Prototyping Interaction with Everyday Artifacts: Training and Recognizing 3D Objects via Kinects

Kexi Liu¹, Dimosthenis Kaleas¹, Roger Ruuspakka¹
Qualcomm Inc¹; cathyl@qualcomm.com

ABSTRACT
In this paper we explore and prototype the interaction with everyday passive objects. We present an approach for 3D object training and recognition, leveraging Kinect sensors: the Dominant Orientation Templates (DOT) method allows real-time object training and multiple Kinects speed up the training process by learning the object from multiple viewpoints simultaneously with the object background removed. A proof-of-concept usage scenario, 3D real-time Lego building instruction, has been developed based on this approach: the system learns the individual Lego pieces and Lego building steps in advance; the users thus construct the Lego model with 3D visual hints attached to present Lego pieces.

Author Keywords
Dominant Orientation Templates; 3D object training and recognition; Kinect; TUI; interaction design; interactive Lego building instructions

ACM Classification Keywords
H5.2. Information interfaces and presentation: User Interfaces.

General Terms
Design, Experimentation

INTRODUCTION
In order to provide new services beyond the objects’ traditional purpose, people implant sensing, computing and actuation into objects. However, traditional everyday objects, with a pleasant visual design and simple functionality, still remain the majority of the objects around, and are sometimes preferred over their advanced versions. Part of the reasons are that attaching tags/sensors on physical objects takes up object surface real estate, changes its outer appearance, and requires users’ extra care. In this sense, relying on special artifacts limits the widespread use of the services and functionalities, as they are not suitable for many everyday situations in which we interact with arbitrary artifacts that are not part of the interface.

While keeping the original form factor of everyday objects, users wish to occasionally bridge the objects physical form to their digital counterparts, so that richer functionalities could be provided. A regular apartment mailbox key might give you the number of mails in your mailbox when you inquire; a “memory” cup displays the photos of your loved ones at times; a living room wall clock shows you today’s schedule when you check.

To be able to interact with conventional objects in an unconventional way, users need novel and appropriate interface which do not require an intensive learning phase [18], considering that elderly and children are notably interesting user groups that have the potential to profit from such an interface paradigm. We propose a template-based method to learn and recognize everyday 3D objects, leveraging Kinect IR sensors. The training and recognition is real-time, robust to small deformation and translation and workable with poor-textured objects. The Kinect IR sensors can further enable the object training without background, and learn the object from multiple viewpoints simultaneously. This way the object’s outer appearances, which can have an important impact on the overall perception of an everyday object, remain unchanged. The recognition can connect the physical object with related digital information naturally.

The ability to detect 3D objects in the user’s vicinity offers a natural way of interaction and strengthens the role of the webcam in a large number of application scenarios. In this paper, we developed a proof-of-concept example application: 3D Lego building instructions. Lego building has been joyful and inspiring experience for kids and enthusiasts. With 3D object recognition method, the system learns the individual Lego pieces and Lego building steps in advance, and when the users are constructing Lego models, 3D visual instruction hints show up at the upper right corner of each Lego piece. If the users snap the two Lego pieces correctly, the system will recognize it as a new Lego piece. Through this example scenario, we discuss the limitations
of the current method and possible future research directions.

RELATED WORK
Technology-augmented artifacts are used for the binding of physical representations with the underlying digital information. Spime builder [4] is a computer vision system that allows users to model three-dimensional objects and environments in real-time using physical materials and to place hyperlinks in specific areas using laser pointer gestures. PaperView [9] use augmented reality for supplementing physical surfaces with digital information through the use of pieces of plain paper that act as personal, location-aware, interactive screens. MADO [17] constructs a physical object by combining electrical blocks. Then, by connecting MADO Interface to the physical object, they can watch the virtual model corresponding to the physical block configuration. DisplayObjects [1] is a rapid prototyping workbench that allows functional interfaces to be projected onto real 3D physical prototypes. However, DisplayObjects uses a Vicon motion capture system that is very expensive. Easigami [13] presents a reconfigurable system of thin flat polygon pieces connected by electronically instrumented hinges. The circuit board embedded in the paper presented aesthetic design challenges. SurfaceWare [6] examines a class of dynamically tagged passive objects; however, it still requires a change to the object design. Furthermore, research has been done for building instructions by augmented objects [2,3].

3D object recognition has been a very challenging problem. With the object to be recognized presented to a vision system in a controlled environment, the system then locates the previously presented object. This can be done either off-line or in real-time. Template Matching is attractive for object detection because of its simplicity and its capability to handle different types of objects. Dominant Orientation Template (DOT) method [10, 11, 12] neither needs a large training set nor a time-consuming training stage, and can handle low-textured objects, which are difficult for feature points-based methods. The location and pose of the object can also be detected through DOT. We leverage DOT in this paper and consider it has the strength of the similarity measure of [11], the robustness of [5] and the online learning capability of [7, 8]. In addition, by using branch-and-bound method of [16], the method becomes very fast for real-time detection of untextured 3D objects [12].

With the birth of the Kinect style cameras, a new generation of sensing technologies capable of providing high quality synchronized videos of both color and depth has emerged. This technology represents an opportunity to dramatically increase object recognition, manipulation, navigation, and interaction capabilities. Lai et al. introduces a large-scale, hierarchical multi-view object dataset [14] collected using an RGB-D camera, with the description of techniques for RGB-D based on object recognition and detection. Lai et al. also studied both object category and instance recognition using the RGB-D Object Dataset [15], a large RGB-D (color + depth) dataset of everyday objects.

