OBJECT RECOGNITION BASED ON SHAPE AND FUNCTION: INSPIRED BY CHILDREN'S WORD ACQUISITION

By Akihiro Eguchi
Department of Computer Science and Computer Engineering

Faculty Mentor: Dr. Craig Thompson
Department of Computer Science and Computer Engineering

Abstract

This paper explores a new approach to computational object recognition by borrowing an idea from child language acquisition studies in developmental psychology. Whereas previous image recognition research used shape to recognize and label a target object, the model proposed in this study added the function of the object resulting in a more accurate recognition. This study makes use of new gaming technology, Microsoft's Kinect, in implementing the proposed new object recognition model. A demonstration of the model developed in this project properly infers different names for similarly shaped objects and the same name for differently shaped objects.

Introduction

Object recognition is a subfield of machine vision and artificial intelligence. Earlier object recognition research required significant knowledge of the mathematics of image processing and expensive equipment like Light Detection and Ranging (LIDAR) cameras, which are commonly used, or SICK sensors, named after its founder Dr. Erwin Sick, which are often used for logistics applications in industry. The recent introduction by Microsoft of the Kinect for the Xbox 360 has changed the landscape and enabled many researchers to tackle a range of image recognition problems without such extensive knowledge or expensive equipment. Most research on developing object recognition models has been based only on the shape of the objects. Therefore, there has been a difficulty in recognizing objects, such as a uniquely designed chair, or different objects that have similar shapes.

In order to solve this problem, this study focuses on a strategy that human children use to solve problems. For example, in the field of psychology, many studies have focused on language acquisition by children. Results indicate that there exist two strong biases children use when they try to learn names of objects: *shape bias* and *function bias* (Landau, Smith, & Jones, 1988; Nelson, Russell, Duke, & Jones, 2000). Furthermore, recent studies confirm that 2 to 3 year old children exhibit great skills in learning associations between novel functions and target objects, which supported the possible important role of function in learning about an object (Eguchi, 2012). On the other hand, most past studies of object recognition in computer science have focused on shape bias (Belongie, Malik, & Puzicha, 2002; Toshev, Makadia, & Daniilidis, 2009); only a few studies have focused on the function of objects to recognize objects (Grabner, Gall, & Van Gool, 2011). In order to learn the name of object, such as a chair, if a computer can learn the name not only based on its shape but also based on its functionality (i.e., a place to sit), then the program can perform more flexible object recognition in a manner more similar to humans.

The objective of this research is to combine the Kinect sensor with machine learning techniques to implement an object recognition model that uses both *shape bias* and *function bias* to learn the names of objects in a manner similar to how human children acquire names of

objects.

Background

Machine Learning

Machine learning is an important subfield of artificial intelligence aimed at giving computers a way to learn many kinds of things without explicitly being programmed. Examples of machine learning occur in domains like autonomous vehicles, checker playing, signal processing, and market simulation (Eguchi & Nguyen, 2011). Machine learning is also commonly used in the image processing area, especially for recognition of objects, including the human face and everyday objects.

Machine learning is based on two types of training: supervised learning and unsupervised learning. In supervised learning, training data always give the right answer "y" to the corresponding input "x" so that the program determines a pattern to predict the function f(x) = y. For unsupervised learning, a training set does not contain the output "y", but instead, the computer must figure out the hidden structure of the data. This paper uses supervised learning because the names of objects will be explicitly provided by the trainer.

A primitive model of machine learning based on experience is called rote learning. In this model, the program stores all input "x" and associated output "y" in the memory. While this is a very simple algorithm, this type of learning only works for small discrete numbers of possible input and output. Another way is to use neural networks. In this approach, all Boolean values of input vector "x" will be nodes in the input layer of the network and the corresponding output "y" will be at the output layer. There can be several layers of nodes between the input layer and the output layer, and with many times of iteration, the algorithm figures out a strong association of each node in the network. This is a generalizable technique; however, the problem is that if the number of nodes in the training set becomes large, then the time to build the network will significantly increase. Additionally, overfitting is a problem in these techniques (i.e., generalizing the solution too much, resulting in failure of correct identification). Even if the model can perfectly predict an output "y" from an input "x" for a training set of data, it does not guarantee that the model is perfect for other data as well.

