2008 IEEE/RSJ International Conference on Intelligent Robots and Systems
Acropolis Convention Center
Nice, France, Sept, 22-26, 2008

Visual Recognition of Grasps for Human-to-Robot Mapping

Hedvig Kjellström Javier Romero Danica Kragić
Computational Vision and Active Perception Lab
Centre for Autonomous Systems
School of Computer Science and Communication
KTH, SE-100 44 Stockholm, Sweden
hedvig,jrgn,dani@kth.se

***Abstract*— This paper presents a vision based method for grasp classification. It is developed as part of a Programming by Demonstration (PbD) system for which recognition of objects and pick-and-place actions represent basic building blocks for task learning. In contrary to earlier approaches, no articulated 3D reconstruction of the hand over time is taking place. The indata consists of a single image of the human hand. A 2D representation of the hand shape, based on gradient orientation histograms, is extracted from the image. The hand shape is then classified as one of six grasps by finding similar hand shapes in a large database of grasp images. The database search is performed using Locality Sensitive Hashing (LSH), an approximate k-nearest neighbor approach. The nearest neighbors also give an estimated hand orientation with respect to the camera. The six human grasps are mapped to three Barret hand grasps. Depending on the type of robot grasp, a precomputed grasp strategy is selected. The strategy is further parameterized by the orientation of the hand relative to the object. To evaluate the potential for the method to be part of a robust vision system, experiments were performed, comparing classification results to a baseline of human classification performance. The experiments showed the LSH recognition performance to be comparable to human performance.**

I. Introduction

Programming service robots for new tasks puts significant requirements on the programming interface and the user. It has been argued that the Programming by Demonstration (PbD) paradigm offers a great opportunity to unexperienced users for integrating complex tasks in the robotic system [1]. The aim of a PbD system is to use natural ways of human-robot interaction where the robots can be programmed for new tasks by simply observing human performing the task. However, representing, detecting and understanding human activities has been proven difficult and has been investigated closely during the past several years in the field of robotics [2], [3], [4], [5], [6], [7], [8].

In our work, we have been studying different types of object manipulation tasks where grasp recognition represents one of the major building blocks of the system [1]. Grasp recognition was performed using magnetic trackers [7], together with data gloves the most common way of obtaining the measurements in the robotics field. Although magnetic trackers and datagloves deliver exact values of hand joints, it is desirable from a usability point of view that the user demonstrates tasks to the robot as naturally as possible; the use of gloves or other types of sensors may prevent a natural grasp. This motivates the use of systems with visual input.

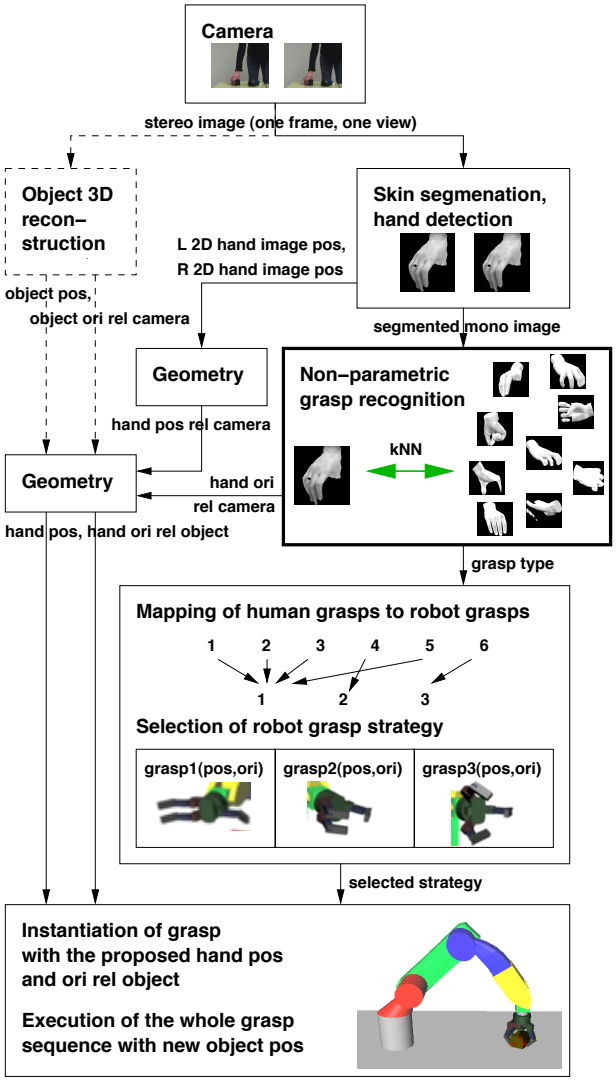
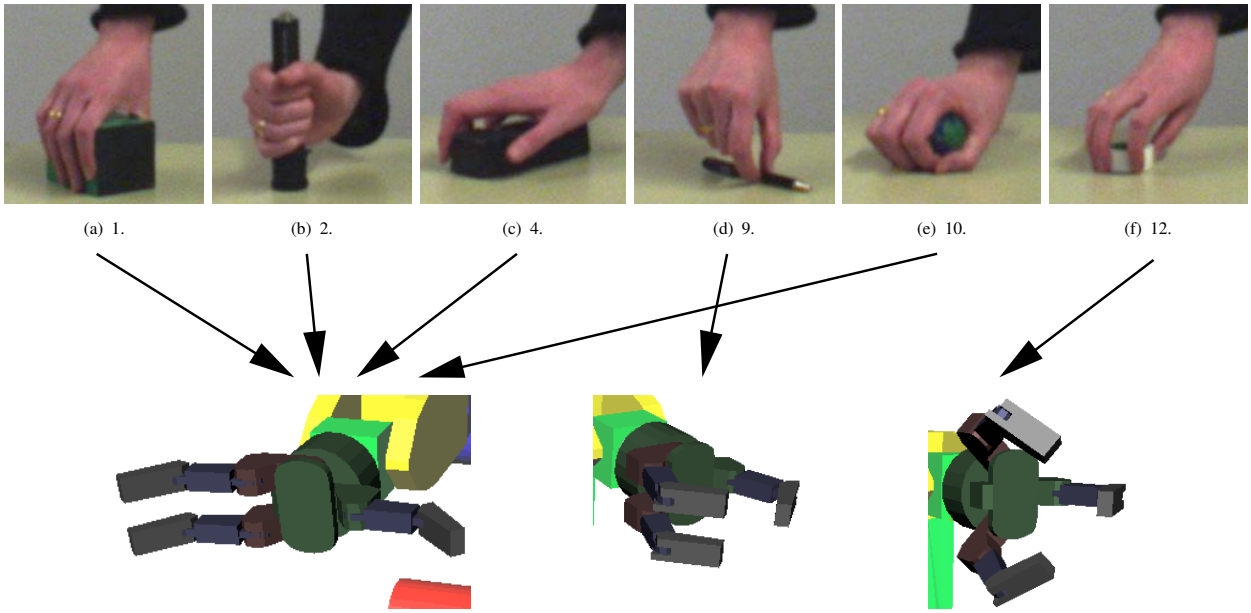


