Coordinating Touch and Vision to Learn What Objects Look Like

Walter A. Talbott[1,4], Ian Fasel[2], Javier Rodriquez Molina[3], Virginia de Sa[1], and Javier Movellan[4]

[1]Department of Cognitive Science, University of California San Diego La Jolla, CA 92093
[2]Department of Computer Science, University of Arizona, Tucson, AZ 85721
[3]CalIT2, University of California San Diego, La Jolla, CA 92093
[4]Machine Perception Lab, University of California San Diego, La Jolla, CA 92093

Abstract

We use contemporary machine learning methods to explore Piaget's idea that active interaction across modalities may be the engine for constructing our knowledge about objects. We identified the existence of modality-specific invariances as a potential mechanism by which Piaget's ideas may be implemented in practice. For example, object segmentation and pose invariant recognition are very difficult in the visual domain but trivial in the tactile/proprioceptive domain; touching an object easily delineates its physical boundaries. We can also grab an object and rotate it without modifying the proprioceptive and tactile information from our hands. We hypothesize that this information may provide invariants that could be useful for training a visual system to recognize and segment objects.

To explore this hypothesis we developed the instrumentation necessary to simultaneously collect tactile, proprioceptive and visual information of a person interacting with everyday objects. We then developed a system that learns pose invariant visual representations using proprioceptive and tactile information as the only training signal.

The classifiers that developed from this approach were accurate and robust to variations in pose and to a wide range of occlusions. They were more accurate (average 2AFC= 0.98) than the classifier trained with human-specified location information (average 2AFC = 0.93). This suggests a specific mechanism using multi-modal information could to construct knowledge about objects, as originally proposed by Piaget.

Introduction

Pose invariant object recognition is one of the most difficult computer vision problems, since objects can change appearance drastically depending on the orientation. Particularly difficult is to learn the appearance of objects in an unsupervised manner, i.e., without any labels telling us where the objects of interest are or whether they are present in the image at all. Yet humans appear to have no problems learning the visual appearance of objects relatively independent of their pose and in a fully unsupervised manner. Developmental psychologists like Piaget (Piaget, 1953), have long argued that infants construct knowledge about objects based on the mutual interaction (assimilation and accommodation) between different modalities (grasping, sucking, looking). Here we take some first steps toward understanding how this may work in practice. We focus on the interaction between tactile and visual modalities and note that the hands may provide invariants that could be used to train the visual system. In particular, by grabbing an object and moving it around, it is possible to create dramatic changes in the visual appearance of the object while maintaining invariant information from the tactile and proprioceptive (joint angles) sensors in the hand.

Recently several research groups have begun to explore the interactions between sensory modalities for improving the performance of systems that interact with objects. In (Grzyb, Chinellato, Morales, & Pobil, 2009), the grasping actions of a robotic system are planned by a visual algorithm that estimates the shape of unmodeled objects. After this initial planning, the tactile sensors guide the final approach of the grasp to correct errors in the visual estimation. Motor commands and visual input are used in (Gold & Scassellati, 2009) to learn a representation of a robot's arm through contingencies between the commands and observed motion. In (Saxena, Wong, & Ng, 2008), a camera is positioned, by a robotic arm, in several orientations, and the proprioceptive information from the arm at each orientation helps locate a grasping point on an object. (Orabona, Caputo, Fillbrandt, & Ohl, 2009) and (Noceti et al., 2009) directly train a supervised mapping between the visual and haptic sensors from human interaction with an object. This mapping is used to estimate missing tactile input from the visual input, and classification using the reconstructed input and the visual information is more accurate than with the visual input alone. In (Fitzpatrick & Metta, 2003), object properties are learned through interaction by making a robot poke the object and examining changes to the visual scene after contact. Also, (Sinapov & Stoytchev, 2009) and (Bergquist et al., 2009) train a robot to categorize household objects using visual paired with acoustic and proprioceptive cues during supervised interaction, where object labels were given to the system by humans.

