

# Autonomous Segmentation of Human Action for Behaviour Analysis

J. E. Hunter<sup>1</sup>, D. M. Wilkes<sup>1</sup>, D. T. Levin<sup>2</sup>, C. Heaton<sup>2</sup>, M. M. Saylor<sup>2</sup>

<sup>1</sup>Center for Intelligent Systems

Vanderbilt University

Nashville TN 37235-0131 USA

{jonathan.e.hunter; mitch.wilkes}@vanderbilt.edu

<sup>2</sup>Department of Psychology and Human Development

Vanderbilt University

Peabody College #512

Nashville TN 37203-5701 USA

{daniel.t.levin; caroline.heaton; megan.saylor}@vanderbilt.edu

## Abstract

*To correctly understand human actions, it is necessary to segment a continuous series of movements into units that can be associated with meaningful goals and subgoals. Recent research in cognitive science and machine vision has explored the perceptual and conceptual factors that a) determine the segment boundaries that human observers place in a range of actions, and b) allow successful discrimination among different action-types. In this project we investigated the degree to which specific movements effectively predict key sub-events in a broad range of actions in which a human model interacts with objects. In addition, we aimed to create an accessible tool to track human actions for use in a wide range of machine vision and cognitive science applications. Results from our analysis suggest that a set of basic movement cues can successfully predict key sub-events such as hand-to-object contact, across a wide range of specific tasks, and we specify parameters under which this prediction might be maximized.*

## 1. Introduction

A key goal of both cognitive scientists and researchers in machine vision is to understand how the actions of sentient agents such as humans are processed, identified, and understood [1-10]. The most salient challenge inherent to this process is the need to segment a continuous set of visual movements into meaningful discrete actions. The need to segment movements into meaningful actions is similar to the need to segment a continuous speech stream into discrete meaningful words, and both cases involve defining small atomic units that can be identified, and grouped into larger meaningful units. Here, we briefly review research exploring this problem within the cognitive science and machine vision literatures then present an experiment testing the utility of a machine vision approach to action perception in a setting that would be most useful for future cognitive science research on action perception.

An important limitation of machine vision research has been that the actions used to train and test these automated

systems have either been very limited in scope, and possibly poorly representative of real-world action analysis, or they have been broader, but more focused on identifications of very basic gross bodily movements (such as standing or sitting), or cyclic actions (such as walking). Little research has focused on the kinds of action that might characterize the early human social learning environment which includes many face-to-face interactions with people as they look at and manipulate objects. Another issue with the research in machine vision systems is that it often relies on an array of input devices or a highly calibrated system of cameras to provide information for segmenting and identifying actions that would be impractical for use in most labs that study human behavior. A large percentage of the information in human learning interactions is gathered strictly from vision across a broad field of view. Therefore, one of the goals of the current project is to create a visual tracking system that can be adapted for easy use in fields outside of engineering. This system has numerous uses and can be applied across many interdisciplinary research fields. The system should be robust in its abilities, yet cost efficient enough to be affordable to most any lab.

## 2. Background/Related Work

### 2.1. Behavioral results

Early cognitive science research in action perception focused both on basic action segmentations, and the role of more abstract expectations in interpreting action. The segmentation research demonstrated that observers could reliably provide segmentation markers that seemed to coincide with breaks between actions, and that the units defined by these breakpoints were psychologically salient. For example, when subjects view stills from moments selected as action breakpoints, they are better able to reconstruct the narrative sequence that the actions instantiate. More recent research has confirmed that action segments can be organized hierarchically, with large goal-defined actions (e.g. emptying the dishwasher) subsuming actions that represent subgoals (e.g. putting a single dish away)[1].

Much of this research has assumed that human-generated action segmentations represent the combined influences of basic perceptual cues such as changes in the direction of moving body parts, and more complex cognitive constraints such as an understanding both of context-consistent sequences of actions, and of the actor's goals. For example, in one recent study [7], subjects were asked to segment the movements of a two simple shapes on a computer screen. One group of subjects was told that the movements were generated by two people playing a game, and the other group was told (correctly) that the movements were randomly generated. Both groups then segmented the actions. Results indicated that the segmentations were predicted by a number of basic movement features such as direction changes and the mean proximity of the two objects. However, these basic movement features predicted segmentations most strongly when subjects believed that the movements were random. According to the researcher, this occurred because subjects in the person condition focused more on abstract conceptual goals and less on specific movement features than subjects in the random condition.

