Learning to Recognize Human Action Sequences

Chen Yu and Dana H. Ballard

Department of Computer Science
University of Rochester
Rochester, NY, 14627
{yu,dana}@cs.rochester.edu

Abstract

One of the major sources of cues in developmental learning is that of watching another person. An observer can gain a comprehensive description of the purposes of actions by watching the other person's detailed body movements. Action recognition has traditionally studied processing fixed camera observations while ignoring non-visual information. This paper explores the dynamic properties of eye movements in natural tasks: eye and head movements are quite tightly coupled with actions. We present a method that utilizes eye gaze and head position information to detect the performer's focus of attention. Attention, as represented by eye fixation, is used for spotting the target object related to the action. Attention switches are calculated and used to segment the action sequence into action units which are recognized by Hidden Markov Models. An experimental system is built for recognizing actions in the natural task of "stapling a letter", which demonstrates the effectiveness of the approach.

1 Introduction

In the drive to understand brain function, a relatively new observation is the role of the body in creating and simplifying brain representations. Although initial attempts to understand actions focused solely on characterizing their effects on the world, it is now appreciated that the accompanying body movements *in situ* provide a helpful code for understanding the effects and purpose of actions in the world.

Most of the actions that make up our lives involve vision. We use eye gaze to locate objects, to direct the hand to a location and to guide the hand to manipulate objects in various ways. Land et al.[1] found that during the performance of a well-learned task(making tea), eyes closely monitor every step of the process although the actions proceed with little conscious involvement. Hayhoe[2] has shown that the eye movements are closely related to the requirements of motor tasks and almost every action in an action sequence is guided and checked by vision, with eye movements usually preceding motor actions.

This paper makes extensive use of the ability to track the course of eye movements in a task and introduces the usefulness of head movements. Particularly in hand-eye coordination tasks, head movements provide valuable cues for the segmentation of such tasks. Gaze, head and hand movements provide a language for interpreting actions in tasks. The goal of this paper is to demonstrate that measurements of these movements can be used to parse typical tasks into actions. The ability to do such parsing is extremely important, as it is a precursor for representing the task linguistically.

Our experiments on stapling a letter (described in Section 6.1) confirms the conclusions by Hayhoe and Land. Furthermore, we notice that at either end of each action, there is almost always an identifiable eye movement(saccade) along with a head movement that switches gaze from one object to another. Within each action, gaze rarely strays from the object of interest though there might be multiple eye fixations on the different parts of the object. In light of this, our hypothesis is that eye and head movements, as an integral part of the motor program of humans, provide important information for action recognition in human activities. We test this hypothesis by developing a method that can segment and recognize action sequences based on eye gaze and head movement.

2 Related Work

Early approaches [3, 4] to action understanding emphasized on reconstruction followed by analysis. More recently, Brand [5] proposes to visually detect causal events by reasoning about the motions and collisions of surfaces using high-level causal constraints. Siskind[6] and Mann et al.[7] present a system that is based on an analysis of the Newtonian mechanics of a simplified scene model. Interpretations of image sequences are expressed in terms of assertions about the kinematic and dynamic properties of the scene. Recently, Ogawara et al.[8, 9] propose a framework of acquiring hand-action models by integrating multiple observations based on geature spotting.

The importance of embodiment is featured in Roy's work [10, 11]. He uses the correlation of speech and vision to associate spoken utterances with a corresponding object's visual appearance. Our work differs from his in that we take an agent-centered view and incorporate an extensive description of the agent's gaze, head and body movements.

3 Overview of Our Approach

Humans perceive an action stream as a sequence of clearly segmented "action units" [12]. This gives rise to the idea that action recognition is to interpret the continuous human behaviors as a sequence of action primitives such as " picking up a coffee pot". To construct a recognition system, we must first detect the time points which correspond to the beginning or the end of the action units. Next, in order to describe "what is happening", these action units need to be recognized as well as the target objects.

In our system, we limit natural tasks to those performed on a table. The system takes the hand positions, the locations of eye and head, and the video sequence captured by a head-mounted camera as input. The output of the system is an action script. Our basic premise is that the eye and head movements are tightly coupled with hand movements in the task. Our approach consists of several stages shown in Figure 1:

1. We first compute eye and head fixations separately using a velocity-based algorithm. The times of action boundaries are extracted by integrating the fixations. Based on these times, the course of hand movements is then partitioned into short segments that correspond to the action units. This is described in Section 4.