Fig. 1. Human grasps are recognized and mapped to a robot. From *one time frame* of video, the hand is localized and segmented from the background. The hand orientation relative to the camera, and type of grasp is recognized by nearest neighbor comparison of the hand view with a database, consisting of synthesized views of all grasp types from different orientations. The human grasp class is mapped to a corresponding robot grasp, and a predefined grasp strategy, *the whole approach-grasp-retreat sequence*, for that grasp is selected. The strategy is parameterized with the orientation and position of the hand relative to the object, obtained from the hand and object positions and orientations relative to the camera. (In our experiments, the object position and orientation were obtained by hand.)



(a) 1. (b) 2. (c) 4. (d) 9. (e) 10. (f) 12.

(g) Barret Wrap. (h) Barret Two-finger Thumb. (i) Barret Precision Disc.

Fig. 2. The six grasps (numbered according to Cutkosky's grasp taxonomy [9]) considered in the classification, and the three grasps for a Barret hand, with human-robot class mappings ((a,b,c,e)→(g), (d)→(h), (f)→(i)) shown. a) Large Diameter grasp, 1. b) Small Diameter grasp, 2. c) Abducted Thumb grasp, 4. d) Pinch grasp, 9. e) Power Sphere grasp, 10. f) Precision Disc grasp, 12. g) Barret Wrap. h) Barret Two-finger Thumb, i) Barret Precision Disc.

Vision based recognition of a grasping hand is a difficult problem, due to the self occlusion of the fingers as well as the occlusion of the hand by the grasped object [10], [11], [12], [13]. To simplify the problem, some approaches use optical markers [14], but markers make the system less usable when service robot applications are considered. We therefore strive to develop a markerless grasp recognition approach.

Figure 1 outlines the whole mapping procedure. Although the scientific focus of this paper is on the classification on human grasps, the classification method should be thought of as part of the whole mapping procedure, which consists of three main parts: The human grasp classification, the extraction of hand position relative to the grasped object (with object detection not implemented for our experiments), and the compilation of a robot grasp strategy, parameterized by the type of grasp and relative hand-object orientation and position, described in Section VI.

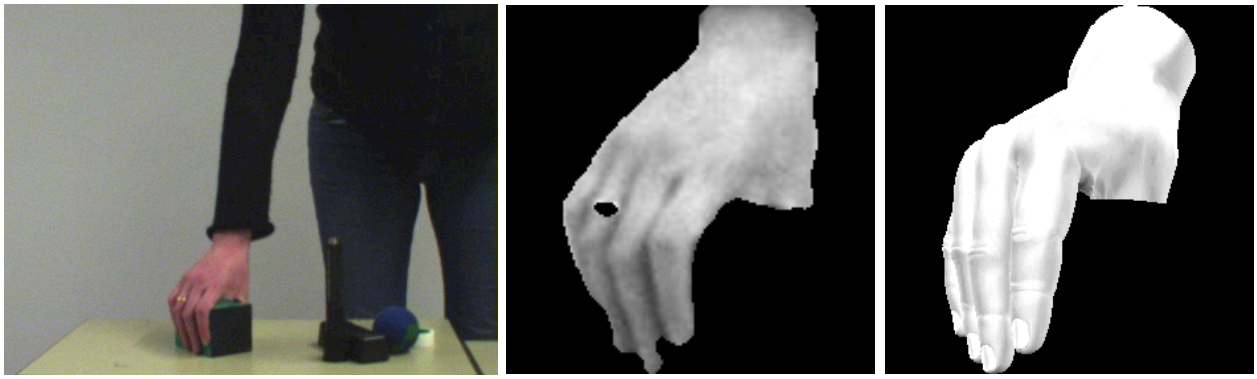
The main contribution of this paper is a non-parametric method for grasp recognition. While articulate 3D reconstruction of the hand is straightforward when using magnetic data or markers, 3D reconstruction of an unmarked hand from images is an extremely difficult problem due to the large occlusion [10], [11], [12], [13], actually more difficult than the grasp recognition problem itself as discussed in Section II. Our method can classify grasps and find their orientation, from a single image, from any viewpoint, without building an explicit representation of the hand, similarly to [12], [15]. Other grasp recognition methods (Section II) consider only a single viewpoint or employ an invasive sensing device such as datagloves, optical markers for motion capture, or magnetic sensors.

The general idea to recognize the human grasp and select a precomputed grasping strategy is a secondary contribution of the paper, since it differs from the traditional way to go about the mapping problem [7]; to recover the whole 3D pose of the human hand, track it through the grasp, and then map the motion to the robot arm. A recognition-based approach such as ours avoid the difficult 3D reconstruction problem, and is also much more computationally efficient since it only requires processing of a single video frame.

