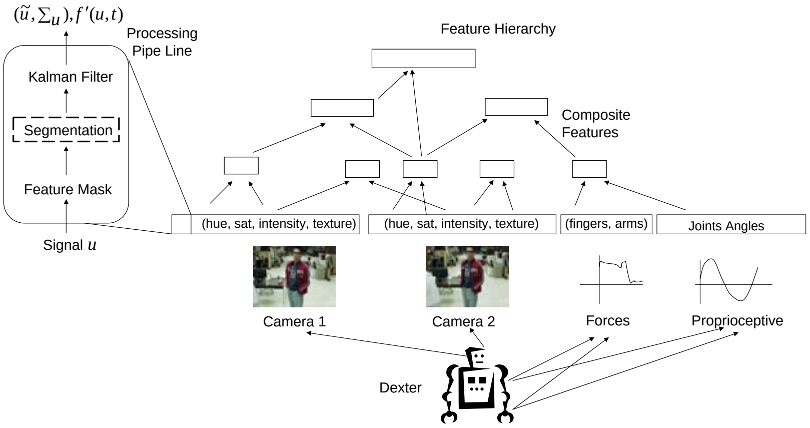

Figure 1: The sensory input processing pipeline: the robot perceives the world through a broad range of features, in visual, proprioceptive and force signals

For this work, only the visual channels are used. The hue, saturation and intensity (HSI) color space is discretized into 18 channels of hue (value ranges from 0 $\sim$ 180), 10 channels of saturation and 10 channels of intensity. An example output of these channels through the sensory processing pipeline is shown in Figure 2. These features are coarse and independently produce an ambiguous summary of the scene. However, in combination, we demonstrate that they effectively discriminate between separable individuals in a small data set. When the simple color, motion, and structural features are not sufficient to make the distinction, then texture features can be computed in the ambiguous regions to improve discriminative power. For this paper, only the coarse features are studied as a first step.

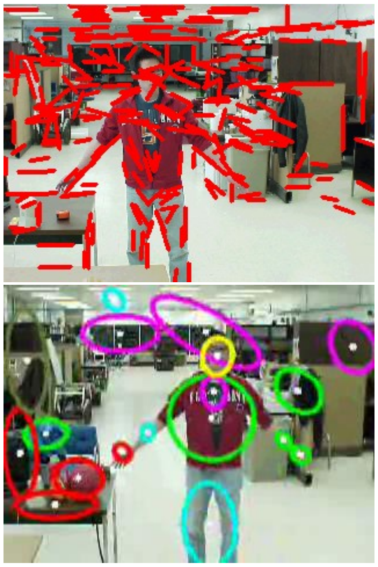

Figure 2: Feature comparison. (a) extracted edge features (b) extracted color features proposed in this work.

3.2 A Constellation of Features Model for Human Detection

Primitive features from the sensory processing pipeline can be combined to form constellations of features and aggregated into abstract features. As the abstraction process continues, feature hierarchy is formed such that complex objects can be modeled. For instance, to model a human, the lowest level of abstract features are modeled as rigid body parts of the human (e.g. lower arm, upper arm), while the next level up is the kinematic multi-body segments of the human (e.g. arm), and finally these kinematic segments form the highest level "human" abstract feature. To aggregate lower level features into an abstract feature, this work employs a probabilistic constellation of features approach proposed by Fergus et al. [4].

In Fergus's approach, objects are modeled as P parts, and for each part, shape (X), appearance (A) and relative scale (S) models are learned. For a given object model with parameters Θ, to determine object presence/absence, a Bayesian decision R is made, s.t.

$$\begin{aligned} R &= \frac{Pr(Object|X,S,A)}{Pr(NoObject|X,S,A)} \\ &\approx \frac{Pr(X,S,A|\theta)Pr(Object)}{Pr(X,S,A|\theta_{bg})Pr(Noobject)} \end{aligned}$$

where Θ_{bg} is the background model The likelihoods are factored into appearance (A), shape (X) and relative scale (S) components, and assuming variables A, X and S are statistically independent from each other:

$$\begin{aligned} & Pr(X,S,A|\Theta) \\ = & \sum_{h\in H} Pr(X,S,A,h|\Theta) \\ = & \sum_{h\in H} Pr(A|X,S,h,\Theta)Pr(X|S,h,\Theta)Pr(S|h,\Theta)Pr(h\Theta) \\ = & \sum_{h\in H} Pr(A|h,\Theta)Pr(X|h,\Theta)Pr(S|h,\Theta)Pr(h|\Theta) \end{aligned}$$

where each of the factored components is modeled as Gaussian distributions, and h is a hypothesis vector (length P)

regarding whether a detected feature belong to a certain part. Given the observed N (maximum) features, each entry in h is between 0 and N that locates a feature to a model part. The set H is all valid combinations of features to the parts, and therefore $|H|$ is $O(N^P)$. For large N, the process of evaluating all possible combinations of H to compute probability $Pr(X, S, A|\Theta)$ is expensive.

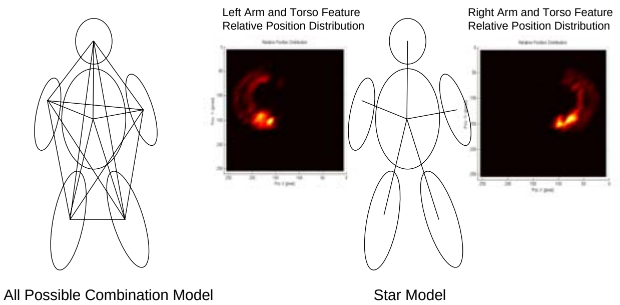
Left Arm and Torso Feature
Relative Position Distribution
Right Arm and Torso Feature
Relative Position Distribution
All Possible Combination Model
Star Model

Figure 3: Using the star model (right), in which the position and scale distribution of each feature is encoded with respect to a reference feature. For instance, relative position distributions between an arm feature and the torso feature is illustrated in the figure. This distribution is a curve because the arms often move up and down during training. Such distribution is not Gaussians. Therefore, distributions in this work are represented using discretized accumulator arrays. With this approach, the number of hypothesis H can be de reduced to $O(N^2P)$ instead $O(N^P)$ in the case of the all pairs model (left) used in the original Fergus paper.

For this work, several simplifications are made to achieve a real-time implementation. First, the use of simple features such as color blobs that reduces N. To further reduce computational complexity, instead of considering all possible combinations (Figure 3), a star model is used where a reference feature is selected (the most stable invariant feature) such that only hypotheses with respect to the reference feature are considered. This is similar to Fergus's later work where his analysis shows that the computation complexity of this approach is reduced to $O(N^2P)$. Lastly, instead of learning Gaussian mixture models, which is computationally expensive, a discretized nonparametric modeling approach is employed (Figure 3).

3.3 A Bayesian Belief Network Approach for Recognition

After a human has been identified from the scene, the corresponding set of blobs are found. Assuming the presence of one human at a time, the robot with a stereo pair of cameras can triangulate and compute the human's position (x, y, z) in Cartesian coordinates. Figures 4 shows the graphical model of the proposed belief network for human recognition. The leaf nodes in the network consist of the observations: i.e. (x, y, z) Cartesian position of the human, relative scale of the observed blobs. Note that more blob properties such as relative distance between blobs and ID of the color features can also be added as leaf nodes to increase the discriminative power of the network. However, for simplicity, only scale is included in this version of the network.

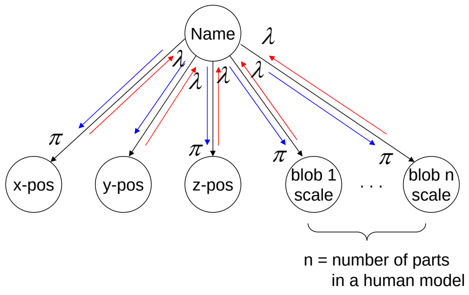

Figure 4: Bayesian network for human recognition. The root node of the network is the class labels of each subject, and the leaf nodes correspond to observation features include (x, y, z) Cartesian position of the human, and scale values of the observed blobs

The Cartesian position nodes encode activity information, in terms of frequently visited locations by a particular human. The relative scale and position of the observed blobs encodes specific shape information of a human, i.e., some subjects are taller than others, while some are larger. Finally, blob's color IDs can also encode clothing habit information of a subject human. These features are often ignored by conventional human recognition techniques as they are not discriminative enough when the database consists of thousands of individuals. However, for the purpose of the domain proposed by this work, they are particularly useful for coarse classification of a small set of humans whose faces may not be visible, but they have distinctive clothing preferences in color distributions.