2. A sequence of feature vectors extracted from the hand positions of each segment is sent to pre-trained Hidden Markov Models(HMMs) to recognize the actions in the task. This is described in Sections 5.1 and 5.2.

3. Snapshots at the beginning and the end points of each action are analyzed. By using eye gaze as a cue, the object involved in the action is detected and segmented from the background image. The object is recognized by calculating color and shape histograms. It is then straightforward to generate action scripts by integrating the types of motion and the target objects. We describe this in Section 5.3.

4 Segmentation of Action Sequences

The segmentation of human action sequences has been a topic of considerable interest in computer vision. For example, Kuniyoshi et al.[3] focus on detecting meaningful changes in the environment to check for action switches. Ogawara et al.[13] attempt to recognize and extract meaningful segments by gesture spotting. The method of Rui et al.[14] is based on detecting temporal discontinuities of the spatial pattern of image motion that captures the action.

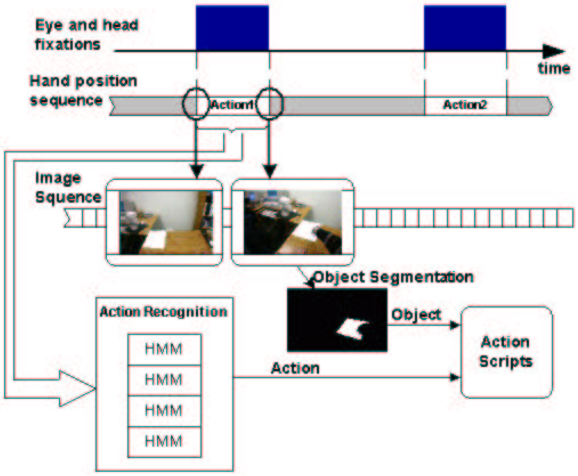

Figure 1: **A summary of our approach**

The novel approach of this paper is to segment the continuous actions in natural tasks by detecting agent-centered switches of attention. Based on the fact that eye and head movements are closely linked to attention[1], we develop a method to detect attention by integrating eye gaze and head position information. In our experiments, we noticed that the temporal relations between eye fixating and hand manipulating are quite tight and predictable, with vision leading action. For example, in the simple action of "picking up an object", the performer rotates his head toward the object first, followed by fixating the eye on the target and moving the arm to approach and grasp it. At the end of the grasping action, the head and eye always move toward another location, which indicates the beginning of the next action.

We observed that actions can occur in two situations: during head fixations and during eye fixations. For example, in a "picking up" action, the performer focuses on the object first then his motor system moves the hand to approach it. During the procedure of approaching and grasping, the head moves toward to the object as the result of the upper body movements but eye gaze remains stationary on the target object. The second case includes such actions as "folding a piece of paper" where the head fixates on the object involved in the action. During the head fixation, eye-movement recordings show that there can be a number of eye fixations ranging from 1 to 6. For example, when the performer folds a piece of paper, he spends 5 fixations on the different part of the paper and 1 look-ahead fixation to the location where he will place it after folding. In this situation, the head fixation is a better cue than eye fixations to segment the actions.

Based on the above analysis, we developed an algorithm for action segmentation, which consists of the following three steps:

1. **Head fixation finding** is based on the positions and orientations of the head. We use (x, y) position on the table plane and 3D orientations to calculate the velocity profile of the head, as shown in the first two rows of Figure 2.

2. **Eye fixation finding** is accomplished by a velocity-threshold-based algorithm. The algorithm significantly reduces the size and complexity of eye data by removing raw saccade(rapid eye movement) points and collapsing raw fixation points into a single representative tuple. A sample of the results of eye data analysis is shown in the third and fourth rows of Figure 2.

3. **Action Segmentation** is achieved by analyzing head and eye fixations, and partitioning the sequence of hand positions into the action segments(shown in the bottom row of Figure 2) based on the following three cases:

   - Within the head fixation, there are one or more than one eye fixations. This corresponds to actions, such as “folding”. “Action 3” in the bottom row of Figure 2 represents this kind of action.
   - During the head movement, the performer fixates on the specific object. This situation corresponds to actions, such as “picking up”. “Action 1” and “Action 2” in the bottom row of Figure 2 represent this class of actions.
   - During the head movement, eyes are also moving. It is most probable that the performer is switching attention after the completion of the current action.