The grasp recognition problem is here formalized as the problem of classifying a hand shape as one of six grasp classes, labeled according to Cutkosky's grasp taxonomy [9]. The classes are, as shown in Figure 2a-f, Large Diameter grasp, Small Diameter grasp, Abducted Thumb grasp, Pinch grasp, Power Sphere grasp and Precision Disc grasp.

The input to the grasp classification method is a single image (one time instance, one camera view point) from the robot's camera. The hand is segmented out using skin color segmentation, presented in more detail in Section III. From the segmented image, a representation of the 2D hand shape based on gradient orientation histograms is computed as presented in Section IV. A large set of synthetic hand views from many different viewpoints, performing all six types of grasps has been generated. Details are given in Section III. The new hand shape is classified as one of the six shapes by approximate k-nearest neighbor comparison using Locality Sensitive Hashing (LSH) [16]. Along with the grasp class, the estimated orientation of the hand relative to the camera is obtained by interpolating between the orientations of the found nearest neighbors. This is presented in Section V.

Experiments presented in Section VII show the method to



(a) Input image $\mathbf{I}$. (b) Hand image $\mathbf{H}$. (c) Synthetic hand image $\mathbf{H}^{\mathrm{synth}}$.

Fig. 3. Processing of image data. a) Input $\mathbf{I}$ from the robot, grabbed with an AVT Marlin F-080C camera. b) Segmented hand image $\mathbf{H}$. c) Synthetic view of hand $\mathbf{H}^{\mathrm{synth}}$, generated in Poser 7.

perform comparably to humans, which indicates that it is fit to be included into complex vision system, such as the one required in a PbD framework.

II. Related Work

Classification of hand pose is most often used for gesture recognition, e.g. sign language recognition [12], [17]. These applications are often characterized by low or no occlusion of the hands from other objects, and a well defined and visually disparate set of hand poses; in the sign language case they are designed to be easily separable to simplify fast communication. Our problem of grasp recognition differs from this application in two ways. Firstly, the grasped object is usually occluding large parts of the grasping hand. We address this by including expected occlusion in our dataset; occluding objects are present in all example views (Section III). Secondly, the different grasping poses are in some cases very similar, and there is also a large intra-class variation, which makes the classification problem more difficult.

Related approaches to grasp recognition [14], [18] first reconstruct the hand in 3D, from infrared images [18] or from an optical motion capture system which gives 3D marker positions [14]. Features from the 3D pose are then used for classification. The work of Ogawara et al. [18] views the grasp recognition problem as a problem of shape reconstruction. This makes their results hard to compare to ours. In addition, they also use a wide baseline stereo system with infrared cameras, which makes their approach difficult to adopt in a case of a humanoid platform.

The more recent work of Chang et al. [14] learns a non-redundant representation of pose from all 3D marker positions – a subset of features – using linear regression and supervised selection combined. In contrast, we use a completely non-parametric approach where the classification problem is transformed into a problem of fast LSH nearest neighbor search (Section IV). While a linear approach is sufficient in the 3D marker space of Chang et al. [14] , the classes in the orientation histogram space are less Gaussian shaped and more intertwined, which necessitates a non-linear or non-parametric classifier as ours.

Using 3D motion capture data as input, Chang et al. [14] reached an astonishing recognition rate of up to 91.5%. For the future application of teaching of service robots it is however not realistic to expect that the teacher will be able or willing to wear markers to provide the suitable input for the recognition system. 3D reconstructions, although with lower accuracy, can also be achieved from unmarked video [19], [20]. However, Chang et al. [14] note that the full 3D reconstruction is not needed to recognize grasp type. Grasp recognition from images is thus an easier problem than 3D hand pose reconstruction from images, since fewer parameters need to be extracted from the input. We conclude that the full 3D reconstruction is an unnecessary (and error prone) step in the chain from video input to grasp type.

Our previous work [7] considered an HMM framework for recognition of grasping sequences using magnetic trackers. Here, we are interested in evaluating a method that can perform grasp classification based on a single image only, but it should be noted that the method can easily be extended for use in a temporal framework.

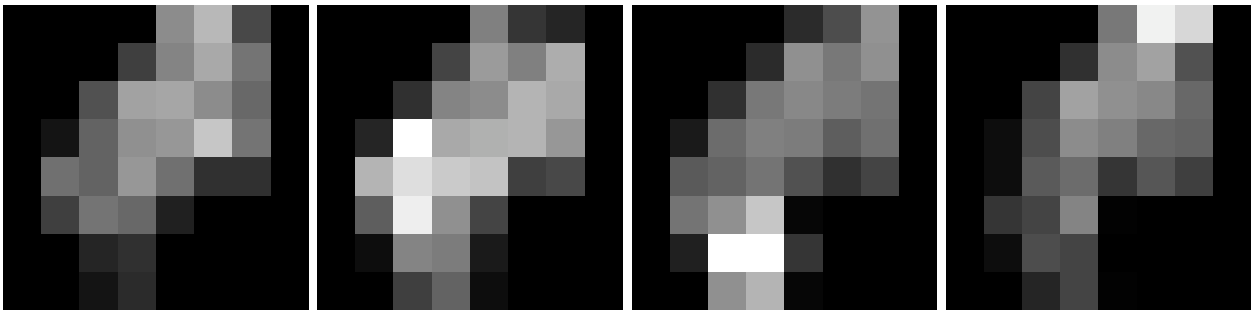
III. Extracting the Hand Image

Since the robot grasp strategies are predefined, and only parameterized by the hand orientation, position and type of grasp, there is no need for the human to show the whole grasp procedure; only one time instance is enough (for example, the image that is grabbed when the human tells the robot ”now I am grasping”).

The input to the recognition method is thus a single monocular image $\mathbf{I}$ from the a camera mounted on the robot. For our experiments, we use an AVT Marlin F-080C camera. An example of an input image is shown in Figure 3a. Before fed into the recognition, the image is preprocessed in that the grasping hand is segmented from the background.

