Learning to Identify and Track Imaginary Objects Implied by Gestures

Andreya Piplica and **Alexandra Olivier** and **Allison Petrosino** and **Kevin Gold**
Department of Computer Science
Wellesley College
Wellesley, MA 02481

Abstract

A vision-based machine learner is presented that learns characteristic hand and object movement patterns for using certain objects, and uses this information to recreate the "imagined" object when the gesture is performed without the object. To classify the gestures/objects, Hidden Markov Models (HMMs) are trained on the moment-to-moment velocity and shape of the object-manipulating hand. Object identification using the Forward-Backward algorithm achieved 89% identification accuracy when deciding between 6 objects. Two methods for rotating and positioning imaginary objects in the frame were compared. One used a modified HMM to smooth the observed rotation of the hand, with mixtures of Von Mises distributions. The other used least squares regression to determine the object rotation as a function of hand location, and provided more accurate rotational positioning. The method was adapted to real-time classification from a low-fps webcam stream and still succeeds when the testing frame rate is much lower than training.

1. Introduction

Social pretend play facilitates mutual understanding of imagined or hypothetical situations by allowing psuedo-tangible interaction with objects that do not exist. In children, pretend play is not observed until advanced cognitive faculties, such as the ability to ascribe two potentially conflicting sets of properties to the same object, develop (Lillard 1993; Howes 1985; Elder and Pederson 1985). Incorporating others into pretend play is an even more difficult task. Children do not typically begin to engage in social pretend play until after they have mastered both complex social play and solitary pretend play (Howes 1985). In social pretend play, especially pretense without objects, bodily activity facilitates interaction between participants (Lillard 1993). Although pretend play is a complex human behavior, a machine learner can implement some of the basic skills needed to accomplish such a task.

A robotic learner that can recognize pantomimed actions may be able to imitate such actions and interact with the suggested objects (Chella, Dindo, and Infantino 2006). A visual representation of the object would make it easier for the learner to understand where someone was pretending an object to be. A robot that learned through observation could be taught a variety of actions in a short amount of time. These traits make such a learner easy to incorporate in a robotic toy that can engage in a child's pretend play or a video game based on pantomimed actions. Additionally, a robotic learner that classifies objects not by their physical properties but by how they are used may have an advantage over other systems when learning object affordances (Montesano et al. 2008).

A vision-based machine learner is presented that can exhibit some of the most basic features of pretend play, namely, using learned properties of objects and actions to reason about where someone is pretending an object is. It learns characteristic hand and object movement patterns for using certain objects, then uses this information to recreate the "imagined" object when the gesture is performed without the object. This research is a new hybridization of gesture recognition using Hidden Markov Models (HMMs) (e.g., (Lee and Kim 1999)), and augmented reality, which tracks imaginary objects but typically assumes rather than decides what object is being manipulated (Azuma 1997).

This machine learner uses HMMs to learn several actions by observing gesture patterns in videos of the action performed with an object. HMMs are common tools for gesture recognition because they rely on probabilistic rather than deterministic reasoning and because of their ability to make predictions in real time (Starner and Pentland 1997). It also learns how the object is positioned and rotated with respect to the hand. When shown video of one of the actions being performed without an object, the learner will choose which HMM most likely describes that action and fill in an image of the imagined object. The accuracy of the learner's action classifications of recorded was tested. In addition, the learner's real-time classification abilities were observed. Two different approaches to rotating the imagined object image, one based on least squares regression and the other based on the von Mises distribution, were compared to determine which provided more accurate rotation. Correct positioning and rotation, which are part of the problem of registration in augmented reality systems (Azuma 1997), are necessary for realistic interaction with imagined objects.

By making use of HMMs, this work was adaptable to real-time recognition; the present work introduces experiments testing whether HMMs trained offline with videos could per-

form the same recognition in real-time with a much-reduced frame rate.

Related Works

Hidden Markov Models have been used since 1992 for human action recognition (Yamato, Ohya, and Ishii 1992). Variations on Hidden Markov Models are commonly used to tackle specific challenges in gesture recognition. Parameterized HMMs, for example, recognize gestures that can vary in parameters like direction and size (Wilson and Bobick 1999). Coupled HMMs can model actions with several underlying processes that interact at the same time but may have different state structures (Brand, Oliver, and Pentland 1997). Unmodified HMMs (Starner and Pentland 1997) and desk-based monocular cameras have been used for recognizing sign language (Starner, Weaver, and Pentland 1998). Color-based skin segmentation, especially with Kalman filters for hand tracking, are effective for isolating hands in video sequences even if the hands are in front of the face (Imagawa, Lu, and Igi 1998).