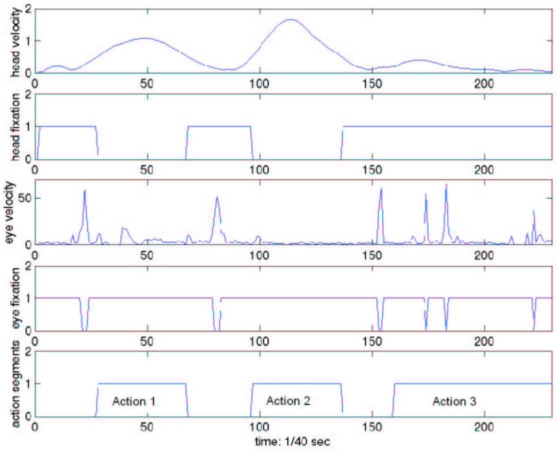

Figure 2: **Segmenting actions based on head and eye fixations The first two Rows:** Point-to-point velocities of head data and the corresponding fixation groups(1–fixating, 0–moving).**The third and fourth rows:** Eye velocity data and the eye fixation groups(1–fixating, 0–moving) after removing saccade points. **The bottom row:** The results of action segmentation by integrating eye and head fixations.

5 Recognition of Human Actions

5.1 Feature Vector Selection

We collect the raw position (x, y, z) and the rotation (h, p, r) data of the hands from which feature vectors are extracted for recognition. In our system, we want to recognize the types of motion not the accurate trajectory of the hand. The same action performed by different people varies. Even in different instances of a simple action of “picking up” performed by the same person, the hand goes roughly in a different trajectory. This indicates that we can not directly use the raw position data to be the features of the actions. As pointed out by Campbell et al.[18], features designed to be invariant to shift and rotation perform better in the presence of shifted and rotated input. The feature vectors should be chosen such that large changes in the action trajectory produce relatively small excursions in the feature space, while the different types of motion produce relatively large excursions. In the context of our experiment, we calculated three element feature vectors consisting of the hand’s velocity on the table plane($d\sqrt{x^2 + y^2}$), the position in the z-axis, and the velocity of rotation in the 3 dimensions($d\sqrt{h^2 + p^2 + r^2}$).

5.2 Hidden Markov Models

Hidden Markov Models(HMMs) have been widely used in speech recognition with great success. Recently, HMMs have been applied within the computer vision community to address problems where time variation is significant, such as action recognition[15, 16] and gesture recognition[17].

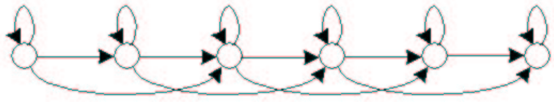

Figure 3: **The HMM used for recognition**

In our experiment, the five actions we sought to detect are: “picking up”, “placing”, “lining up”, “stapling” and “folding”. We model each action as a forward-chaining continuous HMM with the exception of the action of “picking up” and “placing”. These two actions share the same HMM because they involve qualitatively very similar hand movements and cannot be classified using only movement measurements. We explain how these two actions are distinguished in Section 5.3. Each HMM consists of 6 states, each of which can jump to itself and the next two forward-chaining states(Figure 3). The states and transition probabilities are determined by the Baum-Welch algorithm during the HMM training process. In the recognition phase, given a sequence of feature vectors extracted from hand positions, the system determines which HMM most likely generates those observations by calculating the log-probability of each HMM and picking the maximum.

5.3 Object Spotting and Recognition

So far, this paper has described the segmentation and recognition of hand motions characterized by a definite space-time trajectory. To generate a qualitative description of human action sequences, it is necessary to recognize the target objects dealt with in the actions. Instead of time-consuming computation of the image sequence for object spotting, we only analyze snapshots at eye gaze position at the beginning and the end of each action. The leftmost image in Figure 4 shows an example of the snapshot with the eye position at the end point of “picking up” action.

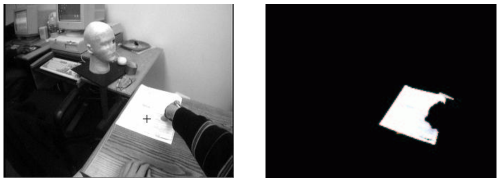

Figure 4: **Left:** The snapshot image in grey scale with eye position(black cross). **Right:** The object extracted from the left image.