One primitive way to deal with the overfitting problem is via a K-nearest neighbor (K-NN) clustering. Like rote learning, it stores all sets of <x, y> in the memory, but in the testing phase, the algorithm takes k <x, y> vectors from the memory of which the "x" value is most similar to the target input "x". Then, the algorithm determines the predicted output "y" by taking a majority vote. Because of its simplicity, this method contains many problems and is seen as old-fashioned. Similar methods include a support vector machine, which can be regarded as a refined version of K-NN clustering.

Microsoft Kinect for the Xbox 360

The Kinect is a sensing technology originally created for a controller for the Microsoft gaming console Xbox 360; it was released to the public in November 2010. The features of this sensor include dynamic depth image retrieval, human body recognition, skeletal joint tracking, and a multi-array microphone, all at a cost of approximately $150. Because of the high versatility and the low cost of the sensor, many researchers have experimented with the Kinect in their projects. Even before the Kinect SDK was officially released in June 2011, Kinect videos were already appearing on Internet video streaming websites like YouTube.

This research study depends directly on several of the features of the Kinect to build its object recognition model. The Kinect can retrieve pixel-by-pixel distance map using its infrared sensor and a photo image using an RGB camera, which represents an image with combination of

three different colors: red, green, and blue. It provides a feature that recognizes players' real-time motion and posture. It employs a machine learning technique to learn the shape of individuals so that it can extract the image of a human from a background or tell multiple humans apart (Girshick, Shotton, Kohli, Criminisi, & Fitzgibbon, 2011). Additionally, by building a classifier to identify body parts, this Kinect sensor can track 20 different joints on a human body.

Study in Developmental Psychology

In the field of psychology, researchers study the strategy or mechanism that children use to acquire their first languages. They develop theories and models from different perspectives. They attempt to simulate language acquisition by implementing such models on a computer. For infants who have not yet acquired language, findings show that they have special skills, like pattern recognition, that are acquired before later learning more complex structures of linguistic communication (Saffran, Aslin, & Newport, 1996).

For infants and toddlers to learn the names of objects, a shape bias is one of the strongest cues used to categorize novel objects; in other words, infants tend to generalize the name of an object based on similar shapes of the objects (Landau, Smith, & Jones, 1988). Also, research has demonstrated that just with the shape bias, three-to-four month old infants can categorize objects in a similar way with a basic categorization by semantic meaning (Eimas & Quinn, 1994).

Function-bias is another cue for object-name learning. Similar research shows that two year-old children who learned the name of an object of which they can easily infer the function or use, also generalized the name to other similarly functioning objects (Nelson, Russell, Duke, & Jones, 2000).

Object Recognition in Computer Science

In the field of computer science, researchers have studied ways of labeling names of objects by running a machine-learning technique on three dimensional point cloud data retrieved from both real world environments and Google 3D warehouse, which stores 3D models of objects. In this way, Google 3D warehouse can teach the name of objects in the real world based on the shape of the model objects (Lai & Fox, 2010).

An interesting question posed by Grabner, Gall, and Gool (2011) was "What makes a chair a chair?" Their group has implemented something similar to function-bias to recognize a particular object, such as a chair. First, based on a *shape-bias*, their group generalized the shape of chairs. Then, they defined a chair as something we can sit on, and built a model that can infer if the target object is *suitable for sitting on* or not. They first tested the procedure in a 3D virtual world setting, and then successfully applied the same method in a real world environment where data were reconstructed with 80 images to recognize chairs even if the shape is unusual.