Here we focus on the problem of using tactile and proprioceptive information from the hand to train a visual system to recognize objects. First, unsupervised methods are used to cluster the tactile and proprioceptive sensory data while freely interacting with two objects: a drinking glass and a plate. The cluster labels provided by the hand are then used to train a weakly-supervised visual object recognition algorithm. This approach does not use human labels as a training signal, but instead attempts to learn objects in an unsupervised way, which sets it apart from the previous work in multi-modal object recognition.

Data Collection

To collect the data required for the project, we developed a suite of sensors. We first attached 12 piezoresistive pressure sensors to a sports glove. The pressure sensors were located at the fingertips, in the palm at the base of each finger, and the last two at the bottom of the palm, as shown in Figure 2. These sensors vary their resistance based on the force applied on their surface. The touch data was collected at a 1KHz sampling rate.

We also equipped the glove with six PhaseSpace motion capture (*Phasespace Motion Capture*, n.d.) LEDs, placed on the back of the hand, wrist, forearm, and elbow. Each of these markers provides a measurement of the three dimensional position of the arm and hand in space with a 400Hz sampling rate. These sensors were then integrated on top of a Cyberglove (*Cyberglove Systems*, n.d.), which provided measures of the angles of 18 joints of the hand and fingers at a 150Hz sampling rate.

Visual input was captured using a head-mounted camera. Since the objects were not fixed in place during the interactions, it was important to ensure that the visual input matched what the human was seeing while manipulating the objects. To this end, the output from the head-mounted camera was directed to a pair of VGA glasses that gave the only view the human had of the interaction. Data were recorded at 20 Hz with a resolution of 640x480 pixels. Three PhaseSpace LEDs were placed on the head to capture the 3D position and orientation of the head with respect to the world (See Figure 3).

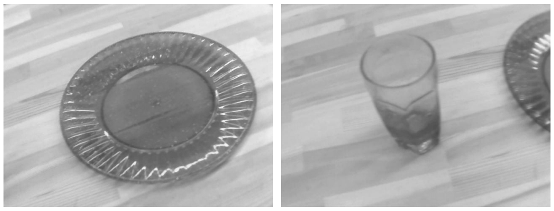

Figure 1: Images of the two objects, a drinking glass and a plate. Because they were both on the table during the interactions, images can include both objects.

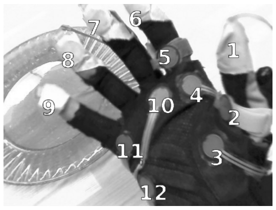

Figure 2: Layout of the 12 force sensors on the fingertips and palm.

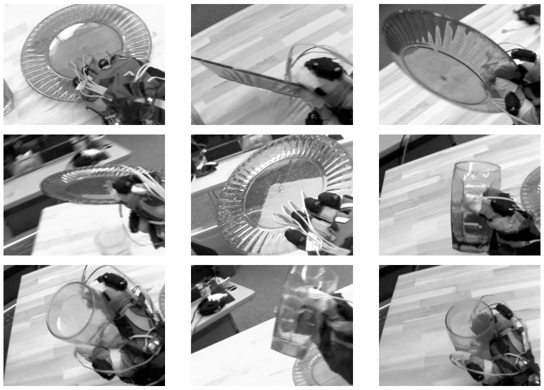

Figure 3: Example images from the interactions with the objects. Interactions were unscripted, with varying backgrounds, and could include both objects in the image. The objects were additionally present in multiple orientations.

The goal of the study was to investigate mechanisms by which a modality (tactile/proprioceptive) may train another modality (visual) in an unsupervised manner so as to construct object concepts. To this end we started with what we thought would be a relatively simple problem, to visually recognize the presence of two objects: a drinking glass and a plate. We collected data from 6 minutes of unscripted human interactions with these objects, shown in Figure 1.

After the data were collected, each sensor was downsampled to the 20Hz rate of the video capture, and the images were converted to grayscale at a resolution of 320x240 pixels. To make the motion capture information invariant to location in space, the three dimensional coordinates were converted into angles between vectors defined by the points tracked by the markers on the arm. The data processing leaves 7521 samples with 5 motion capture angle readings, 18 cyberglove readings, 12 force sensor readings, and the corresponding 320x240 grayscale image. Of these samples, 1152 were during grasps of the drinking glass, and 1602 from grasps of the plate.

