6-2007

Objects in Action: An Approach for Combining Action Understanding and Object Perception

Abhinav Gupta
University of Maryland - College Park, gabhinav@andrew.cmu.edu

Larry S. Davis
University of Maryland - College Park

Published In
Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on , 1- 8.

Objects in Action: An Approach for Combining Action Understanding and Object Perception

Abhinav Gupta and Larry S. Davis
Department of Computer Science
University of Maryland
College Park, MD 20742
{agupta,lsd}@cs.umd.edu

Abstract

Analysis of videos of human-object interactions involves understanding human movements, locating and recognizing objects and observing the effects of human movements on those objects. While each of these can be conducted independently, recognition improves when interactions between these elements are considered. Motivated by psychological studies of human perception, we present a Bayesian approach which unifies the inference processes involved in object classification and localization, action understanding and perception of object reaction.

Traditional approaches for object classification and action understanding have relied on shape features and movement analysis respectively. By placing object classification and localization in a video interpretation framework, we can localize and classify objects which are either hard to localize due to clutter or hard to recognize due to lack of discriminative features. Similarly, by applying context on human movements from the objects on which these movements impinge and the effects of these movements, we can segment and recognize actions which are either too subtle to perceive or too hard to recognize using motion features alone.

1. Introduction

We describe a Bayesian approach to the joint recognition of objects and actions based on shape and motion. Consider two similarly shaped objects such as the spray bottle and the drinking bottle shown in Figure 1. It is difficult to discriminate between the two objects based on shape alone. However, they are functionally dissimilar, so contextual information from human interactions with them can provide functional information for recognition. However, similar human movements can convey different intentions, depending on the contextual information provided by the environment and the objects on which these movement impinge. For example, while the movement $< hand - waving >$ would indicate spraying if a person is holding a spray bottle, it would imply signaling if the person instead carried a road-sign or a flag. Therefore, action recognition requires contextual information from object perception.

Figure 1. Importance of interaction context in recognition of object. While the objects might be difficult to recognize using shape features alone, when interaction context is applied the object is easy to recognize.

Another important element in the perception of human interactions with objects is the effect of manipulation on objects, which we will refer to as "object reaction". While interaction movements might be too subtle to observe with computer vision, the effects of these movements can be used to provide information on functional properties of the object.

We present a computational approach for perception of human interactions with objects. The approach models the contextual relationships between four perceptual elements of human object interaction: object perception, reach motion, manipulation motion and object reaction. These relationships enforce spatial, temporal and functional constraints on object recognition and action understanding.

The significance of the approach is twofold: (1) Human actions and object reactions can be used to locate and recognize objects which might be difficult to locate or recognize otherwise. Human actions and object reactions can also be used to infer object properties, such as weight. (2) Object context and object reactions can be used to recognize actions which might otherwise be too similar to distinguish or too difficult to observe.

1.1. Psychological Evidence of Action/Object Interactions in Human Perception

Early psychological theories of human information processing regarded action and perception as two separate processes [15]. However, recent investigations have suggested the importance of action in perceiving and recognizing objects (especially manipulable objects like tools) [4]. The evidence for such theories comes from neuropsychological studies where even passive viewing of manipulable objects evokes cortical responses associated with motor processes.

With the discovery of *mirror neurons* [9, 25] in monkey, there has been renewed interest in studying the relationships between object recognition, action understanding and action execution [9, 20, 10]. With the same neurons involved in execution and perception, a link between object recognition and action understanding has been established [20] in humans. Gallese et. al [9] showed that movement analysis in humans depends on the presence of objects. The cortical responses for goal directed actions are different from the responses evoked when the same action is executed but without the presence of the object.

Recent studies in experimental psychology have also confirmed the role of object recognition in action understanding and vice-versa. Helbig et. al [11] show the role of action priming in object recognition and how recognition rates improve with action-priming. In another study, Bub et. al [3] investigated the role of object priming in action/gesture recognition. While passive viewing of an object did not lead to priming effects, priming was observed when humans were first asked to recognize the object and then recognize the action.

While most of this work suggests the existence of interaction between object and action perception in humans, they have not examined the nature of the interaction between action and object recognition. Vaina et. al [30] address this through the study of pantomimes. They ranked the properties of objects that can be estimated robustly by perception of pantomimes of human-object interaction. They discovered that the weight of an object is most robustly estimated, while size and shape are harder to estimate. In an another study, Bach et. al [1] proposed that when action involving objects are perceived, spatial and functional relations provide a context in which actions are judged.

