Understanding Human Behaviors Based on Eye-head-hand Coordination

Chen Yu and Dana H. Ballard

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

Abstract. Action recognition has traditionally focused on processing fixed camera observations while ignoring non-visual information. In this paper, we explore the dynamic properties of the movements of different body parts in natural tasks: eye, head and hand movements are quite tightly coupled with the ongoing task. In light of this, our method takes an agent-centered view and incorporates an extensive description of eye-head-hand coordination. With the ability to track the course of gaze and head movements, our approach uses gaze and head cues to detect agent-centered attention switches which can then be utilized to segment an action sequence into action units. Based on recognizing those action primitives, parallel hidden Markov models are applied to model and integrate the probabilistic sequences of the action units of different body parts. An experimental system is built for recognizing human behaviors in three natural tasks: "unscrewing a jar", "stapling a letter" and "pouring water", which demonstrates the effectiveness of the approach.

1 Introduction

Humans perceive an action sequence as several action units[1]. This gives rise to the idea that action recognition is to interpret continuous human behaviors as a sequence of action primitives. However, we notice that a sequence of action primitives is not the final outcome of visual perception. Humans have the ability to group those units into a high-level abstract representation that corresponds to tasks or subtasks. For example, in our experiments, one subject performed some natural tasks while the other subject was asked to describe the actions of the performer. The verbal descriptions of the speaker mostly correspond to subtasks but not action units. For instance, the speaker would say "he is unscrewing a jar", but would not describe activities in such details as "his hand is approaching a jar", "then he is grasping it" and "he is holding the jar while unscrewing it". Thus, the speaker conceptualizes the sensory input into the abstract level corresponding to tasks or subtasks, then verbalizes the perceptual results to yield utterances. Based on this observation, we argue that to mimic human capabilities, such as describing visual events verbally, the goal of action recognition is to recognize not only action primitives but also tasks and subtasks. In light of this, this work concentrates on recognizing tasks instead of action primitives.

Recent results in visual psychophysics[2, 3] indicate that in natural circumstances, the eye, the head, and hands are in continual motion in the context of ongoing behavior.

This requires the coordination of these movements in both time and space. Land et al.[2] found that during the performance of a well-learn task(making tea), the eyes closely monitor every step of the process although the actions proceed with little conscious involvement. Hayhoe[3] has shown that eye and head 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 and head movements usually preceding motor actions. Moreover, their studies suggested that the eyes always look directly at the objects being manipulated. In our experiments, we confirm the conclusions by Hayhoe and Land. For example, in the action of "picking up a cup", the subject first moves the eyes and rotates the head to look towards the cup while keeping the eye gaze at the center of view. The hand then begins to move toward the cup. While the subject grasps the cup, the eyes are fixating it to guide the action.

Despite the recent discoveries of the coordination of eye, head and hand movements in cognitive studies, little work has been done in utilizing these results for machine understanding of human behavior. In this paper, 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 segments action sequences based on the dynamic properties of eye gaze and head direction, and applies Parallel Hidden Markov Models(PaHMMs) to integrate eye gaze and hand movements for task recognition.

2 Related Work

Early approaches [1] to action understanding emphasized on reconstruction followed by analysis. More recently, Brand [4] proposes to visually detect causal events by reasoning about the motions and collisions of surfaces using high-level causal constraints. Mann and Siskind[5] 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. Presently, Hidden Markov Models(HMMs) have been applied within the computer vision community to address action recognition problems in which time variation is significant. Starner and Pentland[6] have developed a real-time HMM-based system for recognizing sentence-level American Sign Language(ASL) without explicitly modeling the fingers. Wilson and Bobick[7] have proposed an approach for gesture analysis that incorporates multiple representations into the HMM framework. Our work differs from theirs in that we take an agent-centered view and incorporate an extensive description of the agent's gaze, head and hand movements.

3 Attention-based Action Segmentation

The segmentation of a continuous action stream into action primitives is the first step towards understanding human behaviors. With the ability to track the course of gaze and head movements, our approach uses gaze and head cues to detect agent-centered attention switches which can then be utilized to segment human action sequences.

In our experiments, we notice that actions can occur in two situations: during eye fixations and during head fixations. For example, in a "picking up" action, the performer

focuses on the object first then the motor system moves the hand to approach it. During the procedure of approaching and grasping, the head moves towards 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 "pouring water" in which 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. For example, when the performer is pouring water, he spends 5 fixations on the different parts of the cup and 1 look-ahead fixation to the location where he will place the water pot after pouring. In this situation, the head fixation is a better cue than eye fixations to segment the actions.