Head mounted displays have been used to allow people to interact with virtual spaces in real time (Butz et al. 1999), but no such work involves robots interacting with virtual objects. Typically, robots learn about object affordances through direct interaction (Sinapov and Stoytchev 2007), though some work has involved humans interacting with virtual objects to teach robots how to interact with real world objects (Bentivegna and Atkeson 2000).

2. Methods

The machine learner must perform several sub-tasks to accomplish the overall goal of identifying pretend actions and filling in pretend objects. First, for each action, an HMM is trained from a video of a person performing that action with an appropriate object. The active hand must be isolated in each frame of the video so information about the discrete state of the hand can be used in an HMM. Once trained, the HMMs are used to identify an action from either a recorded video or from a real-time image stream. A binary image of the imagined object is placed in the recorded frames. Methods based on least squares regression and the von Mises distribution are compared to see which provides a more accurate orientation of the object.

Isolating the Active Hand

In order to recognize the actions studied here, an image of the active hand (the hand in direct contact with the object) must be isolated so features about its shape and position can be extracted. RGB images taken from the camera are convolved with a sharpening filter. The sharpened images are converted to Y'UV color space to perform color segmentation based on skin color. Color spaces that account for both luminance and chrominance such as Y'UV have a high rate of accurate classification of skin color. In addition, Y'UV color space is robust to many shades of skin, both dark and light (Kakumanu, Makrogiannis, and Bourbakis 2007). For this experiment, skin colors are found in the range $Y' < 0.8$, $-0.2 < U < 0$, $V > 0$, which covers bright, mostly pink and red colors (Figure 1).

After the color thresholds create a binary image of the skin colored segments, the image is dilated and eroded to create contiguous segments. Of these skin segments, the active hand and the face tend to be the largest two segments. The face is assumed not to move in the video, so once it is found in the first frame, the skin segments in that region can be ignored in subsequent frames. Actions were performed with the active hand starting to the lower left of the face, though not always in the same position. Thus the two largest skin segments in the first frame were compared, and the segment higher and further right in the frame was determined to be the face. After the first frame, the facial skin segments are blacked out, leaving the hand the largest skin segment. Properties about the active hand were extracted from this largest segment.

Defining the Actions

HMMs use discrete states to probabilistically describe an action over time (Baum et al. 2007). Here, the shape and motion of the hand determine the discrete states. Each state has three features – the hand shape (either open or closed), the hand's vertical motion between frames, and the hand's horizontal motion between frames. Motion is classified as either positive, negative, or still. These eighteen discrete states define the transition and emission matrices that make up the model for each action. The HMMs were trained using the Baum-Welch algorithm (Baum et al. 2007) to perform expectation-maximization (Dempster, Laird, and Rubin 1977). One HMM was trained from each of the eighteen training videos.

In order to decide which of six possible actions is occurring at a given time, the Forward-Backward algorithm determines which of the six models is most likely to describe the actions leading up to the current time. The actions explored here are brushing teeth, drinking from a cup, hitting something with a hammer, petting a stuffed dog, scooping with a shovel, and writing with a marker. The likelihoods for all eighteen models were propagated forward. At each time step, the average likelihood for each action was computed from the likelihoods of the three models for that action. Like HMMs, the Forward-Backward algorithm can update in constant time (Starner and Pentland 1997), making it useful for real time applications.

Placing an Object Image

The final task for the pretending machine learner is to place an image of an object in each video frame. The image should be placed in the space where the performer is pretending there is an object, and it should be positioned and rotated realistically. To do this, the machine learner must learn how the object should be positioned and rotated with respect to the hand's position and rotation. Positioning is learned by observing videos of an action performed with an object. Least squares regression finds a function mapping hand centroid position to the displacement of the object centroid from the hand centroid.

Figure 1: Stages of the skin segmentation process for one video frame (1). Skin colored regions are detected with a filter (2). The face region is blocked out, leaving the hand as the largest skin colored region (3).