The network is trained using a standard belief propagation technique as described in Pearl's book [16]. First, a conditional probability table, that stores the probabilities of each value of the leaf nodes (Y) of the network given the value of its parent (X), is computed. Evidence enters the network through each leaf node's λ vector. A root node, X, receives this information in a λ message transmitted by its children, Y, where

$$\lambda_Y(x) = \sum_y \lambda(y)P(y|x), where, x \in X.$$

The root node calculates the product of its π messages, whose components correspond to the prior probabilities of each x, and every λ message from its children. The resulting vector is the network's belief, which corresponds to the "believability" of an event x occurring. This belief network crudely attempts to take into account past evidence by using the belief calculated at previous time steps to weight the belief vector caulated for the current time step.

4. RESULTS

4.1 Detection

For training, humans walk in front of the robot, sometimes wave arms up and down to allow the robot perceive their kinematic range. For this process, a single camera is needed, and features are extracted through the signal processing pipeline. Motion is used to allow the robot focus on blobs on the human body as the background objects do not move during training.

After training, the learned human model is tested in a naturally cluttered scene, without relying on background subtraction. In this case, the robot is exposed to features from the background as well as features on the human, the noisy data introduces a great deal of ambiguity and there can be many blobs that are potential torso features (Figure 5(a)).

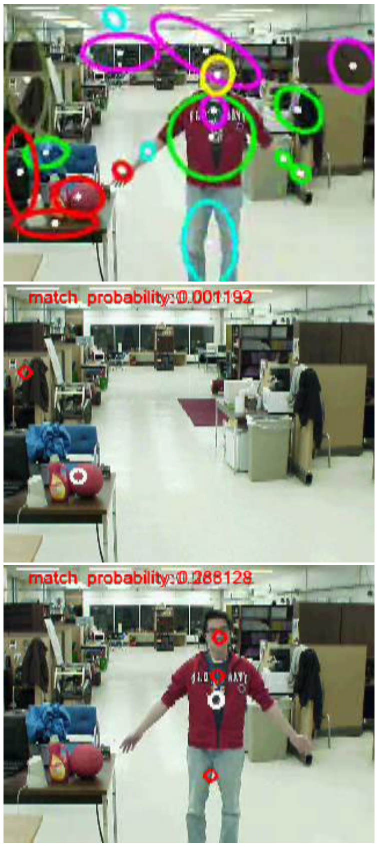

Figure 5: Human detection test. (a) Features extracted in a natural scene with cluttered background, (b) current most probable location (white circle) of a human, though probability is low (0.01192), (c) with presence of a human, a proper estimate is made with a higher confidence(0.288129).

As shown in Figure 5(b), the object on the table contains the relevant set of features that makes it a possible candidate as a human (indicated by the white circle). However, since the features' spatial arrangements do not match the kinematic structure of a human, the match probability is low. When the human subject walks in, the estimate quickly responds to the human. The red circles are features that are found to match the part-based kinematic human model. Each feature found adds support to the overall human model match probabilistically. As a result, the match probability is much higher (Figure 5(c)).

Due to the use of simple approach for modeling and inference, this human detection module runs in real-time at frame-rate (15 fps), and is used as a pre-processing step for the recognition component, to discard outlier features in the background and resolve correspondence issue during triangulation.

4.2 Recognition

Experiments of recognition are carried out in a 10-by-5 meter space in front of the robot. To simulate humans daily activities in a lab environment in this small and confined space, 10 subjects of convenience are drafted from the department, and they are asked to walk in one of three patterns over a span of 2 meters in front of the robot. These patterns, include straight back and forth between two points, figure-eight and zigzag, are used to simulate idiosyncrasies of human behavior on a small scale. Although the system can run in real-time, for qualitative analysis, a data set of 10 subjects walking in different patterns is collected. 10% of the collected data is used for offline cross validation while the rest is used for training.