Learning to Recognize Objects

The problem of recognizing the visual appearance of the target objects turned out to be much more difficult than we had originally anticipated. In the past, we have worked with standard datasets popular in the computer vision literature in which the task is to recognize hundreds of object categories (Fasel, 2006) and so were surprised by the difficulty of our dataset. We do not have a clear explanation yet, but believe that while our database includes only two target objects, it presents some difficult challenges: (1) The objects appeared in a wide range of poses and locations in the image plane. In many cases, the pose was such that the drinking glass looked more like a plate and viceversa (see Figure 4). (2) The hand was visible and occluded different parts of the object. The recognition system had to implicitly factor out the hand and later recognize the target objects on their own. (3) In a large proportion of images both objects were visible simultaneously. (4) The target objects were transparent.

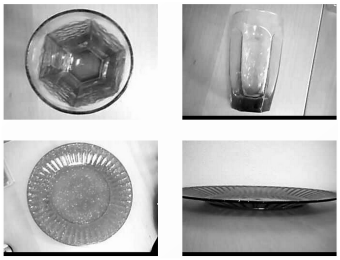

Figure 4: Images of the two objects, a drinking glass and a plate. From certain angles, these objects look more similar to each other than to themselves at different angles.

Unsupervised Clustering

We first tried clustering the data using K-means and spectral clustering. K-means is one of the simplest and most popular approaches to unsupervised clustering. Spectral clustering makes use of the spectrum of the similarity matrix of the data and is one of the most popular algorithms used for unsupervised image segmentation. We found that, in this case, spectral clustering provided marginally worse results than K-means and thus here we focus on the results with the K-means clustering.

Tactile Clustering

The 5 motion capture angles, the 18 cyberglove readings, and the 12 force-sensor readings were normalized and combined into a single vector for each sample. Clusters were then computed using $K = 3$, with an aim of capturing the two object classes plus a background class for when the hand was not interacting with an object.

After combining all of the tactile sensors and clustering, the samples were separated into groups as shown in Table 1. A measure of accuracy can be computed between the two clusters that correspond to samples when the glass and plate are held. The measure takes the sum of the samples that are truly glasses in what is interpreted as the glass cluster (the cluster with more glass images) and the samples that are truly plates in what is interpreted as the plate cluster, and divides by the total number of samples in both clusters. This measure gives the clustering an accuracy of 0.8553.

Thus, the clusters appear to represent the object categories, albeit imperfectly. These clusters are based on the haptic information, and therefore do not reflect which objects are actually in the image. Table 2 shows, for each of the two clusters with highest force sensor values, the number of images that contain each object. Notice that while these clusters were not generated from visual information, they separate the objects well; only 4.5% of the images with one object are incorrectly clustered.

Table 1: Results of clustering samples based on haptic information. Entries show number of samples in each cluster separated by true label.

	Actually Holding		
	Glass	Plate	Neither
Cluster 1	836	72	39
Cluster 2	3	1387	262
Cluster 3	313	143	4466

Table 2: Number of objects present in the images of each cluster. Numbers here represent images that contain part or whole of each object.

	Neither	Glass	Plate	Both
Cluster 1	2	279	70	596
Cluster 2	3	0	1193	456

Visual Clustering

Table 3 contains the same data as Table 2 but for the clusters generated from visual information alone. Cluster 2 may be taken to represent plates, but neither of the other clusters represents a drinking glass cluster. Similar results were obtained with spectral clustering approaches. Thus, for this particular dataset, vision alone does not seem to easily separate the data into the target object clusters.

Table 3: Content of the clusters obtained from clustering the visual data (used to train the Plate-Vision and Glass-Vision classifiers) used as reference for performance. Entries are the number of images in each cluster that contain images of each, neither, or both objects. For training the classifiers, Cluster 1 and Cluster 2 were chosen.

	Neither	Glass	Plate	Both
Cluster 1	76	450	756	1539
Cluster 2	27	118	1332	424
Cluster 3	74	418	967	1340

No matter which clustering algorithm is used, the images from the clusters can contain both visual objects. Given the noisy, unlabeled data of transparent objects at different orientations, it seems a difficult task for a classifier to learn to distinguish the objects from these data.