A. Segmentation of hand images

The hand segmentation could be done using a number of modalities such as depth (estimated using stereo or an active sensor) or color. We choose to use skin color segmentation;



(a) $\frac{\pi}{8}$. (b) $\frac{3\pi}{8}$. (c) $\frac{5\pi}{8}$. (d) $\frac{7\pi}{8}$.

Fig. 4. Gradient orientation histograms from the hand image $\mathbf{H}$ of Figure 3b, with $B = 4$ bins, on level $l = 1$ in the pyramid of $L = 4$ levels (spatial resolution 8×8). a) Bin 1, orientation $\frac{\pi}{8}$. b) Bin 2, orientation $\frac{3\pi}{8}$. c) Bin 3, orientation $\frac{5\pi}{8}$. d) Bin 4, orientation $\frac{7\pi}{8}$.

the details of the method used are described in [21]. To remove segmentation noise at the borders between background and foreground, the segmentation mask is median filtered three times with a 3×3 window.

The segmented image $\hat{\mathbf{H}}$ is cropped around the hand and converted from RGB to grayscale. An example of the resulting hand image $\mathbf{H}$ is shown in Figure 3b.

B. Generation of synthetic hand images for the classification

The fact that the classification method (Section V) is non-parametric and that no explicit model of the hand is built (Section IV) means that a very large set of examples, from many different views, is needed for each grasp.

As it is virtually intractable to generate such training sets using real images, we use a commercial software, Poser 7, to generate synthetic views $\mathbf{H}^{\text{synth}}$ of different hand configurations. Poser 7 supplies a realistic 3D hand model which can be configured by bending the finger joints. For our purposes, the model was configured by hand into the 6 iconic grasps, which were a little exaggerated to provide clear distinctions between the classes. 900 views of each configuration were generated, with viewing angles covering a half-sphere in steps of 6 degrees in camera elevation and azimuth; these are the views which can be expected by a robot with cameras above human waist-height. The synthetic hand was grasping an object, whose shape was selected to be typical of that grasp [9]. The object was black (as the background), and occluded parts of the hand as it would in the corresponding real view of that grasp. This will make the synthetic views as similar as possible to the real views (e.g. Figure 3b), complete with expected occlusion for that view and grasp. Figure 3c shows such a database example.

The synthetic images $\mathbf{H}^{\text{synth}}$ can be seen as ideal versions of the segmented and filtered real hand images $\mathbf{H}$. Note that the recognition method is tested (Section VII) using real hand images prepared as described in the previous subsection, and that the synthetic images are used only for the database. Note further that the hand in the database is not the same as the hand in the test images.

IV. Image Representation

For classification of grasps, we seek a representation of hand views (Figures 3b and 3c) with as low intra-class variance, and as high inter-class variance as possible. We choose gradient orientation histograms, frequently used for representation of human shape [22], [23].

Gradient orientation $\mathbf{\Phi} \in [0, \pi)$ is computed from the segmented hand image $\mathbf{H}$ as

$$\mathbf{\Phi} = \arctan(\frac{\partial \mathbf{H}}{\partial y} / \frac{\partial \mathbf{H}}{\partial x}) \tag{1}$$

where x denotes downward (vertical) direction and y rightward (horizontal) direction in the image.

From $\mathbf{\Phi}$, a pyramid with L levels of histograms with different spatial resolutions are created; on each level l, the gradient orientation image is divided into $2^{L-l} \times 2^{L-l}$ equal partitions. A histogram with B bins are computed from each partition. An example of histograms at the lowest levels of the pyramid can be seen in Figure 4.

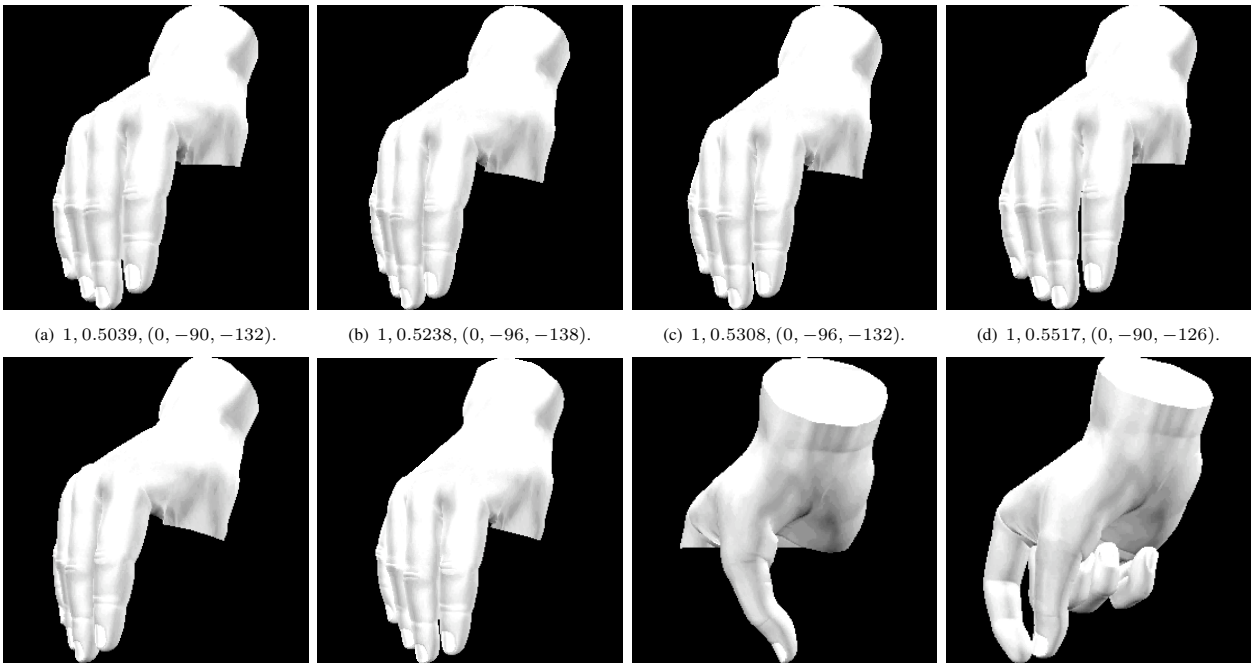
The hand view is represented by $\mathbf{x}$ which is the concatenation of all histograms at all levels in the pyramid. The length of $\mathbf{x}$ is thus $B \sum_{l=1}^{L} 2^{2(L-l)}$. The performance of the classifier is quite insensitive to choices of $B \in [3, 8]$ and $L \in [2, 5]$; in our experiments in Section VII we use $B = 4$ and $L = 3$.