1.2. Related Computational Approaches

Most current computational approaches for object recognition use local static features and machine learning. The features are typically based on shape and textural appearance [5, 17]. These recognition approaches may have difficulty in recognizing manipulable objects when there is a lack of discriminative features. As a result, there has been recent interest in using contextual information for object recognition. The performance of local recognition based approaches can be improved by modeling object-object [18] or object-scene relationships [28]. Torralba et. al used low level image cues [29] for providing context based on depth and viewpoint cues. Hoiem et. al [12] presented a unified approach for simultaneous estimation of object locations and scene geometry.

There has also been work on object recognition based on functional properties. The functional capabilities of objects are derived using characteristics of shape [24, 27], physics and motion [7]. These approaches have been limited by the lack of generic models that can map static shape to function.

Many approaches for action recognition use human dynamics [2]. While human dynamics do provide important clues for action recognition, they are not sufficient for recognition of activities which involve action on objects. Many human actions involve similar movements/dynamics but due to their context sensitive nature have different meanings. Vaina et. al [31] suggested that action comprehension requires understanding the goal of an action. The properties necessary for achieving the goal were called *Action Requirements*. These requirements are related to the compatibility of an object with human movements such as grasp.

There have been a few attempts to model the contextual relationship between object recognition and action understanding. Wilson et. al [32] presented parametric Hidden Markov Model (PHMM) for human action recognition. They indirectly model the effect of object properties on human actions. Davis et. al [6] presented an approach to estimate the weight of a bag carried by a person using cues from the dynamics of a walking person. Moore et. al [16] presented an approach for action recognition based on scene context derived from other objects in the scene. The scene context is also used to facilitate object recognition of new objects introduced in the scene. They did not address the contextual relationship that exists between recognition of the object and the action that acts on the same object. Kuniyoshi et. al [13] presented a neural network for recognition of *true actions*. The requirements for a *true action* included spatial and temporal relationships between object and movement patterns. Peursum et. al [22] studied the problem of object recognition based on interactions. Regions in an image were classified as belonging to a particular object based on the relative position of the region to the human skeleton and the class of action being performed. While the authors recognize the need to apply object context to differentiate similar movements, they assume all similar movements are part of some higher level activity that can be recognized using human dynamics alone. For example, they assume *picking up paper* can be differentiated from *picking up a cup* based on recognizing that a higher level activity such as *printing a document* is being conducted. This is, however, a restrictive approach for two reasons: (a) Actions like *picking* can occur independently too. (b) Recognition of higher level activities is itself a hard problem.

All of these approaches assume that either object recognition or action understanding can be solved independent of

the other. They only model a one-way interaction between them. We next present an approach which unifies the inference process involved in object recognition and localization, action understanding and perception of object reaction.

1.3. Overview of Our Approach

We identify three classes of human movements involved in interactions with manipulable objects that depend on the goal/intention of the movement. These movements are 1) Reaching for an object 2) Grasping an object and 3) Manipulating an object. These movements are ordered in time; manipulation is always preceded by grasping which is preceded by the reach movement[1].

We present a graphical Bayesian model for modeling human-object interactions. The nodes in the belief network correspond to object, reach motion, manipulation motion, object reaction and evidence related to each of these elements.

We consider the interactions between different nodes in the model. Reach movements enable object localization since there is a high probability of an object being present at the endpoint of the reach motion. Similarly, object recognition disables false positives in reach motion detection, since there should be an object present at the endpoint of reach motion (See Figure 2).

Reach motions help to identify the possible segments of video corresponding to manipulation of the object and determine the dominant hand. Manipulation movements provide contextual information about the type of object being acted on. Similarly, object class provides contextual information on possible interactions with them, depending on affordances and function (See Figure 3).

In many cases, similar interactions may produce visually different hand trajectories because of difference in properties of the object. Figure 4 shows the difference in interaction style for $< throw >$ manipulation of heavy and light objects. Therefore, differences in style of execution provide contextual information on properties of objects such as weight.

Object reaction to human action, such as pouring liquid from a carafe into a cup or pressing a button that activates a device, provides contextual information about the object class and the manipulation motion. Our approach combines all these types of evidence into a single video interpretation framework. In the next section, we present a probabilistic model for describing the relationship between different elements in human object interactions.