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

1. **Head fixation finding** is based on the orientations of the head. We use 3D orientations to calculate the speed profile of the head, as shown in the first two rows of Figure 1.
2. **Eye fixation finding** is accomplished by a velocity-threshold-based algorithm. A sample of the results of eye data analysis is shown in the third and fourth rows of Figure 1.
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 1) 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 "unscrewing". "Action 3" in the bottom row of Figure 1 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 1 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.

4 Task Recognition through Parallel HMMs(PaHMMs)

Based on action segmentation, the course of eye and hand movements is partitioned into short segments that correspond to action units. Our method of task recognition is based on recognizing the action units and modeling the probabilistic sequences of those action primitives in tasks. Parallel HMMs consisting of two sets of HMMs are implemented to model the movements of different body parts in parallel. One set is to model eye movements in natural tasks, which is described in subsection 4.1. The other presented in subsection 4.2 uses hand movements as input. At the end point of a task, the probability estimates of two models are combined for recognizing tasks.

4.1 Object Sequence Model Based on Gaze Fixations

There are two kinds of eye movements: saccade and fixation. Saccades are rapid eye movements that allow the fovea to view a different portion of the display. Often a saccade is followed by one or multiple fixations when the objects in a scene are viewed. In the context of performing natural tasks, cognitive studies show that the eyes always look directly at the objects being manipulated[2, 3]. Also, in the computer vision field, the

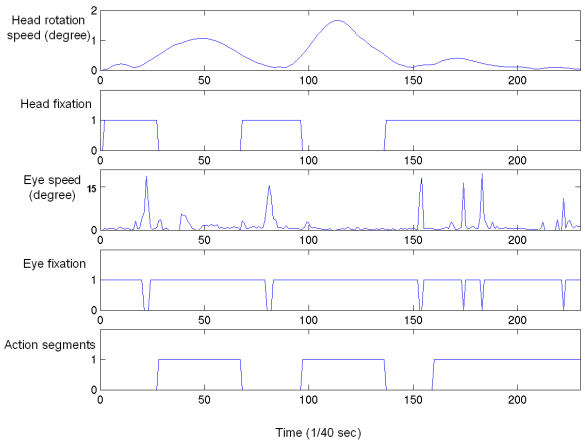

Fig. 1. Segmenting actions based on head and eye fixations The first two Rows: Point-to-point speeds of head data and the corresponding fixation groups(1–fixating, 0–moving).**The third and fourth rows:** Eye movement speeds and the eye fixation groups(1–fixating, 0–moving) after removing saccade points.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

usefulness of object context to perform action recognition is appreciated by the work of Moore et al.[8]. Based on these results, we argue that the sequence of the fixated objects in a natural task implicitly represents the agent's attention in time and provides helpful information for understanding human behaviors.

We develop discrete 6-state HMMs that model the sequences of the fixated objects. The observations of HMMs are obtained by the following steps:

1. **Eye fixation finding** is accomplished by a velocity-threshold-based algorithm. For each action unit obtained from segmentation, there can be a number of eye fixations ranging from 1 to 6.
2. **Object spotting** is implemented by analyzing snapshots with eye gaze positions during eye fixations. Figure 2 shows that the object of agent interest is spotted by using the eye position as a seed for region growing algorithm[9]. Then a color histogram and multidimensional receptive field histogram are calculated from the segmented image and combined to form a feature vector for object recognition. Further information can be obtained from [10].
3. **Observations of HMMs** are obtained by symbolizing the objects involved in tasks. In practice, we notice that there might be multiple eye fixations during an action. Distinct symbols are used to represent the possible combinations of fixated objects. In this way, the discrete observations of HMMs consist of all the objects and their possible combinations in our experiments(described in Section 5), which include "cup", "water pot", "cup+water pot", "stapler", "paper", "stapler+paper", "jar", "lid", "jar+lid" and "nothing".

4.2 Hand Movement Model

Figure 3 illustrates the hierarchical HMMs that are utilized to model hand movements. Firstly, a sequence of feature vectors extracted from the hand positions of each action

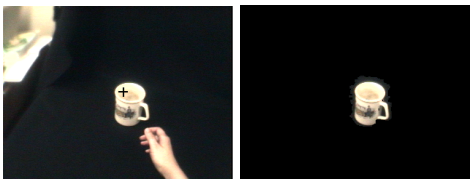

Fig. 2. Left: A snapshot with eye position(black cross) in the action of "picking up". **Right:** The object extracted from the left image.

segment is sent to low-level Hidden Markov Models(HMMs) to recognize the motion type. Then, motion types are used as observations of high-level discrete HMMs whose output will be merged with the models of other HMMs running in parallel.