In our system, the eye tracking device not only provides the video sequence from the “first-person” perspective but also reports the gaze positions in the image, indicating the user’s attention. Thus, we can directly segment the object of user interest by using eye position as a seed for the region growing algorithm[19]. The segmentation result is shown in the right image of Figure 4. A color histogram[20] and multidimensional receptive field histogram[21] are calculated from the segmented image and combined to form the feature vector for object recognition.

Using the results of object spotting and recognition, the system can discriminate “picking up” and “placing” actions, which share the same HMM. For the action of “picking up”, eye gaze remains stationary with respect to the target object both at the beginning and the end of the action. In contrast, in the action of “placing”, eyes look toward the location which is empty at the beginning of the action. In this way, we can distinguish these two actions which have similar hand movements.

Finally, based on the motion types and the target objects, scripts are generated to describe the actions, such as “picking up a letter”.

6 Experiment

6.1 Data Collection and Preprocessing

A Polhemus 3D tracker was utilized to acquire 6-DOF hand and head positions at $40Hz$. The performer wore a head-mounted eye tracker from Applied Science Laboratories(ASL). The headband of the ASL holds a miniature “scene-camera” to the left of the performer’s head that provides the video of the scene from the first-person perspective. The video signals are sampled at the resolution of 320 columns by 240 rows of pixels at the frequency of 15Hz. The gaze positions on the image plane are reported at the frequency of $60Hz$. Before computing feature vectors for HMMs, all position signals pass through a 6th order Butterworth filter with cut-off frequency of 5 Hertz.

The training data of five actions were collected from one subject. We obtained 12 samples for each action. For recognition experiments, three additional people performed the task of stapling a letter which consists of 8 actions (shown in Figure 5). The three participants performed 6, 8 and 8 trials separately. Overall, 176 $((6+8+8)\times 8)$ actions were collected to test the segmentation and recognition performance.

6.2 Results and Analysis

The results of action segmentation and recognition are shown in Table 1. The action recognition accuracy is better than the raw segmentation because the HMMs eliminate some segments produced by mis-segmentation. The error in segmentation is mainly caused by involuntary head movements and unstable eye fixations. For example, when the performer drives the staple to the letter, the head sometimes moves toward to the stapler. Another example is that in the action of “picking up”, the eyes sometimes randomly look toward some other locations, and then come back to fixate on the target object. Table 2 shows the recognition accuracy for each action based on the action segmentation. The relatively low recognition rate of the “lining up” action is caused by the high variation of hand movements between the training data and the test data.

7 Conclusions

This paper describes a novel method to recognize human actions in natural tasks. The approach is unique in

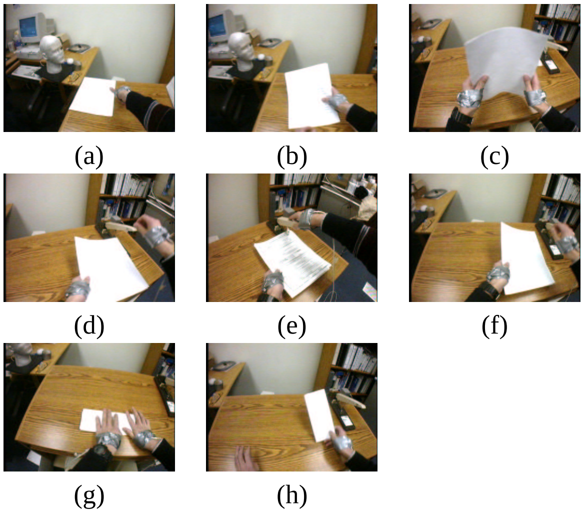

Figure 5: The continuous action sequence in our experiment: (a) picking up the letter (b) placing it to the position close to the body (c) lining up (d) placing it close to the stapler (e) stapling (f) placing it back to the position near the body (g) folding (h) placing it to the target location

Table 1: Results of Action Segmentation and Recognition

	Accuracy
segmentation	83.9%
recognition	91.6%

Table 2: Results of Recognition for five actions

Actions	Accuracy
picking	96.3%
placing	93.6%
lining up	73.2%
stapling	86.2%
folding	83.6 %

that it analyzes both eye gaze and head position to detect the performer's attention. Switches of attention are used for action segmentation and attention during the eye fixations is used for object spotting. The coordination of eye, head and hand movements is utilized for action recognition by integrating multisensory information. We demonstrated our approach in the domain of recognizing the human actions in the task of stapling a letter.