V. Approximate Nearest Neighbor Classification

A database of grasp examples is created by synthesizing $N = 900$ views $\mathbf{H}_{i,j}^{\text{synth}}$ with $i \in [1, M]$, $j \in [1, N]$, from each of the $M = 6$ grasp classes (Section III), and generating gradient orientation histograms $\mathbf{x}_{i,j}$ from the synthetic views (Section IV). Each sample has associated with it a class label $y_{i,j} = i$ and a hand-vs-camera orientation $\mathbf{o}_{i,j} = [\phi_j, \theta_j, \psi_j]$, i.e. the Euler angles from the camera coordinate system to a hand-centered coordinate system.

To find the grasp class $\hat{y}$ and orientation $\hat{\mathbf{o}}$ of an unknown grasp view $\mathbf{x}$ acquired by the robot, a distance-weighted k-nearest neighbor (kNN) classification/regression procedure is used. First, X_k, the set of k nearest neighbors to $\mathbf{x}$ in terms of Euclidean distance $d_{i,j} = \|\mathbf{x} - \mathbf{x}_{i,j}\|$ are retrieved.

As an exact kNN search would put serious limitations on the size of the database, an approximate kNN search method,



(a) $1, 0.5039, (0, -90, -132)$. (b) $1, 0.5238, (0, -96, -138)$. (c) $1, 0.5308, (0, -96, -132)$. (d) $1, 0.5517, (0, -90, -126)$.

(e) $1, 0.5523, (0, -96, -144)$. (f) $1, 0.5584, (0, -102, -132)$. (g) $1, 0.5870, (0, -90, +108)$. (h) $4, 0.6068, (0, -90, +120)$.

Fig. 5. Distance-weighted nearest neighbor classification. a-h) Some of the approximate nearest neighbors to the hand view in Figures 3b, with associated grasp class $y_{i,j}$, distance in state-space $d_{i,j}$, and 3D orientation $\mathbf{o}_{i,j}$.

Locality Sensitive Hashing (LSH) [16], is employed. LSH is a method for efficient ϵ-nearest neighbor (ϵNN) search, i.e. the problem of finding a neighbor $\mathbf{x}_{\epsilon\text{NN}}$ for a query $\mathbf{x}$ such that

$$\|\mathbf{x} - \mathbf{x}_{\epsilon\text{NN}}\| \leq (1+\epsilon)\|\mathbf{x} - \mathbf{x}_{\text{NN}}\| \tag{2}$$

where $\mathbf{x}_{\text{NN}}$ is the true nearest neighbor of $\mathbf{x}$. This is done as (see [16] for details): 1) T different hash tables are created independently. 2) For $t = 1, \ldots, T$, the part of state space in which the dataset $\{\mathbf{x}_{i,j}\}_{i\in[1,M],j\in[1,N]}$ resides is randomly partitioned by K hyperplanes. 3) Every point $\mathbf{x}_{i,j}$ can thereby be described by a K bit binary number $\mathbf{f}_{t,i,j}$ defined by its position relative to the hyperplanes of table t. 4) As the total number of possible values of $\mathbf{f}_{t,i,j}$ is large, a hash function $h(\mathbf{f}_{t,i.j})$ gives the index to a hash table of fixed size H.

The ϵNN distance to the unknown grasp view $\mathbf{x}$ is now found as: 1) For each of the T hash tables, compute hash indices $h(\mathbf{f}_t)$ for $\mathbf{x}$. 2) Let $X_\cup = \{\mathbf{x}_m\}_{m\in[1,N_\cup]}$ be the union set of found examples in the T buckets. The ϵNN distance $\|\mathbf{x} - \mathbf{x}_{\epsilon\text{NN}}\| = \min_{m\in[1,N_\cup]} \|\mathbf{x} - \mathbf{x}_m\|$. In analog, the $\min(N_\cup, k)$ ϵ-nearest neighbors X_k are found as the $\min(N_\cup, k)$ nearest neighbors in $X_\cup$.

The parameters K and T for a certain value of ϵ is dataset dependent, but is learned from the normal data itself [24]. We use $\epsilon = 0.05$.

The computational complexity of retrieval of the ϵNN with LSH [16] is $\mathcal{O}(DN^{\frac{1}{1+\epsilon}})$ which gives sublinear performance for any $\epsilon > 0$. For examples of ϵ-nearest neighbors to the hand in Figure 3b, see Figure 5.

From X_k the estimated class of $\mathbf{x}$ is found as,

$$\hat{y} = \arg\max_i \sum_{j:\mathbf{x}_{i,j}\in X_k} \exp(-\frac{d_{i,j}^2}{2\sigma^2}) \ , \tag{3}$$

i.e. a distance-weighted selection of the most common class label among the k nearest neighbors, and the estimated orientation as

$$\hat{\mathbf{o}} = \frac{\sum_{j:\mathbf{x}_{\hat{y},j}\in X_k} \mathbf{o}_{\hat{y},j} \exp(-\frac{d_{\hat{y},j}^2}{2\sigma^2})}{\sum_{j:\mathbf{x}_{\hat{y},j}\in X_k} \exp(-\frac{d_{\hat{y},j}^2}{2\sigma^2})} \ , \tag{4}$$

i.e. a distance-weighted mean of the orientations of those samples among the k nearest neighbors for which $y_{i,j} = \hat{y}$. (The cyclic properties of the angles is also taken into account in the computation of the mean.) As we can see in Figure 5h, the orientation of a sample from a different class has very low correlation with the real orientation, simply because the hand in a different grasp has a different shape. Therefore, only estimates with the same class label as $\hat{y}$ are used in the orientation regression. All in all, the dependency between the state-space and the global Euler angle space is highly complex, and that is why it is modeled non-parametrically.