Learning The Visual Appearance of Objects

Here we attempted to use the labels obtained from the clustering of a modality to train another modality. Since the clustering methods (tactile or visual) are unlikely to perfectly separate the data into the target object classes, it is important to choose a learning algorithm that can operate in the presence of a large number of errors in the training labels. In addition, the clustering algorithms provide information about image types but no information about where the object is in the image. Thus a learning algorithm is needed that can work with such weak labels.

We chose the Segmental Boltzman Fields (SBF) algorithm (Fasel, Fortenberry, & Movellan, 2005; Fasel, 2006). This approach assumes that images are generated as a collection of rectangular patches. Each of these patches generates either the background or the object of interest from the respective distributions describing the pixel values in a patch. To perform inference, each possible patch in an image is assigned a probability based on the likelihood the pixels in the patch were generated by the object distribution. This likelihood is combined with a prior probability that an object was contained in that image patch. Training this model involves the estimation of the likelihood ratio between these two distributions for image patches.

After the tactile/proprioceptive clustering separated the data into three groups, the two groups that had the highest mean readings on the force sensors were used as the labeled groups. This procedure relies on the assumption that the system prefers situations where objects are present in the hand to when the hands are empty. Call the two clusters A and B. Two classifiers were trained. One of these was trained using A as positive, foreground, examples and B as negative, background, examples. The classifier trained this way can then be assumed to recognize objects contained in A. The other classifier was trained using B as positive examples and A as negative examples, so that it will recognize the object in B.

Segmental Boltzman Fields

The model of foreground (images that contain the object of interest) and background images (images to that do not contain the object) specified in (Fasel et al., 2005; Fasel, 2006) gives a log likelihood function for a foreground image x as follows

$$L(x|f,\hat{b}) = log\sum_{j=1}^{n_s} e^{f(x_j)} K_j(\hat{b}) - log\sum_{k=1}^{n} e^{f(\tilde{x}_k)} + log(n)$$

where f and $\hat{b}$ are the foreground and background models, x_j is segment j from image x, n_s is the number of segments in the positive image, n is the number of segments in the background image set, $\tilde{x}_k$ is segment k from the background images, and $K_j(\hat{b})$ is a measure of how well the background explains the image assuming it contains segment j. Additionally, f is of the form

$$f(x) = \sum_{i=1}^{t} \alpha_i h_i(x).$$

In the current case, each $h_i(x)$ is a step function of the output from a Haar-like feature applied to image segment x which can take values of ± 1, and α_i is a real-valued weight. The goal of learning is to find the model f that maximizes the log likelihood L. However, because computing the K terms is intractable, we instead will attempt to maximize

$$L(x|f,\hat{b}) = log\sum_{j=1}^{n_s} e^{f(x_j)} - log\sum_{k=1}^{n} e^{f(\tilde{x}_k)}$$

since we are concerned only with the choice of f, and the K terms are always positive and constant with respect to this choice.

This likelihood function is maximized using functional gradient ascent by boosting the components of the foreground model, $h_i(x)$.

Results

In order to get a standard benchmark, we first developed a plate and a drinking glass detector using a supervised learning approach. For these supervised approaches, the objects were cropped by hand from the images separated by the tactile clustering. The negative sets for these classifiers were the whole, uncropped images from the other cluster (glass for plate, and vice versa). These sets, positive and negative, were given to the SBF algorithm (Fasel, 2006), which learned supervised classifiers from them. The classifiers trained this way are called Supervised classifiers in tables 4 and 5. The performance of the classifiers was measured in terms of the performance in a 2 alternative forced choice task (2AFC).

We then developed visual classifiers using the labels provided by the tactile/proprioceptive clusters. The two classifiers, one for each cluster, were trained on subsets of 64 images chosen at random from the pool of images separated by the clustering. The classifiers were evaluated on two sets of data: The first set contained images of the instrumented hand grasping the objects, similar to the data on which the classifiers were trained. The second set contained images of the objects alone, without the instrumented hand present in the image. It also contained two distractor objects, a pen and a mug. Neither of these distractor objects was present in the training set. This second set was designed to rule out the possibility that the system was learning to recognize the fact that the hands look different when grasping different objects, rather than learning the appearance of the objects themselves. The classifiers trained using the tactile clusters are called Tactile in the tables. Results presented below are from classifiers for each category that had the highest cross-validated performance on the training set.