We are interested in learning more complicated actions in natural tasks. For future work, we will build a library of additional action units, like phonemes in speech recognition. Then, the system should be able to learn to recognize newly encountered actions and tasks.

Acknowledgments

The authors wish to express their thanks to Brian Sullivan, Bonnie Huang, Nathan Sprague and Brandon Sanders for their contribution to the experimental system.

References

[1] Michael land, Neil Mennie, and Jennifer Rusted, "The roles of vision and eye movements in the control of activities of daily livng," *Perception*, vol. 28, pp. 1311–1328, 1999.

[2] Mary Hayhoe, "Vision visual routines: A functional account of vision," *Visual Cognition*, vol. 7, pp. 43–64, 2000.

[3] Y. Kuniyoshi and H. Inoue, "Qualitative recognition of ongoing human action sequences," in *Proc. IJCAI93*, 1993, pp. 1600–1609.

[4] Y. Kuniyoshi, M. Inaba, and H. Inoue, "Learning by watching: Extracting reusable task knowledge from visual observation of human performance," *IEEE Transactions on Robotics and Automation*, vol. 10, pp. 799–822, 1994.

[5] Matthew Brand, "The inverse hollywood problem: From video to scripts and storyboards via causal analysis," in *AAAI*, 1997, pp. 132–137.

[6] Jeffrey Mark Siskind, "Grounding language in perception," vol. 8, pp. 371–391, 1995.

[7] Richard Mann, Allan Jepson, and Jeffrey Mark Siskind, "The computational perception of scene dynamics," *Computer Vision and Image Understanding: CVIU*, vol. 65, no. 2, pp. 113–128, 1997.

[8] Koichi Ogawara, Soshi Iba, Hiroshi Kimura, and Katsushi Ikeuchi, "Recognition of human task by attention point analysis pd," in *International Conference on Intelligent Robot and Systems (IROS)'00*, Kagawa, Japan, Nov 2000, vol. 3, pp. 2121 – 2126.

[9] Koichi Ogawara, Soshi Iba, Hiroshi Kimura, and Katsushi Ikeuchi, "Acquiring hand-action models by attention point analysis," in *Inter. Conf. Robotics and Automations (ICRA)*, Seul, 2001, vol. 4, pp. 465–470.

[10] Deb Roy, "Integration of speech and vision using mutual information," in *Proceedings of Int. Conf. Acoustics, Speech and Signal Processing(ICASSP)*, Istanbul, Turkey, June 2000.

[11] Deb Roy, Bernt Schiele, and Alex Pentland, "Learning audio-visual associations using mutual information," in *International Conference on Computer Vision, Workshop on Integrating Speech and Image Understanding. Corfu, Greece*, 1999.

[12] D. Newtson et al., "The objective basis of behavior units," *J. of Personality and Social Psychology*, vol. 35, no. 12, pp. 847–862, 1977.

[13] Koichi Ogawara, Soshi Iba, Hiroshi Kimura, and Katsushi Ikeuchi, "Recognition of human behavior with 9eye stereo vision and data glove," in *Computer Vision and Image Media*, March 2000.

[14] Y. Rui and P. Anandan, "Segmenting visual actions based on spatio-temporal motion patterns," in *Proceedings of CVPR*, Hilton Head, SC, June 2000.

[15] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," in *IEEE CVPR97*, 1997.

[16] A. Bobick, "Movement, activity and action: the role of knowledge in the perception of motion," in *the Royal Society Workshop on Knowledge-based Vision in Man and Machine*, 1997.

[17] Thad Starner and Alex Pentland, "Real-time american sign language recognition from video using hidden markov models," in *ISCV'95*, 1996.

[18] L. Campbell, D. Becker, A. Azarbayejani, A. Bobick, and A. Pentland, "Invariant features for 3-d gesture recognition," in *Second International Workshop on Face and Geasture Recognition*, Killington, VT, Oct. 1996, pp. 157–162.

[19] Rolf Adams and Leanne Bischof, "Seeded region growing," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, June 1994.

[20] Michael J. Swain and Dana Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, pp. 11–32, 1 1991.

[21] B. Schiele and J. L. Crowley, "Object recognition using multidimensional receptive field histograms," in *Proceedings of European Conf. on Computer Vision*, Cambridge, UK, 1996, pp. 1039–1046.