[1]Our experiments neglect the grasping motion since the hand movements are too subtle to be perceived at the resolution of typical video cameras when the whole body and context are imaged

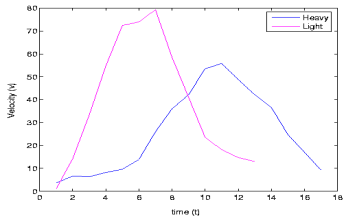


Figure 4. Differences in style based on object properties. In the case of heavier objects, the peak velocity is reached much later as compared to lighter objects. A study on throwing of objects of different weights using 3-mode factorization was reported in [19]

2. Modeling the Object Action Cycle

2.1. The Bayesian Network

Our goal is to simultaneously estimate object type, location, movement segments corresponding to reach movements, manipulation movements, type of manipulation movement and their effects on objects by taking advantage of the contextual information provided by each element to the others. We do this using the graphical model shown in Figure 5.

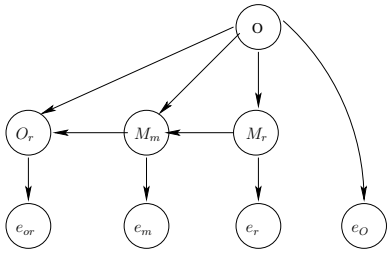
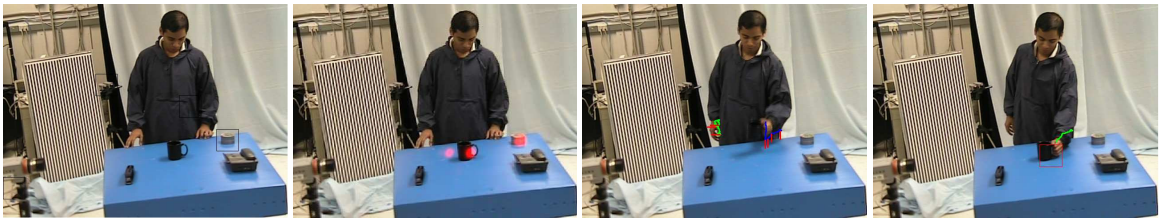


Figure 5. Underlying Graphical Model for Human Object Interaction.

In the graphical model, objects are denoted by O, reach motions by M_r, manipulation motions by M_m and object reactions by O_r. The video evidence is represented by $e = \{e_O, e_r, e_m, e_{or}\}$ where e_O represents object evidence, e_r and e_m represent reach and manipulation motion evidence and e_{or} represents object reaction evidence. Since only changes are observed for measuring object reaction, e_{or} is considered to be independent of O. Using Bayes rule and conditional independence relations, the joint probability distribution can be decomposed as[2]:

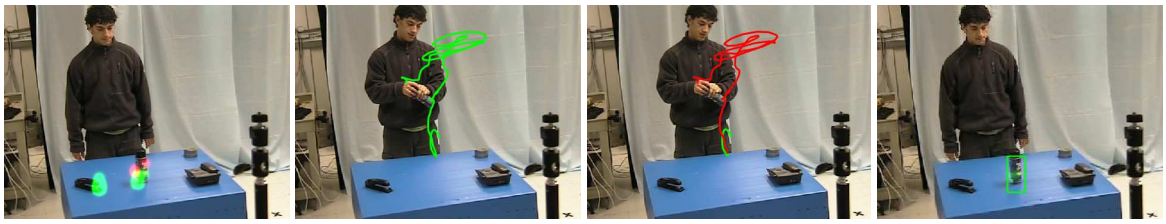
$$P(O, M_r, M_m, O_r|e) \propto P(O|e_O)P(M_r|O)P(M_r|e_r)$$
$$P(M_m|M_r, O)P(M_m|e_m)P(O_r|O, M_m)P(O_r|e_{or})$$

[2]All the variables are assumed to be uniformly distributed and hence $P(O)$, $P(M_r)$, $P(M_m)$, $P(O_r)$, $P(e_O)$, $P(e_r)$, $P(e_m)$ and $P(e_{or})$ are constant



(a) Original detector (b) Likelihood $P(O|e_O)$ (c) Reach Motion $P(M_r|e_r)$ (d) $P(O, M_r|e_O, e_r)$

Figure 2. Importance of contextual information involved in reach motions and object perception. (a) Object Detectors tend to miss some objects completely (b) Lowering the detection threshold can lead to false positives in detection (c) Reach Motion Segmentation also suffers from false positives (d) Joint probability distribution reduces the false positives in reach motion and false negatives in object detection.