Determining correct object rotation is not as simple as finding the rotation of the hand and rotating the object to the same degree. Hand rotation measurements based on the orientation of the hand's major axis are often noisy, especially when different light highlights on the hand can obscure its true shape in a skin filter. In preliminary testing, hand angles were often interpreted as offset by 90 degrees from their true angle. A mixture of two von Mises distributions, a variant of the normal distribution for use in rotational coordinates (Bishop 2006), was fit to hand rotation data collected under known rotations to model these discrepancies. It was expected that the distribution would have two peaks when modeling actions with a consistent angle of rotation, one at the correct angle and another at the 90 degree offset, because the rotation reading might occasionally jump 90 degrees when the segmented hand curled into a fist was close to square. For actions with varying rotation over time, the distribution was expected to have peaks at the most common angles and at their 90 degree offsets. A modified Kalman-like filter over time was used to smooth the hand rotation data and provide a more accurate estimate of actual rotation. The transitional model for this dynamic Bayesian model was trained on video of a hand rotating over time; a von Mises distribution for the rotational change from one moment to the next was fit to this data to obtain a transitional model that could smooth the frame-to-frame readings of the hand rotation. The observation model, the aforementioned mixture of two Von Mises distributions, was then fit to recordings of the hand under known rotations. Functions for the von Mises distribution were obtained through a publicly available circular statistics toolbox (Berens and Velasco 2009).

When this smoothing over time was still not enough to provide consistent rotational readings (see experiment), a different approach was tried. For the set of actions studied here, it was hypothesized that object rotation could be inferred from the hand's centroid position rather than from its angle of rotation, which changed slightly but consistently with each rotation. It was hypothesized that this change over time would be less susceptible to skin segmentation noise, because while finding the rotation of a major ellipses of a color blob can be highly susceptible to noise and inconsistencies at the edges, the centroid is an average of many pixels of data, which tends to wash out errors. Pretend motions that suggest an action or object are often stereotyped and repetitive (Acredolo and Goodwyn 1988), so object rotations are likely to follow a consistent pattern as the hand cycles through the stages of the motion. The rotation pattern can then be generalized by using least squares regression to find a mapping from hand centroid position to the angle of object rotation.

3. Experiments

Training

Using a Logitech Quickcam Orbit AF grabbing 640×480 pixels at 30 fps, three people each recorded six twenty second videos, which were used to train the HMMs. Participants performed the following actions while holding an object appropriate to the action: drinking from a cup, petting a stuffed dog, swinging a hammer, writing with a marker, scooping with a shovel, and brushing teeth with a toothbrush. In each frame of the training videos, the active hand was isolated using the skin segmentation algorithms. An HMM was trained for each action based on the discretized videos.

In addition to training the HMMs, the videos with objects were used to gather information about how each object should be positioned and rotated with respect to the hand. Least squares regression and the von Mises distribution provided two possible approaches to object rotation, the first based on hand position and the second based on hand rotation.

Offline Testing

The three participants performed the same six actions for twenty seconds without the accompanying objects. The forward-backward algorithm was used to calculate the likelihood of each HMM model. The recorded videos were then used as the basis for creating two separate videos with the imaginary object filled in – one using the von Mises smoothing, and another using least squares regression, as described above.

In order to judge the comparative accuracy of the least squares and von Mises rotation methods, the recorded objectless videos were filled in with the image of the correct object for that video. Two new sets of videos were made, one using each rotation method. An independent coder judged whether the least squares rotation or von Mises rotation looked more accurate given the hand's orientation in

a random sample of 40 frames from each of the eighteen videos.

Online testing

Though the forward-backward algorithm can in theory propagate the likelihood of each model forward in time in constant time as images are being recorded from the camera, in practice a reduced frame rate resulted from attempting to perform these calculations as the camera was running while still providing a real-time update. Each model was therefore tested in real-time to determine whether the frame dropping required to provide real-time feedback (producing a practicable frame rate of roughly 4 fps) would affect recognition.

4. Results

HMM Classification

In the eighteen videos with imagined objects, the system chose the correct action sixteen times, yielding an 89% correct classification rate. The two mistaken classifications both misclassified an action as scooping; the true actions were brushing and petting.

The likelihood of each action model at each frame was tracked when videos were recorded in real-time (Figure 2) . By the end of 80 frames, the correct action was identified in each video. For some actions, such as hammering and shoveling, the correct model achieved a clearly higher likelihood than the others in less than 20 frames. For other actions, such as tooth brushing, it took longer, about 50 frames, for the correct action model to become apparent. In all the videos except one, the correct action became apparent after 50 frames or less. In the writing video, however, writing was the most likely action at the end of 80 frames, but in the majority of frames, shoveling was more likely. The two had very similar likelihoods throughout the recording.