We now give a brief description of the method for motion type recognition. Further information can be obtained from [10]. The six actions we sought to recognize in our experiments were: "picking up", "placing", "holding", "lining up", "stapling" and "unscrewing". We model each action as a forward-chaining continuous HMM plus a HMM for any other motions. Each HMM consists of 6 states, each of which can jump to itself and the next two forward-chaining states. Given a sequence of feature vectors extracted from hand positions, we determine which HMM most likely generates those observations by calculating the log-probability of each HMM and picking the maximum.

High-level HMMs model the probabilistic sequences of motion types in different tasks. The outputs of low-level HMMs, motion types, are used as the observation sequences of high-level HMMs, each of which is composed of 5 hidden states. The states and transition probabilities are determined by the Baum-Welch algorithm during the HMM training process.

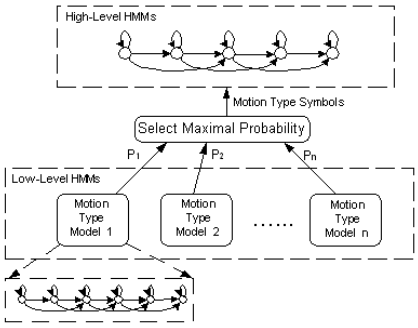

Fig. 3. The hierarchical HMM for action recognition consists of the low-level HMMs for motion types and the high-level HMMs for modeling the sequences of motion types in tasks.

4.3 Integration of Eye and Hand Movements Using PaHMMs

PaHMMs were first suggested by Bourlard and Dupont[11] in subband-based speech recognition. They divided the speech signals into subbands that were modeled independently. The outputs of subbands were then merged to eliminate unreliable parts of the

speech signal. Vogler and Metaxas[12] first introduced PaHMMs in the computer vision field. They developed PaHMMs to model two hand movements for American Sign Language Recognition.

PaHMMs model C processes with C independent HMMs with separate output. The HMMs for the separate processes are trained independently to determine the parameters of each HMM. In the recognition phase, it is necessary to integrate information from the HMMs representing different processes. Using the likelihood-based criterion, we want to pick the model M^k maximizing:

$$\max_k \ logP(O_1, ..., O_C | M_1^k, ..., M_C^k) \tag{1}$$

where the kth PaHMM consists of $M_1^k, ..., M_C^k$, each of which is a HMM. $O_1, ..., O_C$ are observation sequences. Since each process is supposed to be independent to others, we can represent equation 1 as

$$\max_k \sum_{i=1}^{C} log(P(O_i | M_i^k)) \tag{2}$$

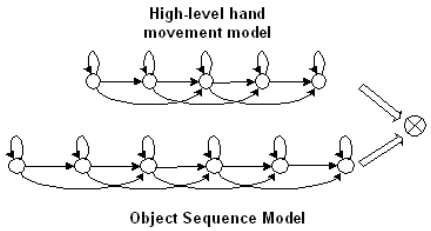

Fig. 4. Parallel HMMs: The streams of body movements are processed in parallel and integrated to yield global scores and a global recognition decision.

Figure 4 shows the approach to integrate eye and hand movements. In the merging state, the probabilities of individual HMMs are combined to yield global scores. When outputs are linearly combined, the expected error will decrease, both in theory and in practice. Therefore, the combination strategy used here is the linear weighted average:

$$\sum_{i=1}^{C} w_i log(P(O_i | M_i^k)) \tag{3}$$

where $w_i \in [0, 1]$ is a fixed weight for each stream. w_i reflects the extent to which the stream contains features that are useful for recognition. The weighting factors are computed by using maximum likelihood from the training data.

5 Experiments

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 a 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

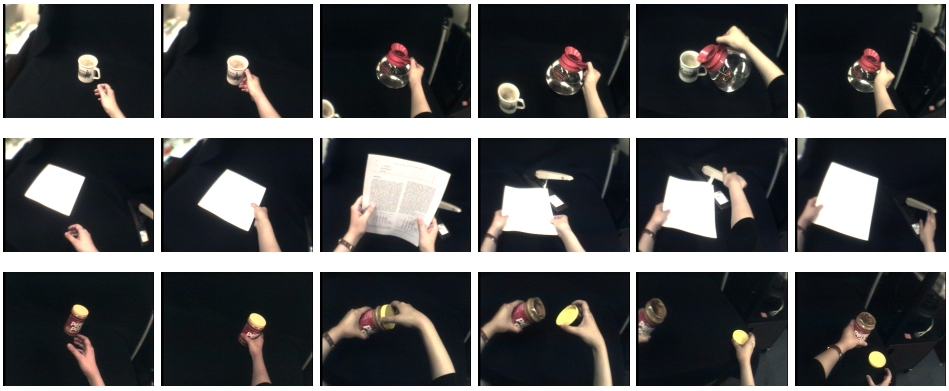

Fig. 5. Snapshots of three continuous action sequences in our experiments. **Top row:** Pouring water. **Middle row:** Stapling a letter. **Bottom row:** Unscrewing a jar.