(a) Likelihood $P(O|e_O)$ (b) Interaction Motion (c) Segmented Motion (d) Belief: $Bel(O)$

Figure 3. Importance of contextual information from interaction motion in object class resolution. In this experiment, object detectors for cups and spray were used. (a) The likelihood value of a pixel being the center of cup and spray bottle is shown by intensity of red and green respectively. (b) Hand trajectory for interaction motion (includes reach and manipulation). (c) The segmentation obtained. The green track shows the reach while the red track shows the manipulation.(d) Likelihood values after belief propagation. By using context from interaction with the object, it was inferred that since the object was subjected to a wave like motion, it is more likely a spray bottle.

2.2. Object Perception

Each object has an associated type which represents the class to which the object belongs. In addition to type, we estimate location and some physical properties.

The approach is independent of the specific object detection algorithm employed. We employ a variant of the histogram of oriented gradient(HOG) approach from [5, 33]. Our implementation uses a cascade of adaboost classifiers in which the weak classifiers are Fischer Linear Discriminants. This is a window based detector; windows are rejected at each cascade level and a window which passes all levels is classified as a possible object location.

Based on the sum of votes from the weak classifiers, for each cascade level, i, we compute the probability $P_i(w)$ of a window, w, containing the object. If a window were evaluated at all cascade levels, the probability of it containing an object would be $\prod_{i=1}^{L} P_i(w)$. However, for computational efficiency many windows are rejected at each stage of the cascade. The probability of such a window containing an object is computed based on the assumption that such windows would just exceed the detection threshold of the remaining stages of the cascade. Therefore, we also compute a threshold probability(Pt_i) for each cascade level i. This is the probability of that window containing an object whose adaboost score was at the rejection threshold. If a detector consists of L levels, but only the first l_w levels classify a window w as containing an object, then the overall likelihood is given by:

$$P(O = \{obj, w\}|e_O) = \prod_{i=1}^{l_w} P_i(w) \prod_{j=l_w+1}^{L} (Pt_j) \qquad (1)$$

2.3. Human Movements

2.3.1 Reach Motion

The reach motion is described by three parameters: the start time (t_s^r), the end time (t_e^r) and the 2D image location being

reached for (l_r). The velocity profile of a hand executing ballistic movements like reach or strike has a characteristic 'bell' shaped profile. Using features such as time to accelerate, peak velocity and magnitude of acceleration and deceleration, the likelihoods of reach movements can be computed from hand trajectories (See [23]).

However, there are many false positives because of errors in measuring hand trajectories. These false positives are removed using contextual information from object location. In the case of point mass objects, the distance between object location and the location being reached for should be zero. For a rigid body, the distance from the center of the object depends on the grasp location. We represent $P(M_r|O)$ using a normal function, $\mathcal{N}(|l_r l_o|, \mu, \sigma)$, where μ and σ are the average distance and variance of the distances in a training database between grasp locations and object centers.

2.3.2 Manipulation Motion

Manipulation motions also involve three parameters: start time (t_s^m), end time (t_e^m) and the type of manipulation motion/action (T_m) (such as answering a phone, drinking etc). We need to compute $P(M_m|e_m)$, the likelihood of a manipulation given the evidence from hand trajectories.

There are many methods for gesture recognition using hand trajectories [2]. The framework described above is independent of the specific action recognition approach employed. We use discrete HMM's for obtaining the likelihoods , $P(M_m|e_m)$.

We first obtain a temporal segmentation of the trajectory based on limb propulsion models. This segmentation is required for computing the discrete representation of manipulation motion and to find possible starting and ending times of the manipulation movement. There are two models for limb propulsion in human movements: ballistic and mass-spring models [26]. Ballistic movements involve impulsive propulsion of the limbs (acceleration towards the target followed by deceleration to stop the movement). In the mass-spring model, the limb is modelled as a mass connected to a springs. Therefore, the force is applied over a period of time.

Each manipulation motion is segmented into atomic segments based on the propulsion models described above. We use the segmentation algorithm described in [23]. The algorithm decomposes manipulation motion trajectories into ballastic and mass-spring motion segments. Each segment is then replaced by a discrete alphabet defined as the cross-product of type of propulsion(ballistic/mass-spring) and the hand locations at the end of the motion segments, represented with respect to the face. By using alphabets for atomic segments we transform a continuous observation into a discrete symbol sequence. This is used as input to obtain the likelihoods of different types of manipulation motion from their corresponding HMM's.