Rotation Method Comparison

An independent coder judged that the least squares method provided more accurate rotations than the von Mises method in 464 out of 720 random frames (Figure 3). These results indicate that least squares provides statistically more accurate rotation (p = 0.001). However, least squares did not always provide more accurate rotations than the von Mises method. For the stuffed dog, von Mises was judged more accurate in 97.5% of frames. Von Mises was also judged more accurate in more frames for the marker, but this disparity is well within the realm of chance ($p > 0.1$). Least squares rotation was used to reproduce the videos from each participant with the imagined objects filled in. (Figure 4)

5. Discussion

An HMM-based machine learner provides accurate classification of actions performed without objects based on training of the same actions performed with objects. Of the two proposed approaches to rotating the image of the imagined object, the least squares method is preferable to the von Mises based smoothing. The instance of positioning the stuffed dog was the only object for which the von Mises

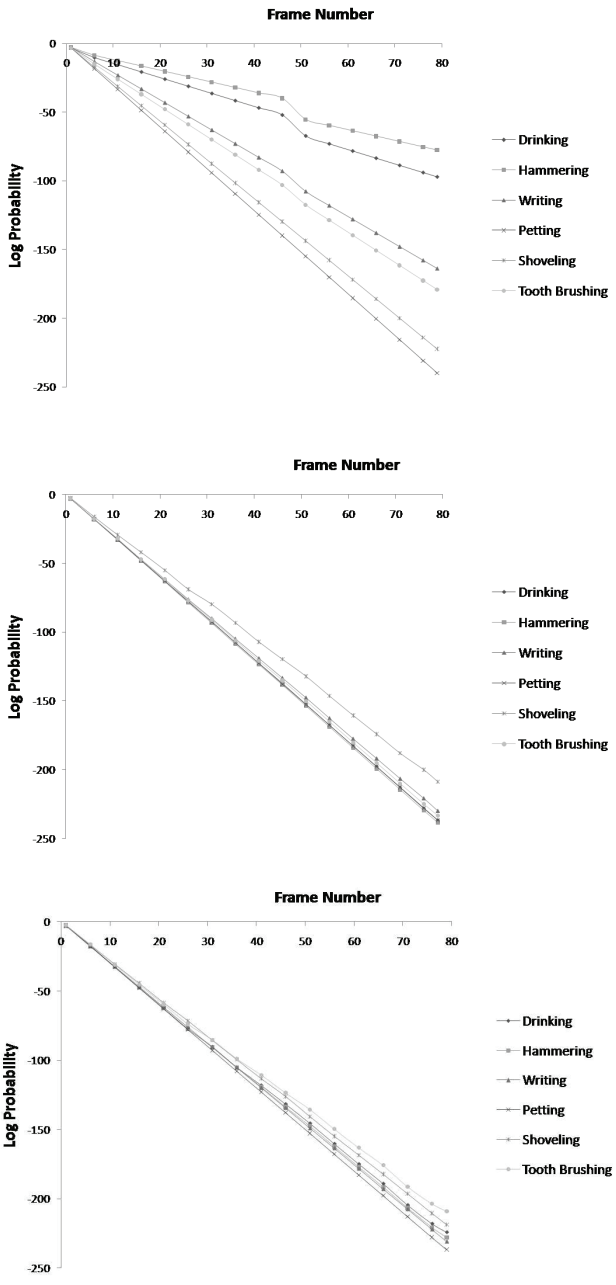

Figure 2: Likelihood estimates over time for real-time videos of hammering (top), shoveling (middle), and brushing (bottom) with imagined objects.