The above study was one of the first to explore the specific features that predict human-generated action segmentations, but it was limited because the stimuli used were, by necessity, relatively artificial. Although this produced the desired unambiguous result of conceptual attribution, it leaves open questions about the degree to which these movements would predict action segmentations in more realistic actions involving a full human figure interacting with objects.

To explore the features that might predict action segmentations in a more ecological context, we completed an analysis of segmentations for a wide range of realistic actions in which a set of human models was videotaped completing a series of ten different tasks with a range of objects [17]. Instead of using basic movement features to predict segments, we defined a set of more meaningful subactions that were hand coded. These included hand-to-object contacts, object-to-object contacts, occlusions, and eye movements. We found that multiple regressions based on these subactions predicted up to 82% of the variance in the number of breakpoints entered (by eight judges) in each one-second bin.

This previous study suggests that segmentation of natural actions might benefit from an explicit attempt to link the basic movements and direction changes that are usually the basis of machine vision approaches with more meaningful subactions prior to attempts at trajectory grouping. Accordingly, the system we describe in the next section was used to parse a broad set of face-to-face actions. These parses were then tested for their ability to predict hand coded subactions similar to those coded in our previous project.

## 2.2. System basis

The hardware for the system is a digital camcorder that feeds into either a laptop or desktop computer. Since the system uses a single camera to collect information, the video information will be from a single uncalibrated camera view point. The initial visual system is also described in Tugcu's dissertation [14].

The system we used starts by training on the objects of interest that are hand defined and provided by the user. These objects are sent to the system by selecting regions of the object from the 480 row x 720 column image. To represent these objects, our system uses feature vectors composed of a high dimensional HSV color space histogram along with a Laplacian texture measure. The region is broken into 7x7 blocks of pixels. Each block has its colors represented by the color histogram and is stored as a feature vector of that particular object.

After a training database is created for all the desired objects, this database is compared to a new 480x720 image that is broken into overlapping 7x7 regions. To lessen the amount of processing necessary for each comparison an approximate nearest neighbor tree is constructed and used to classify each pixel of the processed image [14,15].

Using the resulting 118x178 segmented images, location information for the different objects is extracted.

## 3. Methods

In this experiment, we trained a system using a set of 10 face-to-face tasks, each performed by 10 models. The system processed the entire set of frames for the videos of each model performing each task, segmenting the frames, storing them, and then using them as the basis for motion tracking. The videos were marked by a human rater (JH) for specific subevents, to test the degree to which extracted motions could effectively predict subevents such as hand-to-object contacts, and specific gaze events

### 3.1. Behavioral Tasks

The tasks are various assembly and sorting types shown in Table 1.

Table 1: *Experiment Tasks*

Tasks	Description
Task 1: The assembly of 3 flashlights	The flashlights are fully dismantled. The participant must construct all 3 flashlights.
Task 2: The assembly of 3 item baskets	The basket, lid, tissue paper, and green Legos are in their own groups. There is also a stamp block. The user must construct a basket by placing a Lego, tissue paper and the lid in that order. The basket is finalized by stamping it with the stamp block. This is to be repeated for the next two baskets.
Task 3: The assembly of pipe structure	Four cylinder shaped pipes and two junction pipes are connected in a particular manner. All the pipes must be used to construct a structure.
Task 4: The sorting and filling of containers	Six containers are stacked on top of each other and must be rearranged in a particular order with Lego blocks placed inside each one in a particular order.
Task 5: The filling of containers with Legos	Three yellow containers are to be filled with one color of Legos.

	The pile of Legos consist of 3 different colors and are all piled together.
Task 6: The removal of Legos from containers and storing into another container	Three containers are located side by side. The two periphery containers contain Legos to be moved to the center container.
Task 7: Occlusion Movement	Two Legos and a occluding object are on the table. The Legos are to be moved behind the occluding object, then moved to other sided of the occluding object.
Task 8: Occlusion Assembly	A T-shaped structure is created from Legos in plain sight of camera. Then, another T-shaped structure is created behind the occluding object. Finally, the occluding object is moved away.
Task 9: Assembly of 3 T-shaped structures	The different-colored T-shaped Lego structures are constructed.
Task 10: Lego Stacking	Legos are to be stacked until all of the Legos are used, or the structure collapses. There are two attempts at this task.