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 $5Hz$.

In this study, we limited the possible tasks to those on a table. The three tasks we sought to detect were: "stapling a letter", "pouring water" and "unscrewing a jar". Figure 5 shows snapshots captured from the head-mounted camera when a subject performed three tasks. We collected 108 action sequences, 36 for each task. The first 18 sequences of each task are used for training and the rest for testing.

The results of task recognition are shown in Figure 6. To evaluate the performance of PaHMMs, we also test the recognition rates of using object sequence HMMs and hand HMMs individually. PaHMMs provide a clear advantage compared with other approaches. We also note that object sequence HMMs outperform hand movement HMMs. This demonstrates that temporal object sequences implicitly indicate the performer's focus of attention during action execution and provide important information for machine understanding of human behavior.

6 Conclusion

This paper describes a novel method to recognize human behaviors in natural tasks. The approach is unique in that the coordination of eye, head and hand movements is utilized for task recognition. The integration of multistream body movements is achieved by PaHMMs, in which different streams of body movements are processed in parallel and integrated at the end point of a task. The advantages of this method are twofold. Firstly, the movement sequences of different body parts are not restricted to the same sampling rate and the underlying HMMs associated with individual sequences do not necessarily have the same topology. Secondly, merging different sources of body movements can improve the recognition rate in the way that possibly occurring noise in one stream does not degrade the performance so much since other uncorrupted streams yield sufficient information for recognition.

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 recog-

	Accuracy
PaHMMs	96.3%
Object Sequence HMMs	90.7%
Hand HMMs	78.5%

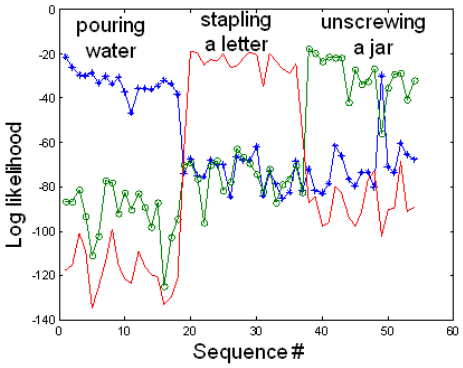

Fig. 6. Left table: the results of task recognition. **Right plot:** per-task sequence log likelihood. The sequences are sorted for ease of comparison. The left third represent the task of "pouring water", the middle third represent the task of "stapling a letter", and the right third represent the task of "unscrewing a jar".

nition. As a result, when a performer works on different kinds of tasks over extended durations(e.g., over the course of an hour), the system could learn to recognize newly encountered actions and tasks without human involvement.

References

1. Kuniyoshi, Y., Inoue, H.: Qualitative recognition of ongoing human action sequences. In: Proc. IJCAI93. (1993) 1600–1609
2. Land, M., Mennie, N., Rusted, J.: The roles of vision and eye movements in the control of activities of daily livng. Perception **28** (1999) 1311–1328
3. Hayhoe, M.: Vision visual routines: A functional account of vision. Visual Cognition **7** (2000) 43–64
4. Brand, M.: The inverse hollywood problem: From video to scripts and storyboards via causal analysis. In: AAAI. (1997) 132–137
5. Mann, R., Jepson, A., Siskind, J.M.: The computational perception of scene dynamics. Computer Vision and Image Understanding: CVIU **65** (1997) 113–128
6. Starner, T., Pentland, A.: Real-time american sign language recognition from video using hidden markov models. In: ISCV'95. (1996)
7. Wilson, A., Bobick, A.: Learning visual behavior for gesture analysis. In: Proceedings of the IEEE Symposium on Computer Vision, Florida, USA (1995)
8. Moore, D., Essa, I., Hayes, M.: Exploiting human actions and object context for recognition tasks. In: In Proceedings of IEEE International Conference on Computer Vision 1999 (ICCV 99), Corfu, Greece (1999)
9. Adams, R., Bischof, L.: Seeded region growing. IEEE Transaction on Pattern Analysis and Machine Intelligence **16** (1994)
10. Yu, C., Ballard, D.H.: Learning to recognize human action sequences. In: Proceedings of the 2nd International Conference on Development and Learning, Boston, U.S. (2002) 28–34
11. Bourlard, H., Dupont, S.: Subband-based speech recognition. In: Proc. ICASSP '97, Munich, Germany (1997) 1251–1254
12. Vogler, C., Metaxas, D.N.: Parallel hidden markov models for american sign language recognition. In: ICCV (1). (1999) 116–122