The standard deviation σ is computed from the data as

$$\sigma = \frac{1}{\sqrt{2}MN} \sum_i \sum_{j_1,j_2\in[1,N],j_1\neq j_2} \|\mathbf{x}_{i,j_1} - \mathbf{x}_{i,j_2}\| \ , \tag{5}$$

the mean intra-class, inter-point distance in the orientation histogram space [25].

The obviously erroneous neighbors in Figures 5g and 5h could maybe have been avoided with a larger

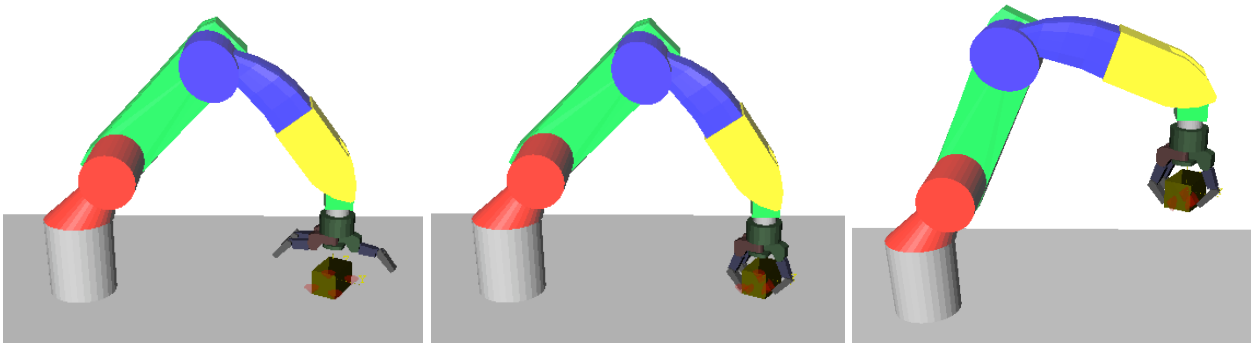


Fig. 6. Barret Wrap grasp, carried out on the same type and size of object as the human Large Diameter grasp shown in Figure 3b.

database containing hands of varying basic shape, such as male/female/skinny/fat/long-fingered/short-fingered hands. The hand in the test images (Figure 3b) is considerably different from the synthetic Poser 7 hand (Figures 3c, 5), and thus their 3D shapes are different even though they take the same pose. This poses no problem to the method in general; since the approximate kNN classification/regression has a sub-linear complexity, the database can be increased considerably to a limited computational cost.

VI. Example-Based Mapping of Grasp to Robot

To illustrate how the grasp classification can be employed for human-to-robot mapping in a pick-and-place scenario, a simulated robot arm is controlled with parameterized predefined grasping strategies as illustrated in Figure 1.

A human-to-robot grasp mapping scheme is defined depending on the type of robot hand used; here we use a Barret hand with three types of grasps as shown in Figure 2. The type of robot grasp defines the preshape of the robot hand.

The hand orientation estimate $\hat{\mathbf{o}}$ relative to the camera, along with the hand position estimate and the estimated position and orientation of the grasped object relative to the camera, are used to derive the estimated position and orientation of the human hand relative to the object, as depicted in Figure 1. The estimation of object position and orientation is assumed perfect; this part of the system is not implemented, instead the ground truth is given in the simulations.

In contrary to related grasp approaches [26], the robot here does not explore a range of approach vectors, but instead directly imitates the human approach vector, encoded in the hand position and orientation relative to the object. This leads to a much shorter computational time at the expense of the non-optimality of the grasp in terms of grasp quality. However, since the the selection of robotic preshape has been guided, the stability of the robotic grasp will be similar to the human one, leading to a non-optimal but successful grasp provided that the errors in the orientation and position estimate are sufficiently small.

An analysis of the robustness to position errors can be found in [26]. For an optimally chosen preshape, there is a error window ≥ 4 cm $\times$ 4 cm about the position of the object, within which the grasps are successful. The positioning of the robot hand can also be improved by fusing the estimated human hand position with an automatic selection of grasping point based on object shape recognition [27].

The robustness to orientation errors depends greatly on the type of grasp and object shape. We investigate the robustness of the Barret Wrap grasp with an approach vector perpendicular to the table (Figure 6). We get good results for orientation errors around the vertical axis of up to 15 degrees. As a comparison, the mean regression error of this orientation (Section VII-B) is on the same order as the error window size, 10.5 degrees, which indicates that the orientation estimation from the grasp classifier should be used as an initial value for a corrective movement procedure using e.g. the force sensors on the hand.

VII. Experimental Results

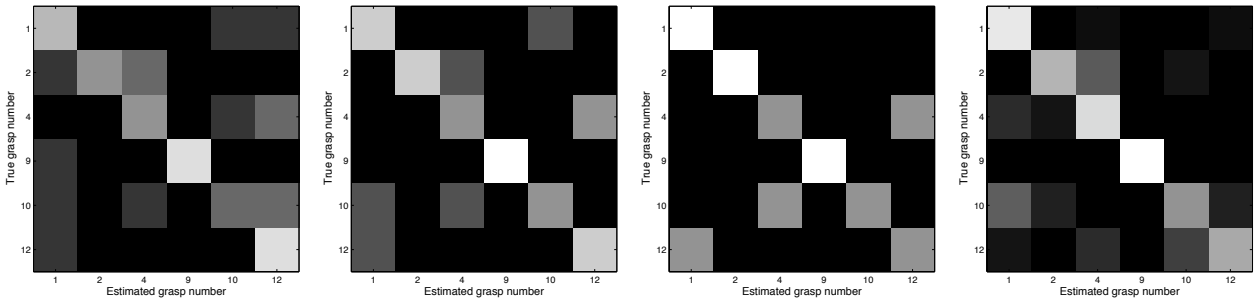
Quantitative evaluations of the grasp classification and orientation estimation performance were made.