### 3.2. System Assumptions

For each task, a model sits at a table with a set of objects and performs some type of assembly task. The current system assumes that there are 2 hands and 1 gaze estimator in the video. Once these three regions are classified by the user, the features of the video are extracted. To facilitate segmentation, the model wore a red glove on their right hand and a purple glove on their left hand and a hat with a lime green strip down the center. The camera faced the model from across the table (Figure 1).

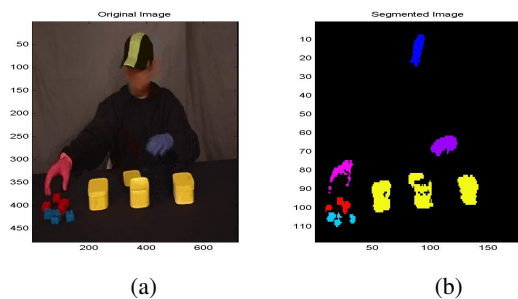


Figure 1: (a) Original Image (b) Segmented Image

Each tree is trained on the set of the same task since each task includes the same set of objects. After processing the frame, the frame number, object name, object center row value, object center column value, max width of object, max height of object, number of pixels of object, and angle displacement from vertical line through centroid of the head are recorded in a text file for the second stage of the analysis.

### 3.3. Behavior Feature Vector

Based on the movements of the segmented objects, a new feature vector was created to analyze the motions of

the person in the video. Many of these feature vectors correspond to some of the stronger features in other studies [11-13, 16]. The 12 computed features are shown in Table 2.

Table 2: Behavior Features

Behavior Feature	Description
Magnitude velocity of Hand 1	This value is calculated by taking the centroid values of hand 1 between successive frames and calculating the Euclidean distance between them. These distances are averaged across the bin
Velocity Stop of Hand 1	This is a binary value that is decided by if the mean velocity of hand 1 is less than 1.5 pixels between successive frames of that bin (1 if true, 0 if false).
Object Contact of Hand 1	This is a binary value that is decided by drawing a line between the centroid of an object and the centroid of hand 1. If the number of pixels that are not classified as the hand or the object in question is less than 2 for all object groups in the image, then hand 1 is considered near an object. (1 if true, 0 if false).
Hand 1/ Object Change	This is an average of the change in the number of pixels represented by objects when hand 1 is within 30 pixels of them.
Magnitude velocity of Hand 2	This value is calculated by taking the centroid values of hand 2 between successive frames and calculating the Euclidean distance between them. These distances are averaged across the bin.
Velocity Stop of Hand 2	This is a binary value that is decided by if the mean velocity of hand 2 is less than 1.5 pixels between successive frames of that bin (1 if true, 0 if false).
Object Contact of Hand 2	This is a binary value that is decided by drawing a line between the centroid of an object and the centroid of hand 2. If the number of pixels that are not classified as the hand or the object in question is less than 2 for all object groups in the image, then hand 1 is considered near an object. (1 if true, 0 if false).
Hand 2/ Object Change	This is an average of the change in the number of pixels represented by objects when hand 2 is within 30 pixels of them.
Gaze Velocity	This is the mean velocity of the estimated gaze angle change across a bin. (Note: Gaze is estimated by using a stripe on the participant's hat. The angle is calculated by estimating the angle between the best fit line for the points of the stripe and the vertical line between the centroid of the stripe)
Gaze Object	This is a binary value that is decided by if the gaze angle is within 10 degrees of an object for at least half the frames of a bin (1 if true, 0 if false).
Gaze Hand	This is a discrete value with possible values of {0, 1, 2, 3}. This value is determined if the gaze estimation is within 10 degrees of : None of the Hands – yields a value 0 Hand 1 alone – yields a value 1 Hand 2 alone – yields a value 2

	Both of the Hands – yields a value 3
Gaze Stop	This is a binary value calculated by if the gaze velocity of a bin is less than the mean gaze velocity of the entire video (1 if true, 0 if false).