Activity Recognition

In everyday activities, humans perform many routine tasks from brushing their teeth to performing heart operations. These tasks involve objects that they see and touch. Consider a log of data trace in the form <time stamp, location stamp, observation> where observations could be Radio Frequency Identification (RFID) reads, smart phone actions like receiving or responding to a text message, Kinect readings, and other types of "sensory inputs". Also, consider a collection of workflows, which are named activities consisting of a set of steps, some of which may result in leaf level trace data. In order to record human activity, previous research studies assigned an RFID tag to objects around individuals and participants wore RFID reader-embedded gloves. Then, after capturing trade data from the participants in the experiment consisting of a log of touch events, the researchers used the log of the order of touched objects to

identify probable workflow activities of daily living, performing this task using dynamic Bayesian networks (Philipose, Fishkin, Perkowitz, Patterson, Fox, Kautz, & Hahnel, 2004). In order to recognize simpler segments of activities, other researchers used four seconds of video recorded action data like walking, jogging, and running, and ran a support vector machine (SVM) learning algorithm to train a classifier to recognize sequences of those primitive actions (Schuldt, Lapteve, & Caputo, 2004).

Approach and Architecture

Architecture of the Object Recognition Model

The proposed object recognition model consists of two distinct methods: to infer the name of object based on the object's shape and to identify its functional use. In a teaching phase, the user sets a target object in front of the Kinect sensor and the program learns the object's shape and name. Then the user can perform some action associated with the object to teach the object's function. Later in the testing phase, the program infers the name of an object presented based on the object's shape and its function.

Plane Surface Removal with RANSAC Algorithm

Kinect can instantly retrieve a depth map and an RGB image from a real world environment, so we might assume that Kinect provides a similar input individuals retrieve from using our eyes. However, the problem is how to separate the target object for Kinect to learn from its background. We can use the already implemented seek-bar to adjust the range of focus. It is then necessary to remove the surface where the object is placed. One simple way to do that is to use the Random Sample Consensus (RANSAC) routine for the 3D point clouds (Figure 1). The idea is that, first the program randomly takes three points from the point cloud to determine a random plane; then, it counts how many points are on the plane. By iterating through this steps many times, the program finds the plane that has the maximum number of points, assuming the plane is the surface where the objects are placed.

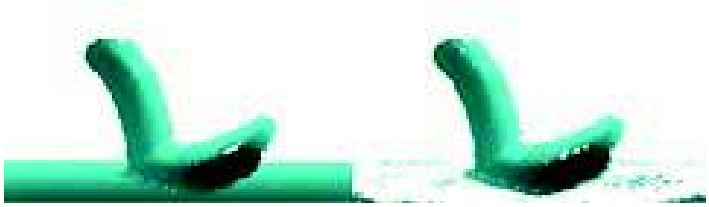


Figure 1. Using RANSAC, background surface was detected and removed.

SVM for Shape Learning

Once we receive a depth image of the target object, the program asks the user to name the object or asks the user to choose the name from a list of names the user has already entered into the program. The user can change the angle of the same target object and label the object with the same name. Then, in the testing session, the program compares the shape of the target object with all the shape information in the memory and chooses the closest shape of objects using SVM to determine the name of the target.

Activity Recognition using the Kinect

A weakness of an object recognition model based only on shape is that although it can recognize an object that has a similar shape to the objects in its memory, it cannot generalize the name to other objects that have the same functionality but a different shape (e.g., a chair and a sofa), or it provides a wrong answer to a similar looking object where the function is different

(e.g., a glass and a vase). This problem can be solved by implementing function bias, which human children use to infer the name of objects.

This study uses the human body recognition feature provided by Kinect SDK to distinguish uses of an object. It enables overlaying the skeletal information of a human body on top of the 3D image reconstructed based on the depth map and the RGB image (Figure 2).