As an additional control, we developed classifiers trained using clusters generated from the images alone, without any tactile information. The images were split into three clusters. Since there was no signal to distinguish which contained an object and which did not, the clusters that best contained instances of the plate and glass were selected by hand, and trained with SBF. The content of these clusters is described in Table 3. These classifiers represent how how well the objects could be learned from the dataset's visual information alone. These are called the Vision classifiers in the tables.

Table 4: Performance of the classifiers on the data containing the hand manipulating the objects. These data were similar to the data on which the classifiers were trained.

2AFC		
Tactile	Vision	Supervised
0.968	0.782	0.938

Table 5: Performance of the two classifiers on the data with the objects alone. These data were tested to show that the classifier had not learned to recognize the hand.

2AFC		
Tactile	Vision	Supervised
0.983	0.529	0.935

Object Localization

The SBF algorithm provides posterior probability maps representing the presence or absence of target objects. The intensity of a pixel on these maps indicates the probability that that particular pixel renders the object of interest. In addition, the algorithm can select the most probable location of the target object.

Figure 5 shows the estimated locations for the classifier trained to identify glasses. The first two images show examples of what the classifier estimated for images with and without the hand. The classifier seems to have picked out the curve of the glass rim in the first image, but manages to cover a more complete area of the glass when the hand is present. In the third image, notice that when the glass is viewed horizontally, the classifier is not confident enough to predict that the object is present in the image.

Figure 6 shows posterior probability maps from the classifier that learned to detect images with plates. The first three images are examples of when the classifier identified the plate well. These images contain the plate at different orientations, and the classifier manages to locate the plate even in a side-on view. The fourth image shows a mistake. The base of the glass and some of the background (not shown) are occasionally selected as the location of the object in the image.

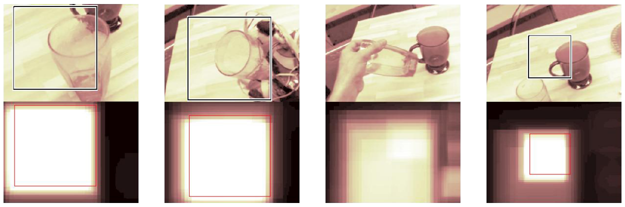

Figure 5: Example localization estimated by the Glass classifier. Each image contains a box around the area it estimates is most likely to have generated the image of the object. The lower section of the images is a heatmap of the probability that an object is present in the image at each location. This heatmap is normalized individually for each image, so direct comparisons between heatmaps are difficult.

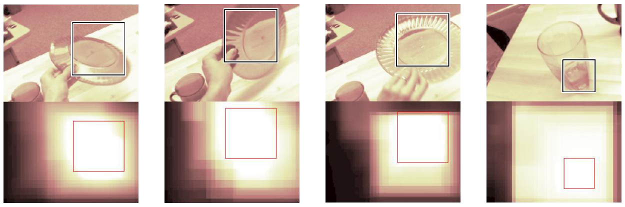

Figure 6: Example localization estimated by the Plate classifier.

Conclusion

The goal of this project was to use contemporary machine learning methods to investigate Piaget's idea that active interaction across modalities may be the engine for the construction of object knowledge (Piaget, 1953). We identified the existence of modality-specific invariances as a potential mechanism by which Piaget's ideas may be implemented in practice. For example, object segmentation and pose invariant recognition are very difficult in the visual domain but trivial in the tactile/proprioceptive domain. This creates an opportunity for tactile information to be used to learn the location and appearance of objects in images.

To test these ideas we developed the instrumentation necessary to simultaneously collect tactile, proprioceptive and visual information of a person interacting with two everyday objects (a drinking glass and a plate). While the dataset we obtained contained only two target objects, there were a large number of occlusions (due to the presence of the hands) and a wide variety of 3D poses. When given the supervised location of the object, the relatively mediocre results (average 2AFC of 0.93) of the resulting classifier indicate that the task was not easy.