Figure 6 shows the learned conditional probability distributions of position and scale features for 10 subjects. The less overlap between the distributions, the more distinguishable the subjects are. However inference does not rely on any particular feature alone. Results show that in this particular dataset, x position does not contribute much in distinguishing between the subjects as there is a significant amount of overlap in the feature distributions. However, y positions and scale features are distinctive in many situations. Even for subjects that are not distinctive in one feature, distinction can be found through other features. For instance, subject 10 is not distinguishable in both the x-position and scale in one dimension, but is much more so in the other dimension of scale and in the y-position.

Table 1 shows the confusion matrix from the evaluation results. The columns correspond to the identification ground truth, and the rows correspond to prediction made by the network. High value in the diagonals of the matrix indicates a correct match between the name label and observed features. Results show that the network performs well on the dataset collected such that 9 out 10 subjects are correctly identified. We can see from Figure 6 that the reason why subject 9 is mis-classified is because its conditional probability distribution is dwarfed by other subjects in every feature. This is to be expected due to the limitation of the resolution of the features we used in the data collection process. With higher resolution, or more features, classifications result can be improved.

5. DISCUSSION AND CONCLUSION

This paper argues that the problems of human detection and recognition, for the purpose of robotics can be approached with simple yet efficient techniques. This comes from the observation that in-home or office assistant robots may only need to personalize its behavior to a handful of people. By combining and taking advantage of coarse information such as motion, activities, shape and color attributes, simple probabilistic inference algorithms can be applied to help a robot to become aware of nearby humans and their identities. Experimental results show that these

Table 1: Confusion matrix of belief distribution for 10 subjects

SubjectID	1	2	3	4	5	6	7	8	9	10
1	495.903	0.000	0.000	0.000	0.000	0.000	0.000	0.000	16.840	0.000
2	0.000	500.000	0.297	0.233	0.089	0.000	0.000	0.000	0.001	0.006
3	0.061	0.000	34.178	0.014	0.054	0.000	0.000	0.000	0.004	0.000
4	0.001	0.000	0.000	499.685	0.000	0.001	0.000	0.000	0.001	0.000
5	0.003	0.000	465.283	0.000	499.736	0.000	0.000	0.000	0.001	0.000
6	0.000	0.000	0.000	0.068	0.000	499.998	0.000	0.000	0.000	0.000
7	3.891	0.000	0.177	0.000	0.098	0.000	499.998	0.000	0.000	0.000
8	0.030	0.000	0.021	0.001	0.018	0.000	0.000	500.000	0.051	0.002
9	0.111	0.000	0.044	0.000	0.005	0.000	0.001	0.000	483.070	0.000
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.029	499.992

simple human models can be used to detect human presence robustly against a naturally clutterd and ambiguous background and perform well in a recognition test consisting of 10 subjects. Since this approach does not rely on the faces for detection or recognition, it can function under situations where conventional techniques would fail. Moreover, the simple model offers dramatic improvement in computation efficiency and can be used for robots to engage real-time interaction with human.

For future work, several problems deserve attention: (1) adding textures [12] to the robot's feature set for extraction as discussed in Section 3.1. Compared to previous methods, we believe real-time performance can be achieved when these texture features are computed within candidate regions selected by the detection component proposed in this paper, instead of over the entire image; (2) handling multiple humans during recognition: currently, it is assumed that only 1 human is present in the environment at any given time during the recognition phase, due to the correspondence issue when triangulating the human's position. We anticipate this issue can be resolved by performing a match procedure on the candidate sets of human blobs such that the closely matched ones from different cameras are paired up for triangulation; (3) show examples of utilizing interaction to bootstrap learning of more complicated tasks; (4) use interaction to expand the robot's knowledge regarding the humans beyond the visual percept, and learn the behavioral uniqueness of humans in order to better distinguish humans from the environment (detection) and from each other (recognition).

6. ACKNOWLEDGMENTS

This material is based upon work supported under Grant ONR N00140710749.

7. REFERENCES

[1] P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible illumination conditions? *International Journal of Computer Vision*, 28(3):245–260, 1998.

[2] I. J. Cox, J. Ghosn, and P. N. Yianilos. Feature-based face recognition using mixture-distance. In *CVPR*, pages 209–. IEEE Computer Society, 1996.

[3] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human face images (invited paper). In J. Bign, G. Chollet, and G. Borgefors, editors, *AVBPA*, volume 1206 of *Lecture Notes in Computer Science*, pages 127–142. Springer, 1997.