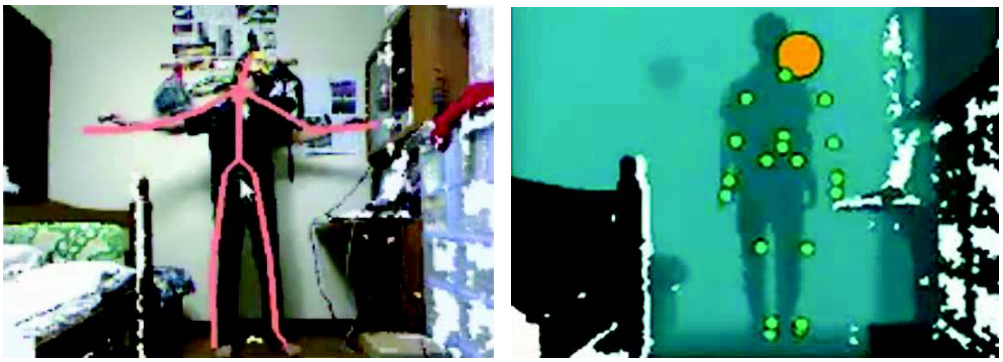


Figure 2. With the data retrieved by Kinect sensor, the program overlays a skeleton on the 3D image and tracks body joint.

In order to implement this kind of functional bias, one has to first recognize human action. By using the feature of Kinect to track 20 different joints of skeletal information, the program records coordinates for each joint every 0.1 seconds for 10 seconds while the teacher performs some activity. Then the system asks the teacher to name the activity. Instead of storing absolute distance of each joint from the Kinect sensor, by storing relative distance of each joint from the coordinate of the head position, the recorded data become independent of the direction the actor is facing when performing the activity. In the testing session, the program again uses the SVM on the stored data to infer the name of a target action.

In an online video demonstration, a teacher demonstrates walking, skipping, and running activities and provides labels for these (Eguchi, 2011). Later, during testing, the user goes through a sequence of walking, skipping or running and the system identifies each sequence of activities. This demonstration replicates the previous results (Schuldt, Lapteve, & Caputo, 2004) but uses the Kinect instead of a more expensive solution.

Object Recognition Model with Shape Bias and Function Bias

In order to use both biases for object recognition, it is necessary to train two different classifiers; one is based on the shape and the other is based on the function. In addition to the feature of activity recognition, users are now prompted to choose the name of the object associated with the action (Figure 3). For example, the action "drinking" can be associated with a cup and a glass.

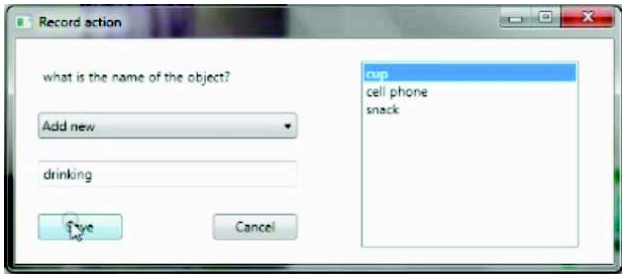


Figure 3. The program prompts the user to choose the name of the object corresponding to an action

Figure 4 shows the overall design of the program. In the program, teaching involves two processes: shape learning and function learning. A teacher shows and labels objects; then the teacher demonstrates the function of each object, names the function, and associates each function with the appropriate name of the used object. Testers can show any object to the program and demonstrate the use so that the program can try to recognize the target object.

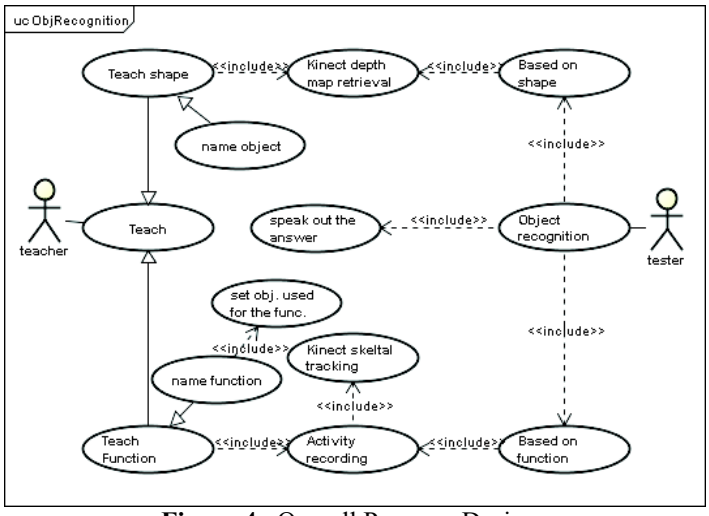


Figure 4. Overall Program Design.