### 3.4. Significant Moment Labeling

The first author marked video frames in which key subevents occurred as defined by their importance in describing the steps needed to complete the task. These moments are chosen to be at the finest resolution of the motions in the task. Selection of these subevents reflected the findings of the study mentioned earlier [17] where the segmentation boundaries corresponded to hand-to-object contact and object-to-object interactions with gaze confirmation (using participant gaze to disambiguate the model’s current focus of attention). A frame was selected as a significant frame if there was hand to object contact; hand induced object to object contact, or releasing of an object. In many cases, key events extended over multiple frames. For example, the tasks often require combination of objects. These combinations require the contact of two objects and applying force to squeeze them together. During the moment of the hands holding the objects and applying the force, all frames depicting this event were marked for the subevent. Depending on the task, the participant, and the bin size, the number of marked bins ranged between one-third to one-half of the total bins in the video.

### 3.5. Bin Size Analysis

All of the features are dependant on bin-size. The videos were recorded with a frame rate of about 30 frames per second. So, the feature vectors were calculated for bin-sizes of 1, 3, 6, 8, 10, and 20 frames. Bins were defined as “marked” if any frame within the bin had been selected by the rater. The bins included sequential sets of bin-size frames with no overlap (e.g. for bin-size 3, the first vector consisted of frames 1-3, the second will consisted of frames 4-6, and so on).

### 3.6. D-Prime Performance Measure

The measure used to determine performance is called d-prime ( $D'$ ). In Table 1, there are 2  $D'$  measures,  $D'_1$  and  $D'_2$ , that are calculated.  $D'_1$  takes the average hit rate and the average false alarm rate of all the then calculates the  $D'$  from that value.  $D'_2$  is calculated by taking the individual  $D'$  for all the available hit rates and corresponding false alarm rates, and taking the average of all the  $D'$  values. Since  $D'_2$  has the possibility of containing infinity values, those individual  $D'$  values are replaced with a value of 0.5 to represent maximum uncertainty. This measure is the average of the  $D'$  values for each individual test set.

### 4. Prediction Analyses

The data were in the format of 100 videos (10 participants doing 10 tasks each). Each video had its set of behavior feature vectors calculated for each of the bin-sizes for analysis. Various regression models were trained and tested on the data to determine the best methods using the  $D'$  measure.

The data were analyzed using a “jack knife” method in which each participant was removed and a regression classifier was trained on the remaining data and tested for its ability to fit removed participant’s data. This was done for all participants using linear, quadratic, and Mahalanobis regression techniques. The same analysis for a task jack knife was performed as well. The results were very similar to the results of the subject jack knife.

Other methods were used as well such as k-nearest neighbors and support vector machines. The k-nearest neighbor method was the closest to matching the best results of the linear regression analysis. Support vector machines (SVM) using the radial basis function performed worse and using the linear svm performed about the same as the linear regression.

Table 3: Comparison of Bin Size and Regression Models using Subject Jack Knife

Binsize (frames)	Linear				Quadratic				Mahalanobis			
	HR ( $\mu, \sigma$ )	FAR ( $\mu, \sigma$ )	$D'_1$	$D'_2$ ( $\mu, \sigma$ )	HR ( $\mu, \sigma$ )	FAR ( $\mu, \sigma$ )	$D'_1$	$D'_2$ ( $\mu, \sigma$ )	HR ( $\mu, \sigma$ )	FAR ( $\mu, \sigma$ )	$D'_1$	$D'_2$ ( $\mu, \sigma$ )
1	(0.869, 0.060)	(0.513, 0.088)	1.088	(1.124, 0.294)	(0.195, 0.163)	(0.042, 0.065)	0.864	(0.886, 0.460)	(0.378, 0.119)	(0.113, 0.048)	0.903	(0.921, 0.265)
3	(0.839, 0.077)	(0.399, 0.090)	1.247	(1.309, 0.357)	(0.842, 0.078)	(0.396, 0.088)	1.267	(1.325, 0.327)	(0.478, 0.131)	(0.146, 0.062)	0.998	(1.038, 0.321)
6	(0.859, 0.079)	(0.385, 0.011)	1.368	(1.411, 0.385)	(0.819, 0.083)	(0.349, 0.096)	1.300	(1.362, 0.338)	(0.488, 0.129)	(0.132, 0.067)	1.085	(1.131, 0.332)
8	(0.843, 0.092)	(0.375, 0.126)	1.326	(1.396, 0.358)	(0.792, 0.097)	(0.328, 0.107)	1.259	(1.332, 0.375)	(0.516, 0.132)	(0.157, 0.077)	1.047	(1.109, 0.385)
10	(0.804, 0.114)	(0.360, 0.134)	1.215	(1.298, 0.385)	(0.738, 0.108)	(0.303, 0.105)	1.152	(1.210, 0.361)	(0.496, 0.131)	(0.161, 0.086)	0.982	(1.053, 0.424)
20	(0.626, 0.127)	(0.432, 0.211)	0.494	(0.472, 0.546)	(0.345, 0.128)	(0.185, 0.148)	0.500	(0.402, 0.400)	(0.395, 0.131)	(0.226, 0.155)	0.487	(0.435, 0.418)