[4] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 264–271, June 2003.

[5] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, Jan. 1999.

[6] Hart, S. S., Sen, and R. Grupen. Generalization and transfer in robot control. In *Epigenetic Robotics Annual Conference*, 2008.

[7] Hart, S. S., Sen, and R. Grupen. Intrinsically motivated hierarchical manipulation. In *Proceedings of 2008 IEEE Conference on Robots and Automation (ICRA)*, 2008.

[8] D. C. Hogg. Model-based vision: A program to see a walking person. *Image and Vision Computing*, 1(1):5–20, Feb. 1983.

[9] S. Ioffe and D. Forsyth. Human tracking with mixtures of trees. pages 690–695.

[10] R. Jafri and H. R. Arabnia. A survey of component-based face recognition approaches. In H. R. Arabnia, M. Q. Yang, and J. Y. Yang, editors, *IC-AI*, pages 103–113. CSREA Press, 2007.

[11] M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(1):103–108, 1990.

[12] J. J. Koenderink. The structure of images. *Biological Cybernetics*, 1984.

[13] M. W. Lee and I. Cohen. Proposal maps driven MCMC for estimating human body pose in static images. In *CVPR (2)*, pages 334–341, 2004.

[14] G. Mori, X. Ren, A. A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *CVPR (2)*, pages 326–333, 2004.

[15] A. V. Nefian and M. H. H. III. Face detection and recognition using hidden markov models. In *ICIP (1)*, pages 141–145, 1998.

[16] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* 1992.

[17] P. Penev and J. Atick. Local feature analysis: A general statistical theory for object representation. *Computational Neural Systems*, 7:477–500, 1996.

[18] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1994.

[19] D. Ramanan and D. Forsyth. Finding and tracking

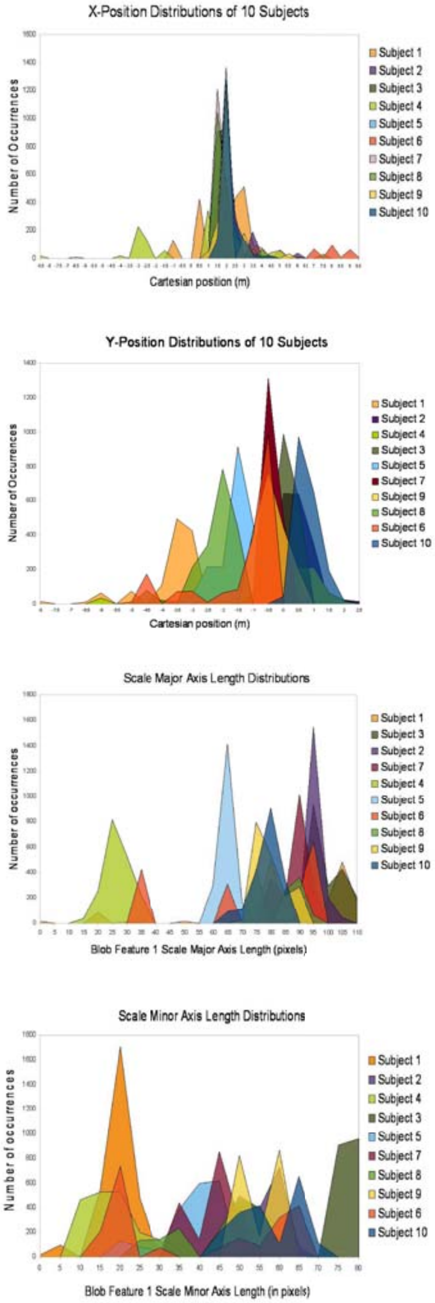

Figure 6: Conditional probability distributions of different features collected from 10 subjects. The less overlap in the distributions, more distinguishable the subjects are. The taller the peak corresponds to a stronger association of a name label to a particular feature value.

people from the bottom up, 2003.

[20] X. F. Ren, A. C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *International Conference on Computer Vision*, pages I: 824–831, 2005.

[21] L. Sigal and M. J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *IEEE Computer Vision and Pattern Recognition or CVPR*, pages II: 2041–2048, 2006.

[22] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, pages 399 – 458, 2003. Also appeared as UMD Technical Report, CS-TR4167, 2000. Revised 2002, CS-TR4167R.