Then, in the testing session, if the inferred name of the target object based on the shape and the function matches, the program tells the user the answer; on the other hand, if it does not match, the program tells the user the uncertainty of the answer by noting:

> *"Maybe the object is [answer based on the shape]. But the object may be a [answer based on the action] because you used the object for [the name of action]".*

If the shape is quite different from the one in the memory, the program will note:

> *"I think the object is [answer based on the action] because you used the object for [name of action]. But it might be a [answer based on the shape] based on the shape."*

An online demonstration is available for individuals (see Eguchi, n.d.).

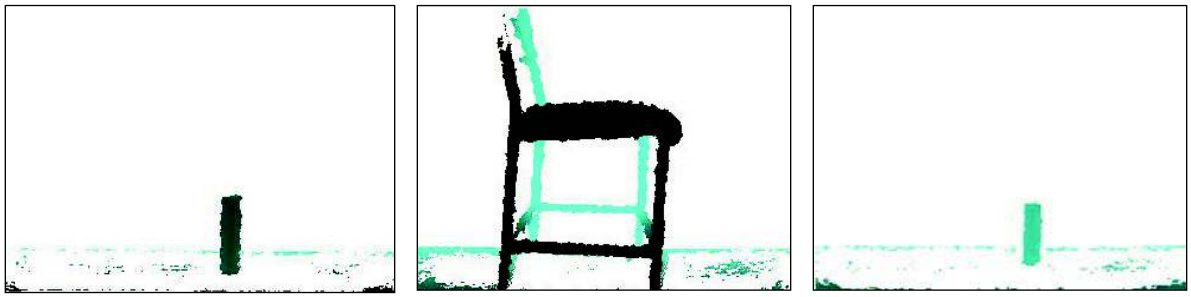
Methodology, Results and Analysis

Methodology

To test the accuracy of the model, the system was tested with two objects that look similar but have a different use (a can of antiperspirant and a can of insecticide), and two other objects that look different but have the same name and function (a conventional chair and an oddly shaped chair). The model was then tested based on three different bases. The first was object recognition based on a shape; the second used activity recognition based on the body joint movement; the third was an object recognition based on both the shape and the function. The tests worked most of the time; however, there were several constraints that need to be addressed.

Results

Object Recognition based on Shape. As indicated in Figures 5, the program identified the shape of an insecticide can, a chair, and an antiperspirant can.



Figures 5. These images illustrate the data used for the shape learning of insecticide, chair, and antiperspirant.

During testing, the program successfully recognized the chair all of the time, but it sometimes confused the insecticide and the antiperspirant because the shape is quite similar. Without training and when asked to name the differently shaped chair (Figure 6) and asked the name of the object, the program failed by noting that the object is antiperspirant because the program knows neither the shape of this type of object nor the use, which is to sit.

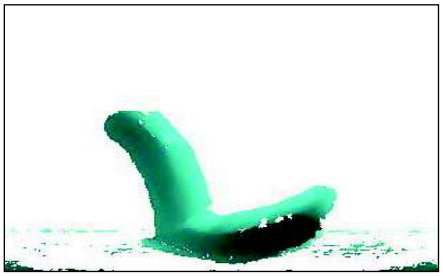


Figure 6. Different shape of a chair was also used in the learning trial.

Object Recognition based on Function. During activity learning, each action of sitting, killing bugs, and deodorizing is associated with a chair, a can of insecticide, and a can of antiperspirant respectively as shown in Figures 7-9.