We then used a simple unsupervised clustering method on the tactile/proprioceptive information to separate the observed data into three clusters. These clusters roughly mapped into episodes of interaction with each of the two target objects and episodes in which the target objects were not present. The information provided by unsupervised clustering in the tactile/proprioceptive channel was then used to train a visual classifier.

The classifiers that developed out of this approach were accurate and robust to variations in pose and to a wide range of occlusions. They were more accurate (average 2AFC= 0.98) than the classifiers trained in a supervised manner (average 2AFC = 0.93). If confirmed by future studies, this would be a remarkable result, suggesting that tactile information may indeed provide critical invariances for the construction of object detectors in the visual domain. Additionally, the superiority of the tactile and visual classifier over the visual only classifier lends support to the idea that multi-modal integration may be better than individual sensory modalities when used to generate object knowledge.

The results are a first step. We need to study how the proposed approach scales up as we add a larger number of objects. In this project, we addressed the perception problem decoupled from the motor control problem, i.e., we let a human move his hands and change his visual input at will. In practice infants face a combined perceptual and control problem and they may use this opportunity to optimize the knowledge gained about objects. In some conditions they may choose to move an object in front of their eyes while maintaining a constant hold of the object. In such cases, the tactile system may provide useful invariances to train the visual system. In other conditions, they may choose to look at a stationary object while touching it in different locations, when the visual system would provide invariances to train the tactile system. One of our immediate goals is to formalize this problem from the point of view of information maximization approaches to motor control (Butko & Movellan, 2010). We are also planning to test such formalism on Diego-San, a humanoid robot we developed to help us understand cognitive development from a computational point of view.

Acknowledgment This research was supported in part by ONR DURIP Award #N000140811114 (H. Poizner, PI).

References

Bergquist, T., Schenck, C., Ohiri, U., Sinapov, J., Griffith, S., & Stoytchev, A. (2009). Interactive object recognition using proprioceptive feedback. In *Proceedings of the iros 2009 workshop: Semantic perception for mobile manipulation.*

Butko, N. J., & Movellan, J. R. (2010, October November). Detecting contingencies: An infomax approach. *Neural Networks*, *23*(8-9), 973–984.

Cyberglove Systems. (n.d.). Available from `http://www.cyberglovesystems.com/`

Fasel, I. (2006). *Learning real-time object detectors: probabilistic generative approaches.* Unpublished doctoral dissertation, UCSD.

Fasel, I., Fortenberry, B., & Movellan, J. (2005, April). A generative framework for real time object detection and classification. *Computer Vision and Image Understanding*, *98*(1), 182–210.

Fitzpatrick, P., & Metta, G. (2003). Grounding vision through experimental manipulation. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, *361*, 2165–2185.

Gold, K., & Scassellati, B. (2009). Using probabilistic reasoning over time to self-recognize. *Robotics and Autonomous Systems*, *57*(4), 384–392.

Grzyb, B., Chinellato, E., Morales, A., & Pobil, A. del. (2009). A 3D grasping system based on multimodal visual and tactile processing. *Industrial Robot: An International Journal*, *36*(4), 365–369.

Noceti, N., Caputo, B., Castellini, C., Baldassarre, L., Barla, A., Rosasco, L., et al. (2009). Towards a theoretical framework for learning multi-modal patterns for embodied agents. In *Proceedings of the 15th international conference on image analysis and processing* (p. 248). Springer.

Orabona, F., Caputo, B., Fillbrandt, A., & Ohl, F. W. (2009, June). A theoretical framework for transfer of knowledge across modalities in artificial and biological systems. *2009 IEEE 8th International Conference on Development and Learning*, 1–7.

Phasespace Motion Capture. (n.d.). Available from `http://www.phasespace.com/`

Piaget, J. (1953). *The origins of intelligence in children.* London: Routledge and Kegan Paul.

Saxena, A., Wong, L. L. S., & Ng, A. Y. (2008). Learning grasp strategies with partial shape information. In *Aaai'08: Proceedings of the 23rd national conference on artificial intelligence* (pp. 1491–1494). AAAI Press.

Sinapov, J., & Stoytchev, A. (2009). From acoustic object recognition to object categorization by a humanoid robot. In *Proceedings of the rss 2009 workshop: Mobile manipulation in human environments.*