For each of the six grasp types, two video sequences of the hand were captured, from two different viewpoints. From each video, three snapshots were taken, one where the hand was starting to reach for the object, one where the hand was about to grasp and one where the grasp was completed. This test set is denoted X.

The test examples from the beginning of the sequences are naturally more difficult than the others, since the hand configuration in those cases are closer to a neutral configuration, thus more alike than the examples taken closer to the completed grasp. It is interesting to study the classification rate for the different levels of neutrality, since it indicates the robustness to temporal errors when the robot grabs the image upon which the classification is based (Section III). In some tests below, we therefore removed the 12 most neutral examples from the test set, denoted X'. In other tests, we kept only the 12 most specific examples, denoted X''.

A. Classification of human grasps: Comparison of LSH and human classification performance

Figures 7a, 7b, and 7c show the confusion matrices for LSH classification of test set X, X', and X'', respectively. Apart from the fact that the performances on X' and X''



(a) LSH, all hand images: 61%. (b) LSH, hand close to object: 71%. (c) LSH, hand grasping object: 75%. (d) Human, all hand images: 74%.

Fig. 7. Confusion matrices for classification of the six grasps. White represents a 100% recognition rate, while black represents 0%. a) LSH performance, all hand images (X): 61% correct classifications. b) LSH performance, images with hand close to object (X'): 71% correct classifications. c) LSH performance, images with hand grasping object (X''): 75% correct classifications. d) Human performance, all hand images (X): 74% correct classifications.

are better than for X, it can be noted that the performance on Pinch grasp (9) and Precision Disc grasp (12) are very good. This is expected since these grasps are visibly very different from the others. Interestingly, it also concords with the mapping to the Barret grasps (Figure 2) in which these grasps have unique mappings while the others all are mapped to the same grasp. Note however that the human grasps map differently to more articulated robot hands.

The error rates alone say little about how the method would perform in a PbD system. The grasp recognition would there interact with methods for object, shape and action recognition, and a perfect performance in an isolated grasp recognition task is probably not needed.

How do we then know what error rate is "enough"? Humans are very good at learning new tasks by visual observation, and reach a near perfect performance on combined object, shape, action and object recognition. Human recognition performance on the same task as our classifier, with the same indata, would thus be a good baseline.

As an important side note, two things can be noted about this comparison. Firstly, in a natural learning situation, a human would use information about the grasped object and the motion of the hand as well. This information is removed for this experiment. As discussed in the Conclusions, we intend to integrate automatic grasp, object and action recognition in the future. Secondly, it is debated how important depth perception is for human recognition; humans perceive depth both through stereo and through prior knowledge about the hand proportions. For this experiment, we disregard depth as a cue in the human experiment.

Figure 7d shows the classification performance of a human familiar with the Cutkosky grasp taxonomy. The human was shown the segmented hand images $\mathbf{H}$ in the set X in random order and was asked to determine which of the six grasp classes they belonged to.

Interestingly, the human made the same type of mistakes as the LSH classifier, although to a lower extent. He sometimes misclassified Power grasp (10) as Large Diameter grasp (1), and Small Diameter grasp (2) as Abducted Thumb grasp (4). This indicates that these types of confusions are intrinsic to the problem rather than dependent on the LSH and training set. Since humans are successful with grasp recognition in a real world setting, these confusions are compensated for in some other way, probably by recognition of shape of the grasped objects. It can also be noted that the human was better at recognizing the most neutral grasps present in X but not in X' or X''.

Overall, the LSH performance is at par with, or slightly worse than human performance. This must be regarded as a successful experimental result, and indicates that the grasp recognition method can be a part of a PbD system with low error rate.

B. Classification of human grasps: Orientation accuracy

Figure 8 shows the mean orientation error for regression with X. The angular displacement of the two coordinate systems corresponds to how far off a robot hand would be in grasping an object without corrective movements during grasping. As noted in Section VI, the orientation estimate from this method should only be regarded as an initial value, from where a stable grasp is found using a corrective movement procedure.

VIII. Conclusions

PbD frameworks are considered as an important area for future robot development where the robots are supposed to learn new task through observation and imitation. Manipulating and grasping known and unknown objects represents a significant challenge both in terms of modeling the observation process and then executing it on the robot.

In this paper, a method for classification of grasps, based on a single image input, was presented. A grasping hand was represented as a gradient orientation histogram; a 2D image-based representation. A new hand image could be classified as one of six grasps by a kNN search among large set of synthetically generated hand images.

On the isolated task of grasp recognition, the method performed comparably to a human. This indicates that the method is fit for use in a PbD system, where it is used in interaction with classifiers of object shape and human actions. The dataset contained grasps from all expected viewpoints and with expected occlusion. This made the

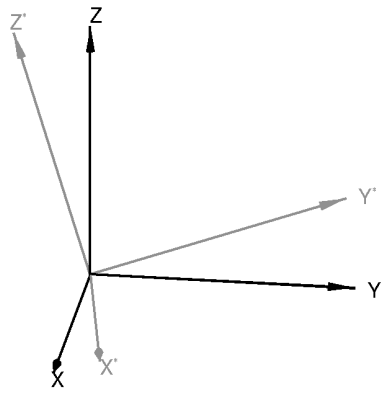


Fig. 8. Mean orientation error, all hand images (X): $(0, 0.29, 0.18)$ radians $= (0, 16.8, 10.5)$ degrees.

method view-independent although no 3D representation of the hand was computed.

The method was considered part of a grasp mapping framework, in which precomputed grasp strategies were compiled based on the detected type of grasp and hand-object orientation.

A. Future Work

It would be interesting to add an object orientation estimation technique to the system, and to execute the grasps on a real robot arm. Furthermore, we will investigate the inclusion of automatic positioning methods into the grasp strategies, as suggested in Section VI.