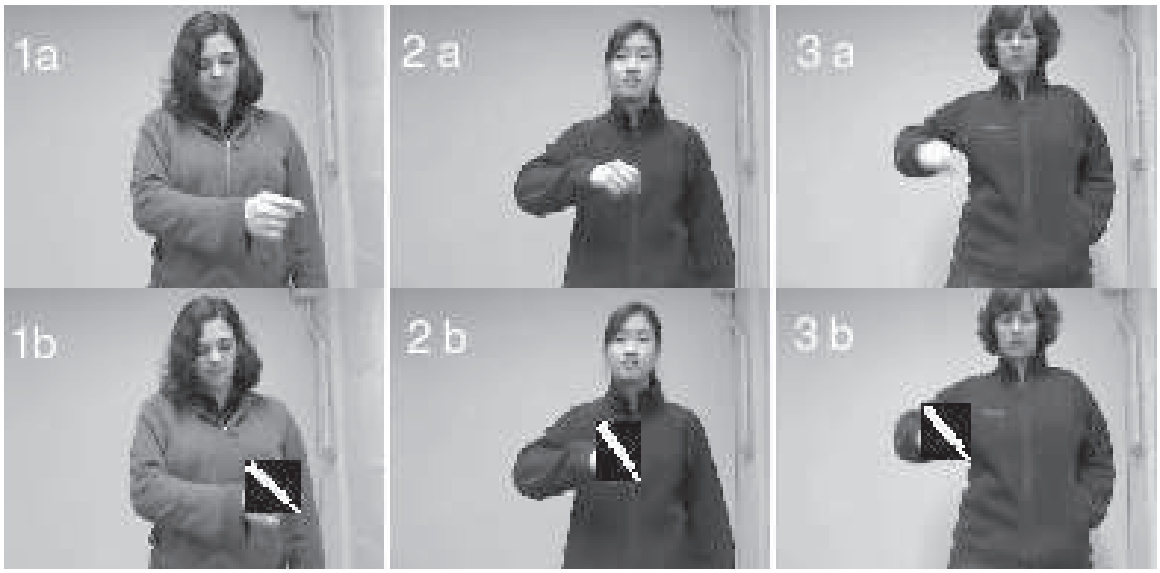

Figure 4: Imagined marker filled in using least squares regression. Analysis of object and hand positioning and rotation data from one participant generalized to allow accurate filling in for three participants.

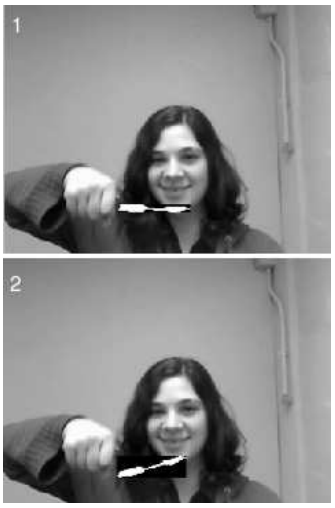

Figure 3: Least squares regression and von Mises distribution approximations for imagined object rotation.

Object	Least Squares	Von Mises
Cup	119	1
Dog	3	117
Marker	57	63
Hammer	83	37
Shovel	97	23
Toothbrush	105	15
Method Total	**464**	**256**

Table 1: Independent coder assessments of least squares and von Mises rotation accuracy.

distribution provided significantly more accurate rotation. The tracked hand in this case was the hand holding the dog, which remained still throughout the video. The von Mises distribution may have provided more accurate results because the hand tended to maintain a constant angle of orientation, so the smoothing could be more effective. The least squares regression primarily worked because the motions in question were repetitive and stereotyped. These attributes, however, are common to young children's social pretend play.

Approaches that attempt to derive the properties of missing objects from the bottom up may not perform as well as approaches that account for repetition, stereotyped behavior, and context. The von Mises approach, which performed poorly, attempted to smooth the rotational data obtained directly from the visual input. In contrast, the least-squares method took advantage of the limited number of possible object positions during these repetitive motions. The observed shift in children's pretend play in which they move from requiring a placeholder object to being able to pretend without any placeholder (Elder and Pederson 1985) may therefore be a function of moving from bottom-up, perceptually driven cognition, where the placeholder object can fill in some of the missing perceptual details, to top-down thinking driven largely by contextual expectations.

The real-time experiment demonstrates the relative robustness of this approach in situations where training conditions are significantly different from testing conditions. The training data for the recognition algorithm contained objects, but all testing was done without them. HMM model likelihood comparison is an attractive approach when testing must occur under different conditions than training; the correct model does not need to see the same observations it was trained with to make accurate classifications, so long as it

fits better than the other models. This approach is also robust against large changes in frame rate. Training was done with videos captured at a high frame rate, but testing was done with real time images captured at a lower frame rate. The velocity of the hand from frame to frame plays only a minor role in the model decisions. That is, the HMM transition probabilities would be different under different frame rates, but models trained with high frame rates still perform reasonably well under much lower frame rates.