Table 4: Top 3 Feature Set Results for Subject Jack Knife

Top 5 Results	Linear				Quadratic				Mahalanobis			
	HR ( $\mu, \sigma$ )	FAR ( $\mu, \sigma$ )	D <sub>1</sub> '	D <sub>2</sub> ' ( $\mu, \sigma$ )	HR ( $\mu, \sigma$ )	FAR ( $\mu, \sigma$ )	D <sub>1</sub> '	D <sub>2</sub> ' ( $\mu, \sigma$ )	HR ( $\mu, \sigma$ )	FAR ( $\mu, \sigma$ )	D <sub>1</sub> '	D <sub>2</sub> ' ( $\mu, \sigma$ )
1 <sup>st</sup>	(0.891, 0.079)	(0.435, 0.134)	1.397	(1.461, 0.391)	(0.908, 0.069)	(0.458, 0.134)	1.431	(1.486, 0.424)	(0.875, 0.082)	(0.407, 0.128)	1.388	(1.439, 0.386)
2 <sup>nd</sup>	(0.889, 0.076)	(0.432, 0.132)	1.391	(1.460, 0.371)	(0.903, 0.070)	(0.450, 0.125)	1.427	(1.458, 0.426)	(0.875, 0.083)	(0.406, 0.128)	1.388	(1.439, 0.386)
3 <sup>rd</sup>	(0.886, 0.080)	(0.426, 0.136)	1.391	(1.467, 0.397)	(0.917, 0.066)	(0.482, 0.132)	1.426	(1.466, 0.437)	(0.875, 0.082)	(0.406, 0.128)	1.388	(1.438, 0.385)

## 5. Results

Overall, predictions of subevents based on the movement and contact variables were moderate, and strongest for 6 frame bins using a linear classifier (as can be shown by Table 3). The data was also analyzed with the k-nearest neighbor method since this method converges to the MLE results. The k number of neighbors was incrementally increased by 50 to a group of 10001. Since the total amount of vectors created for the top bin-size of 6 for the entire data set was 22,862, this max k value would sufficiently capture the maximum D'. Analysis shows D<sub>1</sub>' increases dramatically then saturates at a value of about 1.400 with the value k around 950 nearest neighbors.

To assess the degree to which our 12 predictor variables can be represented by a smaller number of more basic factors, we performed a principle components analysis. First, the entire database is thinned out. The thinned data are the points that have the smaller distances from its nearest neighbor. The thinned data vectors are about half in number compared to the full data set. Fisher's linear discriminant analysis is applied to the thinned data as well as a principal component analysis (PCA). By calculating the eigenvalues and eigenvectors of that cross correlation method, the top 3 eigenvalues that caused the most variance in the data were identified. The eigenvectors corresponding to these eigenvalues were applied multiplied to the data and plotted. Four distinct groups could be seen, each with interesting points tightly clustered and non-interesting points trailing outward.

