Learning to Place Imaginary Objects Implied by Gestures in Video

Andreya Piplica, Alexandra Olivier, Allison Petrosino, and Kevin Gold
Wellesley College Department of Computer Science
Wellesley, MA
apiplica@wellesley.edu

***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.**

I. Introduction

It can be useful for a robot to be able not only to recognize real objects, but also recognize when a person's gestures imply the existence of imaginary objects. Adults often use such gestures to illustrate ideas [1], while children often use them in pretend play [2]. Because these symbolic gestures are so common in communication and play, robots designed to communicate with humans should be able to recognize both the gestures that imply objects and the objects these gestures imply.

An understanding of object-implying gestures would be advantageous in many areas. It would enhance human-robot interaction by helping robots to interpret natural gesture-based communications. A robotic learner that could recognize pantomimed actions may be able to imitate such actions and interact with the suggested objects [3]. 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 [4].

A vision-based machine learner is presented that can use previously learned properties of objects and actions to reason about what object is implied by a gesture and where the imaginary 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., [5]), and augmented reality, which tracks imaginary objects but typically assumes rather than decides what object is being manipulated [6].

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 [7]. 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. 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 [6], are necessary for realistic interaction with imagined objects.

II. 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.

A. 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 [8]. 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 (Fig. 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.

B. Defining the Actions

HMMs use discrete states to probabilistically describe an action over time [9]. 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 [9] to perform expectation-maximization [10]. 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 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 [7], making it useful for real time applications.

C. 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 [11], 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 [12].

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 [2], 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.

III. Experiments

A. Training

Using a Logitech Quickcam Orbit AF grabbing 640 X 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.

B. 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.

IV. Results

A. HMM Classification

In the eighteen videos with imagined objects, the system chose the correct action sixteen times, yielding an 89% correct classification rate (Fig. 2). The two mistaken classifications both misclassified an action as scooping; the true actions were brushing and petting.

B. 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 (Fig. 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

TABLE I. HMM CLASSIFICATION

True Action	1	2	3
Drinking	drinking	drinking	drinking
Petting	scooping	petting	petting
Hammering	hammering	hammering	hammering
Writing	writing	writing	writing
Scooping	scooping	scooping	scooping
Brushing	brushing	brushing	brushing

Figure 2. Table of most likely HMM action classifications, for each action and participant, in videos where actions are performed without objects.

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. (Fig. 4)

TABLE II. ROTATION COMPARISON CLASSIFICATIONS

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

Figure 3. Independent coder assessments of least squares and von Mises rotation accuracy. Numbers represent how many frames each method was judged to provide more accurate rotation based on the orientation of the hand in that frame.

Figure 4. Least squares regression and von Mises distribution approximations for imagined object rotation.

V. DISCUSSION

The preceding experiment showed that HMM-based methods are feasible for identifying pretended objects from gestures. The most difficult part of placing the object's binary image in the frame was determining an accurate rotation, since the orientation of the hand was difficult to determine from the image. Though it seemed reasonable to attempt to infer the object's rotation from the major axis of an ellipsoid fitted to the skin-segmented hand and smooth this rotation over time, in practice, this ellipsoid was too noisy even with smoothing. Our final solution made use of the repetitive nature of the actions, by fitting the rotation to a function of the object's displacement. This would not work for general motions, but does work for highly stereotyped, repetitive motions.

Repetition and stereotyped movements are common to the symbolic gestures used in communications [1] and young children's social pretend play. A learner that exploits these features may be at an advantage when implemented in a robot designed for interaction in contexts where these gestures occur frequently. In the case of pretend play, a shift from bottom-up, perceptually driven thinking to top-down, contextual expectation driven thinking may help explain why children require a placeholder object for pretense before they can pretend without any placeholder [13].

A toy or video game that recognized object-implying gestures could have pre-programmed information about what types of objects it would expect users to gesture with so that it can be used out of the box. In addition, it could also have a learning mode where users demonstrate the actions performed with a new object and use keywords to describe the contexts in which this action might occur. Depending on the type and number of new objects presented, the toy or game may learn entirely new sets of objects in different contexts and still interact with the user.

The learner is currently being adapted to work in real-time so that it can be implemented in communicative and entertainment robots. Robotic toys may be taught how to reach for and interact with imaginary objects so they may be more thoroughly incorporated in a child's pretense. The probabilistic reasoning based on symbolic gestures presented here may inform systems for robotic imitation, a necessary skill for humanoid robots [14].