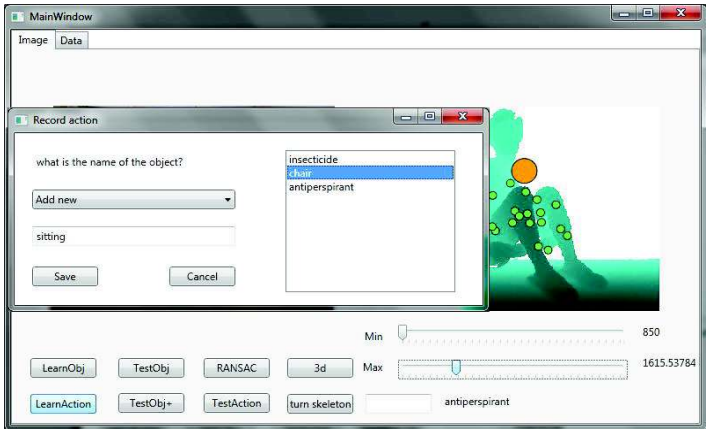


Figure 7. A teacher shows the program how to sit on a chair.

Figure 8. A teacher shows the program how to kill bugs with insecticide

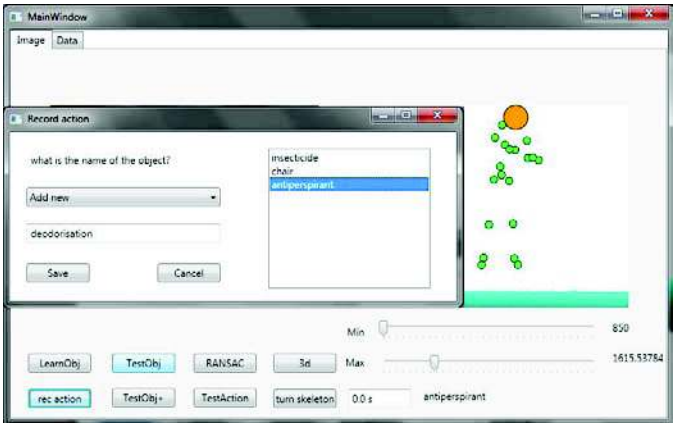


Figure 9. A teacher shows the program how to deodorize with an antiperspirant.

Then, in the testing session, the program successfully identified all of those actions. Additionally, even though the action "sitting" was learned with an unconventional shape of chair, the program successfully answered "sitting" for the sitting action on a different shape of a chair as shown in Figure 10.

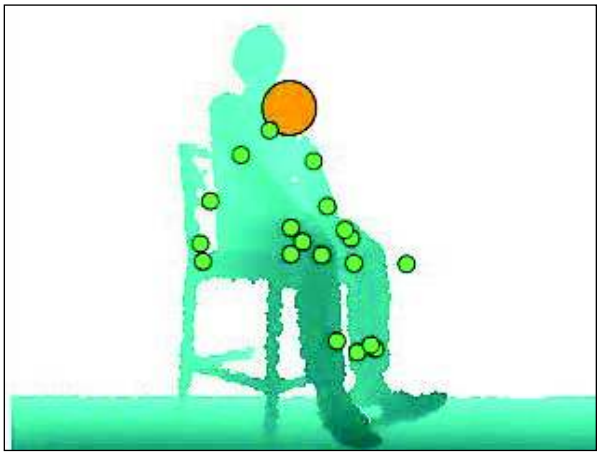


Figure 10. The program successfully generalizes the action recognition of sitting to a differently shaped chair.

Object Recognition based on both Shape and Function. In the final tests, the program uses both the shape and the function to infer the target objects. With the object recognition based only on the shape, an antiperspirant and an insecticide were sometimes confused; however, with this object recognition method, for the antiperspirant, it either correctly answered or stated:

"Maybe the object is insecticide. But the object may be antiperspirant because you used the object for deodorizing".