The classifier will also benefit from a training set with hands of many different shapes and grasped objects of different sizes. Although, this will increase the size of the database, the sub-linear computational complexity of the LSH approximate kNN search ensures that the computation time will grow at a very limited rate.

This paper discussed instantaneous recognition of grasps, recognized in isolation. Most probably, a higher recognition performance can be reached using a sequence of images over time. Moreover, there is a statistical correlation between types of objects, object shapes, human hand actions, and human grasps in a PbD scenario. We are therefore on our way to integrating the grasp classifier into a method for continuous simultaneous recognition of objects and human hand actions, using Conditional Random Fields (CRF) [28].

IX. Acknowledgments

This research has been supported by the EU through the project PACO-PLUS, FP6-2004-IST-4-27657, and by the Swedish Foundation for Strategic Research.

References

[1] S. Ekvall, "Robot task learning from human demonstration," Ph.D. dissertation, KTH, Stockholm, Sweden, 2007.

[2] Y. Kuniyoshi, M. Inaba, and H. Inoue, "Learning by watching," *IEEE Transactions on Robotics and Automation*, vol. 10, no. 6, pp. 799–822, 1994.

[3] S. Schaal, "Is imitation learning the route to humanoid robots?" *Trends in Cognitive Sciences*, vol. 3, no. 6, pp. 233–242, 1999.

[4] A. Billard, "Imitation: A review," in *Handbook of Brain Theory and Neural Networks*, M. Arbib, Ed., 2002, pp. 566–569.

[5] K. Ogawara, S. Iba, H. Kimura, and K. Ikeuchi, "Recognition of human task by attention point analysis," in *IEEE International Conference on Intelligent Robots and Systems*, 2000, pp. 2121–2126.

[6] M. C. Lopes and J. S. Victor, "Visual transformations in gesture imitation: What you see is what you do," in *IEEE International Conference on Robotics and Automation*, 2003, pp. 2375–2381.

[7] S. Ekvall and D. Kragić, "Grasp recognition for programming by demonstration tasks," in *IEEE International Conference on Robotics and Automation*, 2005, pp. 748–753.

[8] S. Calinon, A. Billard, and F. Guenter, "Discriminative and adaptative imitation in uni-manual and bi-manual tasks," in *Robotics and Autonomous Systems*, vol. 54, 2005.

[9] M. Cutkosky, "On grasp choice, grasp models and the design of hands for manufacturing tasks," *IEEE Transactions on Robotics and Automation*, vol. 5, no. 3, pp. 269–279, 1989.

[10] J. Rehg and T. Kanade, "Visual tracking of high dof articulated structures: An application to human hand tracking," in *European Conference on Computer Vision*, vol. 2, 1994, pp. 35–46.

[11] E. Ueda, Y. Matsumoto, M. Imai, and T. Ogasawara, "A hand-pose estimation for vision-based human interfaces," in *IEEE Transactions on Industrial Electronics*, vol. 50(4), 2003, pp. 676–684.

[12] V. Athitsos and S. Sclaroff, "Estimating 3D hand pose from a cluttered image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2003, pp. 432–439.

[13] C. Schwarz and N. Lobo, "Segment-based hand pose estimation," in *Canadian Conf. on Computer and Robot Vision*, 2005, pp. 42–49.

[14] L. Y. Chang, N. S. Pollard, T. M. Mitchell, and E. P. Xing, "Feature selection for grasp recognition from optical markers," in *IEEE International Conference on Intelligent Robots and Systems*, 2007.

[15] H. Murase and S. Nayar, "Visual learning and recognition of 3-D objects from appearance," *International Journal of Computer Vision*, vol. 14, pp. 5–24, 1995.

[16] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *International Conference on Very Large Databases*, 1999, pp. 518–529.

[17] Y. Wu and T. S. Huang, "Vision-based gesture recognition: A review," in *International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction*, 1999, pp. 103–115.

[18] K. Ogawara, J. Takamatsu, K. Hashimoto, and K. Ikeuchi, "Grasp recognition using a 3D articulated model and infrared images," in *IEEE International Conference on Intellingent Robots and Systems*, vol. 2, 2003, pp. 1590–1595.

[19] B. Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla, "Model-based hand tracking using a hierarchical bayesian filter," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1372–1384, 2006.

[20] E. Sudderth, M. I. Mandel, W. T. Freeman, and A. S. Willsky, "Visual hand tracking using non-parametric belief propagation," in *IEEE Workshop on Generative Model Based Vision*, 2004.

[21] A. A. Argyros and M. I. A. Lourakis, "Real time tracking of multiple skin-colored objects with a possibly moving camera," in *European Conference on Computer Vision*, vol. 3, 2004, pp. 368–379.

[22] W. T. Freeman and M. Roth, "Orientational histograms for hand gesture recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 1995.

[23] G. Shakhnarovich, P. Viola, and T. Darrell, "Fast pose estimation with parameter sensitive hashing," in *IEEE International Conference on Computer Vision*, vol. 2, 2003, pp. 750–757.

[24] B. Georgescu, I. Shimshoni, and P. Meer, "Mean shift based clustering in high dimensions: a texture classification example," in *IEEE International Conference on Computer Vision*, vol. 1, 2003, pp. 456–463.

[25] I. W. Tsang, J. T. Kwok, and P.-M. Cheung, "Core vector machines: Fast SVM training on very large data sets," *Journal of Machine Learning Research*, no. 6, pp. 363–392, 2005.

[26] J. Tegin, S. Ekvall, D. Kragić, B. Iliev, and J. Wikander, "Demonstration based learning and control for automatic grasping," in *International Conference on Advanced Robotics*, 2007.

[27] A. Saxena, J. Driemeyer, J. Kearns, and A. Y. Ng, "Robotic grasping of novel objects," in *Neural Information Processing Systems*, 2006.

[28] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *International Conference on Machine Learning*, 2001.