OBJECT RECOGNITION WITH KINECTS

DOT Object Recognition
Our method is based on Dominant Orientation Templates method [12]: The Energy function counts how many orientations are similar between the image and the template. Magnitudes of template gradients are only considered to retain the orientations of the strongest gradients, without using their actual values for matching. Therefore it is robust to illumination change and noise. The template representation is also robust to small image deformation and translation: template calculation returns the set of orientations of the strongest gradients. By contrast, the input image returns only one orientation, the orientation of the strongest gradient. Also, when the object is slightly moved, a series of transformation matrices will be used for the comparison, which returns the maximal value of energy function. Bitwise operation, SSE instructions, which allows performing the same operation on 16 bytes in parallel, and Branch & Bound clustering for the similar templates, can further speed up the calculation. Therefore, DOT 3D object recognition does not require a time consuming training stage, handle poor textured objects in complex situations and provide their 3D pose in real-time.

Kinects with DOT Object Recognition
With the original DOT object recognition method, the object training stage has to be on a uniformed background; however, this restriction can be removed with the use of Kinect sensors. Kinect depth sensors consist of an infrared laser projector and a monochrome CMOS sensor, which captures video data in 3D under any ambient light conditions. Furthermore, the Kinect IR sensor, combined with the adjacent CMOS VGA camera, is capable of providing high quality synchronized videos of both depth and color. With this feature, the background pixels can be removed if their corresponding depth value is greater than a specified threshold. When the object is held by a hand, the static depth threshold could be set and the distance between the hand and the camera should be equal to that threshold, so that everything behind the hand (including the hand itself) is treated as background and only the pixels for the object remain.

Concerning setup automation, we place the object on a turning table, so that the object can be learnt from different view angles. To remove the background plane, three points in the scene are picked to identify the plane. The plane equation in 3D will be calculated, based on the coordinates. For each pixel, if its depth value satisfies the equation, then it will be removed, so that only the pixels of the object are remained in the scene. By removing the object background, the shadow of the object is removed as well, which makes
the learning more precise. From there, we apply DOT method to get the templates.

Although the original DOT method is very fast with binary representation and Branch-and-Bound clustering, the object learning gets even faster with multiple Kinects. In this paper, two Kinects are used to point to the object from different viewpoints. As the object rotates on the turning table, the templates captured respectively by the two Kinects are placed in the same class, so that the learning of the overall templates is quicker, as shown in Figure 1.

**SYSTEM SETUP**

We run our DOT 3D object recognition on a standard PC (Lenovo T410), with 2 Kinects, and a small turning table, as shown in Figure 2. The training and recognition with DOT method goes like this: the target object is presented on top of a turning table, and the Kinects are mounted above (45 degrees), facing each other and both pointing to the target object. On the video streaming of the object scene, first the user picks up the target object (region) by mouse clicking on it; then the dominant gradients of that object are extracted and saved as a template. As the object rotates, Kinects collect the templates of that object from multiple viewpoints simultaneously and save them in the object class, as shown in Figure 3. With a keystroke, the system switches from learning stage to recognition stage, and only one RGB webcam is needed for the recognition stage. Black bounding boxes pop up to show possible object location and pose.

**USAGE SCENARIO: LEGO BUILDING INSTRUCTIONS**

For the current system, the templates of the learnt objects will be saved in the database and reloaded. Those objects could be recognized by RGB cameras as natural ID to identify objects and thus be used for broad applications, such as accessing online information, etc. Object assemblies, e.g., furniture assembly or LEGO game could be challenging problems for newbies. The traditional assembly instructions are still paper printed, which usually only show one view angle of the object parts and the way they are assembled.

**LIMITATIONS AND FUTURE WORK**

Several limitations of the current implementation are: when one Lego piece is significantly bigger than the other piece, the system will mistakenly treat the stacked piece as the previous big single piece, which means if we build a large Lego model, the system will be problematic towards the end; the system can’t identify the Lego piece when it’s occluded by hand. We will address these problems in future development. For the current development, we augmented a simple 3D representation on the recognized object. Now we are developing a 3D augmented reality to represent the object so that the users could view the instructions in the
3D paradigm. Augmented Reality QCAR SDK can possibly be used for the implementation. We are also working on recognizing the genres of objects from depth information.

Figure 4. Lego building instructions based on DOT object recognition

CONCLUSION

DOT 3D real-time object training and recognition with Kinects is an on-going project that seeks an intuitive way to identify everyday passive objects and play with those interactive objects. The traditional way to identify the object is through static or dynamic tagging, e.g., RF tags, or visual tags, which brings up aesthetic design challenges. This system, stemmed from template based image matching, trains everyday passive objects (excluding transparent objects) in real-time, recognizes the objects and their poses despite small deformation and translation, and handles poor-textured objects. The proof-of-concept usage scenario, 3D Lego building instruction, is one practical example to show how users can playfully interact with the identified objects. Our vision for the near future is to realize the object recognition on portable devices and integrate it with augmented reality technology to bring better visual effects and interaction experience.

REFERENCES