Also, for the insecticide, it either correctly answered or noted:

"Maybe the object is antiperspirant. But the object may be insecticide because you used the object for killing bugs".

Similarly, with the previous model, the different shape of chair cannot be properly recognized, but by showing the use of the object, sitting, the program answered:

"I think the object is a chair because you used the object for sitting. But it might be antiperspirant based on the shape."

Table 1 shows the summary of the results.

Table 1. *Results from two different object recognition models.*

	Insecticide	Antiperspirant	Learned chair	Novel chair
Shape bias	Confused with Antiperspirant	Confused with Insecticide	Correctly recognized	Completely failed
Shape bias and Function bias	Correctly inferred	Correctly inferred	Correctly recognized	Narrowed down the possibility

Analysis

Although the object recognition model worked as expected to properly identify the objects learned by shape or function, there were several constraints in the current model. First, because of the use of simple SVM on the 240 × 360 depth grid map, the object has to be set at the exact same position every time to be recognized as the same. This is similarly true for the activity learning. If the action associated with the object does not involve a lot of movement, like sitting, the program works well; however, if the action involves movement like walking, the shifting of timing can be a problem even though averaged position is also used as a key for the recognition. Figure 11 describes the problem of this model by the shift in the timing of input data.

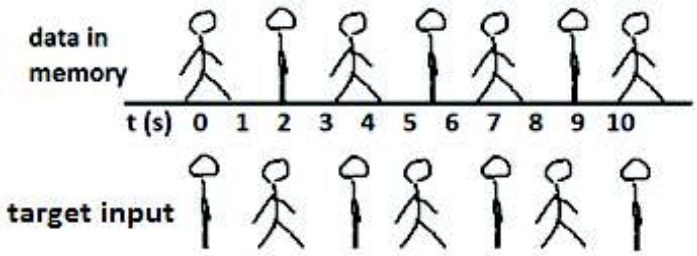


Figure 11. This image illustrates the case where the action is poorly learned.

Conclusion

Summary

This paper proposes a new way of designing a computational object recognition model by introducing knowledge from developmental psychology. Whereas most past object recognition models use the shape of objects for the recognition, the model discussed in this paper also uses the function of the target object for more accurate recognition. The powerful features of the Kinect sensor, like depth map retrieval and human body joint recognition, made the development of this proposing model easier to construct and also much less expensive. Results of testing demonstrate that the model works as expected; results also identified possible improvements for a future model.

Potential Impact

This research indicates that knowing the variety of shapes of common objects (like a chair) is important but knowing the functional use of the object also plays a role in object recognition. The study benefitted from its interdisciplinary use of results from psychology about how children learn to associate words with things that informed the computational model of object recognition.

If computers can learn to recognize everyday objects, many new applications will be enabled. If computers can monitor and recognize sequences of activities, many additional applications will be enabled. For only a few examples, pairing both recognition capabilities may help computers including robots or even other objects (a) recognize objects in a room or on an assembly line, or (b) identify activities helpful in driving a car or watching, recording and providing advice during heart operations.

Future Work

The machine learning technique used for this model can be improved by addressing the constraints of the current model or by replacing the current algorithm with a more advanced one. The name teaching part could be implemented with speech recognition technology so that the teacher does not have to manually type the words.

Improving the accuracy of object recognition would accelerate the idea of building a semantic world with smart objects. Instead of labeling objects with RFID tags as discussed in a previous paper (Eguchi, Nguyen, & Thompson, in press; Eguchi & Thompson, 2011), the object recognition technique can be used to identify any object that then identifies an associated Application Programming Interface (API) so that we can communicate with the object. The information about objects in a certain environment can be linked to the associated map, which can be autonomously constructed by an autonomous floor mapping robot (Nguyen, Eguchi, & Hooten, 2011).

The accuracy of the recognition can be improved by combining the idea with the study of ontology to narrow down the search space of objects that are most likely to exist in a certain environment, like a kitchen, where we are most likely find a sink, a refrigerator, cabinets, and plates (Eno & Thompson, 2011).