Several improvements can be made to this system to make it more robust to a variety of situations and actions. Several assumptions were made about the nature of the environment and the actions performed. The environment was controlled so that only one person was visible to the camera at a time. Actions were designed so that a performer only had one hand moving at a time and did not need to move any other body parts besides arm and hand. A more robust learner could possibly learn more physically involved actions and ignore potential background actions. As the learner learns more actions, it may become harder for it to distinguish between actions with similar motion patterns. Parallel HMMs may be used instead to model complex actions with both arms moving at the smae time and improve recognition robustness in even small state spaces (Vogler and Metaxas 1999). The ultimate goal for this learner is to build a robot arm that can interact with an imaginary object in the pretended space. A robotic learner, especially one that may be used to develop toys or video games, should be able to compensate for these variations in action.

Acknowledgments

We would like to thank the Wellesley College Dean of the College and Norma Wilentz Hess for funding. We would also like to thank all video participants and our independent coder.

References

Acredolo, L., and Goodwyn, S. 1988. Symbolic Gesturing in Normal Infants. *Child Development* 59(2):450–466.

Azuma, R. T. 1997. A Survey of Augmented Reality. *Presence: Teleoperators and Virtual Environments* 6(4):355–385.

Baum, L. E.; Petrie, T.; Soules, G.; and Weiss, N. 2007. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov chains. *Pattern Recognition* 40:1106–1122.

Bentivegna, D., and Atkeson, C. G. 2000. Using primitives in learning from observation. In *First IEEE-RAS International Conference on Humanoid Robots (Humanoids)*.

Berens, P., and Velasco, M. 2009. CircStat2009The Circular Statistics Toolbox from MATLAB. Technical Report 184, MPI.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.

Brand, M.; Oliver, N.; and Pentland, A. 1997. Coupled Hidden Markov Models for complex action recognition. In *CVPR '97*. IEEE.

Butz, A.; Hollerer, T.; Feiner, S.; and MacIntyre, B. 1999. Enveloping users and computers in a collaborative 3d augmented reality. In *IWAR '99*, 35–44. IEEE.

Chella, A.; Dindo, H.; and Infantino, I. 2006. A cognitive framework for imitation learning. *Robotics and Autonomous Systems* 54:403–408.

Dempster, A.; Laird, N.; and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Methodological* 39(1):1–38.

Elder, J. L., and Pederson, D. R. 1985. Preschool Children's Use of Objects in Symbolic Play. *Child Development* 56(2):1253–1258.

Howes, C. 1985. Sharing Fantasy: Social Pretend Play in Toddlers. *Child Development* 56(5):1253–1258.

Imagawa, K.; Lu, S.; and Igi, S. 1998. Color-based hands tracking system for sign language recognition. In *FG '98: Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, 462. Washington, DC, USA: IEEE Computer Society.

Kakumanu, P.; Makrogiannis, S.; and Bourbakis, N. 2007. A Survey of Skin-Color Modeling and Detection Methods. *Pattern Recognition* 40:1106–1122.

Lee, H.-K., and Kim, J. H. 1999. An HMM-Based Threshold Model Approach for Gesture Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21(10):961–973.

Lillard, A. S. 1993. Pretend Play Skills and the Child's Theory of Mind. *Child Development* 64(2):348–371.

Montesano, L.; Lopes, M.; Bernardino, A.; and Santos-Victor, J. 2008. Learning Object Affordances: From Sensory-Motor Coordination to Imitation. *IEEE Transactions on Robotics* 24(1):15–26.

Sinapov, J., and Stoytchev, A. 2007. Learning and generalization of behavior-grounded tool affordances. In *ICDL '07*. IEEE.

Starner, T., and Pentland, A. 1997. RealTime American Sign Language Recognition from Video Using Hidden Markov Models. Technical Report 375, M.I.T Media Laboratory Perceptual Computing Section.

Starner, T.; Weaver, J.; and Pentland, A. 1998. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20:1371–1375.

Vogler, C., and Metaxas, D. 1999. Parallel hidden markov models for american sign language recognition. *Computer Vision, IEEE International Conference on* 1:116.

Wilson, A. D., and Bobick, A. 1999. Parametric Hidden Markov Models for gesture recognition. *IEEE transactions on pattern analysis and machine intelligence* 21(9).

Yamato, J.; Ohya, J.; and Ishii, K. 1992. Recognizing human action in time-sequential images using hidden Markov model. In *CVPR '92*. IEEE.