ACKNOWLEDGMENTS

We thank all the participants in the training and testing videos and Julia Gall, who coded the rotation comparison data. We also thank the Wellesley College Dean of the College and the Norma Wilentz Hess Faculty and Program Fund for providing funding for this project.

REFERENCES

[1] C. L. Nehaniv, "Classifying Types of Gesture and Inferring Intent," Proc. AISB'05 Symposium on Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction, The Society for the Study of Artificial Intelligence and Simulation of Behaviour, pp. 74-81, 2005.

[2] L. Acredolo and S. Goodwyn, "Symbolic Gesturing in Normal Infants," in Child Development, vol. 59, pp. 450–466, April 1988.

[3] A. Chella, H. Dindo, and I. Infantino, "A cognitive framework for imitation learning," Robotics and Autonomous Systems, vol. 54, pp. 403–408, March 2006.

[4] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning Object Affordances: From Sensory-Motor Coordination to Imitation." IEEE Trans. on Robotics, vol. 24, pp.15–26, December 2008.

[5] H. Lee and J.H. Kim, "An HMM-based threshold model approach for gesture recognition," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 21, pp. 961–973, October 1999.

[6] R. T. Azuma, "A Survey of Augmented Reality," in Presence: Teleoperators and Virtual Environments, vol. 6, pp. 355–385, August 1997.

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

[8] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A Survey of Skin-Color Modeling and Detection Methods," in Pattern Recognition, vol. 40, pp. 1106–1122, March 2007.

[9] E. L. Baum, T. Petrie, G. Soules, and N. Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov chains," in The Annals of MAthematical Statistics, vol. 41, pp. 164–171, February 1970.

[10] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society. Series B. Methodological vol. 39 pp. 1–38, 1977.

[11] C. M. Bishop, Pattern Recognition and Machine Learning. New York: Springer. 2006.

[12] P. Berens and M. Velasco, CircStat2009 The Circular Statistics Toolbox from MATLAB, Technical Report 184, MPI, 2009.

[13] J. L. Elder and D. R. Pederson, "Preschool Children's Use of Objects in Symbolic Play," in Child Development, vol. 56, pp.1253–1258, June 1985.

[14] S. Schaal, "Is imitation learning the route to humanoid robots?" in Trends in Cognitive Sciences, vol. 3, pp. 233-242, June 1999.

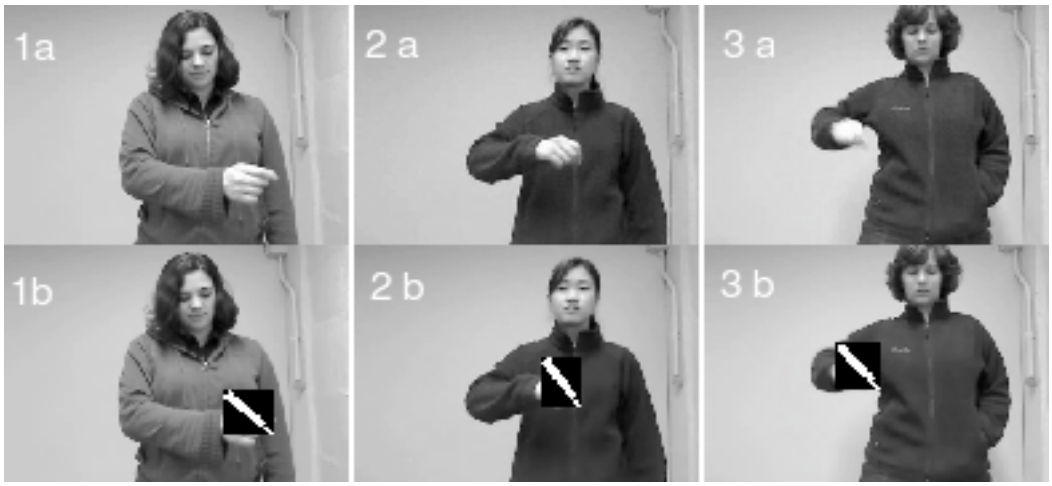

Figure 5. Imagined marker filled in using least squares regression. Analysis of object and hand positioning and rotation data from one participant generalized to allow accurate filling in for three participants.