Also, the activity recognition feature can be improved by parsing or recognizing the logs of trace data observations and workflow rules to identify higher level named activities. For instance, we might be able to understand that someone is packing a truck (a higher level workflow) by observing a sequence of lower level workflow like, <go to object, pick up, move object into truck> triples. In order to accomplish this work, the ideas from formal language (grammars, terminals, rules) might be useful to recognize real world activities from trace observations.

References

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. Pattern analysis and machine intelligence, *IEEE Transactions, 24*(4), 509–522.

Eguchi, A. (2012). *Cultural bias during word learning*. (Unpublished bachelor's thesis). University of Arkansas, Fayetteville, AR.

Eguchi, A. (2011, July 23). Activity learning using Kinect skeletal data. [Video podcast]. Retrieved from http://www.youtube.com/watch?v=AxCn0eKWkiQ

Eguchi, A. (2011, October 31). *Object recognition with shape/function bias*. [Video podcast]. Retrieved from http://www.youtube.com/watch?v=4ia76fzxm68

Eguchi, A. (2011). *Object recognition based on shape and function*. (Unpublished bachelor's thesis). University of Arkansas, Fayetteville, AR.

Eguchi, A., & Nguyen, H. (2011, September). *Minority game: The battle of adaptation, intelligence, cooperation and power*. Paper presented at the 2011 Institute of Electrical and Electronic Engineers International Conference, Szczecin, Poland.

Eguchi, A., Nguyen, H., & Thompson, C. W. (in press). Everything is alive: Towards the future wisdom web of things. *World Wide Web*.

Eguchi, A., & Thompson, C. W. (2011). Towards a semantic world: Smart objects in a virtual world. *International Journal of Computer Information Systems and Industrial Management, 3*, 905-911.

Eimas, P. D., & Quinn, P. C. (1994). Studies on the formation of perceptually based basic-level categories in young infants. *Child Development, 65*(3), 903–917.

Eno, J. D., & Thompson, C. W. (2011). Virtual and real-world ontology services. *IEEE Internet Computing, 15*(5), 46–52.

Girshick, R., Shotton, J., Kohli, P., Criminisi, A., & Fitzgibbon, A. (2011, November). *Efficient regression of general-activity human poses from depth image*. Paper presented at the Institute of Electrical and Electronic Engineers International Conference, Barcelona, Spain.

Grabner, H., Gall, J., & Van Gool, L. (2011, June). *What makes a chair a chair?* Paper presented at the Institute of Electrical and Electronic Engineers International Conference, Colorado Springs, Colorado.

Lai, K., & Fox, D. (2010). Object recognition in 3D point clouds using web data and domain adaptation. *International Journal of Robotics Research, 29*(8), 1019–1037.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*(3), 299–321.

Nelson, D. G. K., Russell, R., Duke, N., & Jones, K. (2000). Two-year-olds will name artifacts by their functions. *Child Development, 71*(5), 1271–1288.

Nguyen, H., Eguchi, A., & Hooten, D. (2011, September). *In search of a cost effective way to develop autonomous floor mapping robots.* Paper presented at the Institute of Electrical and Electronic Engineers International Conference*,* Montreal, Canada.

Philipose, M., Fishkin, K. P., Perkowitz, M., Patterson, D. J., Fox, D., Kautz, H., & Hahnel, D. (2004). Inferring activities from interactions with objects. *IEEE Pervasive Computing, 3*(4), 50–57.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926-1928.

Schuldt, C., Laptev, I., & Caputo, B. (2004, August). *Recognizing human actions: A local SVM approach.* Paper presented at the Institute of Electrical and Electronic Engineers International Conference*,* Cambridge, England.

Toshev, A., Makadia, A., & Daniilidis, K. (2009, June). *Shape-based object recognition in videos using 3D synthetic object models.* Paper presented at the Institute of Electrical and Electronic Engineers International Conference*,* Miami Beach, Florida.