Table 5: Top 3 Feature Sets for Linear Analysis using Subject Jack Knife

Top 3 Feature Sets for Linear Analysis						
1 <sup>st</sup>	10	8	5	4	2	1
2 <sup>nd</sup>	8	7	5	4	2	1
3 <sup>rd</sup>	9	8	7	5	2	1

Table 6: Top 3 Feature Sets for Quadratic Analysis using Subject Jack Knife

Top 3 Feature Sets for Quadratic Analysis						
1 <sup>st</sup>	12	7	5	1	-	-
2 <sup>nd</sup>	12	11	7	5	1	-
3 <sup>rd</sup>	5	1	-	-	-	-

Table 7: Top 3 Feature Sets for Mahalanobis Analysis using Subject Jack Knife

Top 3 Feature Sets for Mahalanobis Analysis						
1 <sup>st</sup>	10	9	8	6	3	2
2 <sup>nd</sup>	10	9	8	7	6	2
3 <sup>rd</sup>	10	9	8	6	2	-

It was decided that further analysis of the feature combinations were to be examined. Exhaustive analyses of all combinations of features (up to 6 total features) were examined in predicting the key moments and the top 5 feature sets were calculated. The purpose of doing this analysis was to determine if a subset of the features used would provide as good or better results from the use of all features. Table 4 has the same format as Table 3 except showing the statistics of each technique's 1st – 3rd best feature combination results for the subject jack knife. Tables 5-7 show the top 3 feature combinations for each of the regression methods.

The most prominent features for the linear set are 1, 2, 5, and 8. These 4 features are found in all instances of the top 5 feature sets. Features 1 and 5 are the velocities of the two hands, feature 2 correspond to the binary feature for stopped hand 1 motion, and feature 8 corresponds to the amount of pixel change around hand 2. In the linear regression case, these features embody the information needed of the hand. Notice that the best feature sets also involve gaze information (feature 10) or rather the gaze toward an object.

The quadratic results are very similar except using even fewer features. Features 1 and 5 are necessary in every instance for the top 5 sets and feature 12 adds the gaze information needed for the top set of features in the subject jack knife analysis.

The Mahalanobis results show a high tendency toward the binary or discrete values. This measure focuses on the hand stop features (2 and 6) but also uses the gaze angular velocity (feature 9) and gaze toward objects (feature 10) to perform at its highest capacity.

The velocity features correlate with grasping since a majority of the grasps require a pause in the hand motion. This same reasoning also explains the correlation between the binary hands stopped features as well. Since grasps occur when the velocity of the hand is low and low gaze motions indicate focusing on an action, these findings support the findings of the Psychology Department analysis of a high correlation between significant moments

in the task with hand grasps and gaze.

## 6. Conclusion

The visual system designed by Tugcu has served as a basis for a new system that is able to segment human action sequences. We were successful in predicting subevents using a set of basic movement features for a wide range of tasks in which a model manipulated objects. A key finding was the consistency of this prediction, both across tasks and across performers, revealed by the small amount of variability in D' measures for the jackknife tests. We also systematically assessed the movement-subevent relationship for a range of bin sizes, and found that 6-frame bins were optimal.

These findings extend previous research on action parsing in three ways. First, we demonstrated the success of a specific action parsing technique. Second, we identified a set of basic movement features that align with subevents in human action. There are two proposals for the way that basic movement features might support parsing of human action [8, 9]. For one, dynamic intentional action may contain statistical structure. Certain motions may co-occur more frequently than others, because they are causally linked to achieving a goal (e.g., in cooking, the motion of slicing a vegetable may be preceded by the motion of grasping a knife, while slicing motions would only rarely be preceded by grasping a towel). Recent research has established that adults use these statistical regularities to group actions into units [10]. In addition to statistical regularities, there may be a predictable configuration of movements that occurs in the physical and temporal characteristics of bodily motion. These configural cues may reliably indicate segment boundaries. To act intentionally, we first locate relevant objects—typically resulting in head turns and associated changes in gaze orientation—and then contact them with our hands. Sensitivity to configural cues may assist with action segmentation if such cues reliably recruit observers' attention at the right moments. The present research provides suggestive evidence for the utility of such configural information in action parsing.

The third contribution this work makes involves isolating an ecologically important set of tasks. These face-to-face interactions are not the typical stimulus for machine vision approaches to action parsing, and yet they are a good candidate for the context in which human infants learn to parse and identify actions, and, ultimately, the intentions and goals of the people they interact with. Accordingly, a potentially interesting next step in this analysis would be to explore correlations between the information available based on a purely perceptual analysis of movements, and infants' ultimate responses to specific behavioral sequences. For example, a basic movement analysis such as the one presented here might serve as a basis for predicting infants' tracking of objects and events. If infants gradually develop a deeper understanding of the goal-directed nature of action during their first year of life, it might be possible to observe initially strong movement-based control over looking that gradually lessens as they complete their first year of life, and focus more on the deeper meaningful structure of actions than on perceptual patterns.

## 7. Acknowledgements

We would like to express our gratitude to the National Science Foundation for its very valuable support of this work under award 0325641 and award 0433653.

## 8. References

- [1] J. M. Zacks and B. Tversky, Event structure in perception and conception. *Psychological Bulletin*, 127, 3-21, 2001.
- [2] J. K. Aggarwal, Problems, ongoing research and future directions in motion research, *Machine Vision and Applications*, Vol. 14, No. 4, pp. 199-201, 2003.
- [3] J. K. Aggarwal, and Q. Cai, Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428-440, 1999.
- [4] H. Buxton, Learning and Understanding Dynamic Scene Activity: A Review. *Image and Vision Computing*, Vol. 21, No. 1, pp. 125-136, 2003.
- [5] J. Baird and D. A. Baldwin, Making Sense of Human Behavior: Acting Parsing and Intentional Inference. In *Intention and Intentionality*. Edited by B. F. Malle, L. J. Moses, and D. A. Baldwin. Cambridge, MA: MIT Press, April, pp. 193-206, 2001.
- [6] J. M. Zacks, B. Tversky, and G. Iyer, Perceiving, remembering and communicating structure in events. *The Journal of Experimental Psychology: General*, 130, 29-58, 2001.
- [7] J. M. Zacks, Using movement and intentions to understand simple events. *Cognitive Science*, 28, 979-1008, 2004.
- [8] J. A. Baird and D. A. Baldwin. Making sense of human behavior. In B. Malle, L. Moses, & D. Baldwin (Eds.). *Intentions and Intentionality: Foundations of social cognition* (pp. 193-206). Cambridge, MA: MIT Press, 2001.
- [9] M. M. Saylor and D. A. Baldwin, D. A. Action analysis and change blindness: Possible links. In D. T. Levin (Ed.), *Thinking and Seeing: Visual Metacognition in Adults and Children* (pp. 37-57). Cambridge, MA: MIT Press, 2004.
- [10] D. Baldwin, A. Andersson, J. Saffran, and M. Meyer (in press), Segmenting dynamic human via statistical structure. *Cognition*.
- [11] C. Yu, and D. Ballard, Learning to Recognize Human Action Sequences, *IEEE International Conference on Development and Learning (ICDL'02)*, Cambridge, MA, , pp. 28-34, 2002.
- [12] A. Madabhushi, and J. K. Aggarwal, Using Head Movement to Recognize Activity, *International Conference on Pattern Recognition*, Vol IV, 698-701, 2000
- [13] S. B. Kang, and K. Ikeuchi, Determination of Motion Breakpoints in a Task Sequence from Human Hand Motion. CRA94, 551-556, 1994.
- [14] M. Tugcu, A Computational Neuroscience Model with Application to Robot Perceptual Learning, Ph.D. Dissertation, Vanderbilt University, 2007
- [15] M. Tugcu, X. Wang, J. E. Hunter, J. Phillips, D. Noelle, and D. M. Wilkes, A computational Neuroscience model of working memory with application to robot perceptual learning, *Third IASTED International Conference on Computational Intelligence (CI)*, Banff, Alberta, Canada, 2007.
- [16] J. E. Hunter, Human Motion Segmentation and Object Recognition using Fuzzy Rules. *Proceedings of 14th Annual IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2005)*, Nashville, TN, pp 210-216, 2005.
- [17] D. T. Levin, J. Hunter, D. M. Wilkes, C. Heaton, and M. M. Saylor, Specifying the looking and reaching actions that predict breakpoint judgments, *Manuscript in Preparation*, 2008.