In addition to computing the likelihood, we need to compute the term $P(M_m|M_r, O)$. Manipulation motion is defined as a 3-tuple, $M_m = (t_s^m, t_e^m, T_m)$. The starting and ending times, t_s^m and t_e^m, depend on M_r but are independent of O. Similarly, the type of manipulation motion, T_m, depends on O but is independent of M_r[3]. Hence, we decompose the prior term as:

$$P(M_m|M_r, O) = P(t_s^m, t_e^m|M_r)P(T_m|O) \quad (2)$$

Assuming grasping takes negligible time, the time difference between the ending time of a reach motion and the starting time of a manipulation motion should be zero. We model $P(t_s^m, t_e^m|M_r)$ as a normal function $\mathcal{N}(t_s^m - t_e^r, 0, \sigma^t)$ where σ^t is the observed variance in the training dataset. $P(T_m = mtype|O = obj)$ is computed based on the number of occurrences of manipulation $mtype$ on object obj in our training dataset.

2.3.3 Hand Trajectories

The likelihood terms for reach and manipulation motion require computation of hand trajectories. To compute hand trajectories, we implemented a variant of [8] for estimating the 2D pose of the upper body. Figure 6 shows the results of the algorithm on few poses.

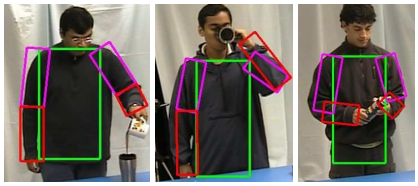


Figure 6. Results of Upper Body Pose Estimation Algorithm.

2.4. Object Reaction

In many cases, the interaction movement might be too subtle for effective measurement. In such cases, the result of interaction can provide context on object type and interaction involved. For example, consider the case of lighting a flashlight. The interaction involved is pressing a button, which is unlikely to be perceived using current computer vision approaches. However, the reaction/result of such an interaction, the change in illumination, is easy to detect. Similarly, the observation of object reaction can provide context on object properties. For example, the observation of the effect of pouring can help making the decision of whether a cup was empty or not.

[3]Type of manipulation also depends upon the direction of reach motion. This factor is, however, ignored in this paper

The parameters involved in object reaction are the time of reaction (t_{react}) and the type of reaction (T_{or}). However, measuring object reaction type is difficult. Mann et. al [14] presented an approach for understanding observations of interacting objects using Newtonian mechanics. However, such an approach can only be used to explain rigid body motions. Apart from rigid body interactions, the interactions which lead to changes in appearances using other forces such as electrical are also of interest to us.

We use the differences of appearance histograms around the hand location as a simple representation for reaction type classification. Such a representation is useful in recognizing reactions in which the appearance of the object at time of reaction, t_{react}, would be different than appearance at the start or the end of the interaction. Therefore, the two appearance histograms are subtracted and compared with the difference histograms in the training database to infer the likelihood of the type of reaction(T_{or}).

In addition, we need to compute the priors $P(O_r|M_m, O)$. Object reaction is defined by a 2-tuple, $O_r = (T_{or}, t_{react})$. Using the independence of the two variables:

$$P(O_r|M_m, O) = P(T_{or}|M_m, O)P(t_{react}|M_m, O) \quad (3)$$

The first term can be computed by counting the occurrences of T_{or} when the manipulation motion is of type $mtype$ and the object is of type obj. For modeling the second term, it was observed that the reaction time ratio, $r_r = \frac{t_{react} - t_s^m}{(t_e^m - t_s^m)}$, is generally constant for a combination of object and manipulation. Hence, we model the prior by a normal function $\mathcal{N}(r_r, \mu_r, \sigma_r)$ over the reaction-time ratio, where μ_r and σ_r are the mean and variance of reaction-time ratios in the training dataset.

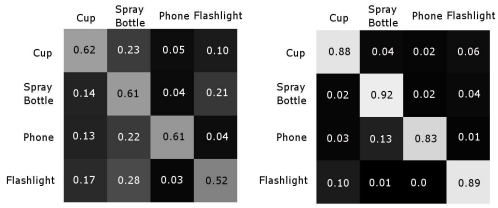
2.5. Training and Inference

We used Pearl's belief propagation algorithm [21] for inference. Training of the model requires training of a HOG based detector for all object classes and HMM models for all classes of interactions. Training for HOG based detector was done using images from various training datasets. HMM models were trained using a separate training dataset. Additionally our model requires co-occurence statistics of object-interaction-reaction combinations, distance between grasp location and object center, and reaction time ratios.

3. Experimental Evaluation

We evaluated our framework on a test dataset of 10 subjects performing 6 interactions with 4 objects. The objects in the test-dataset included cup, spray bottle, phone and flashlight. The interactions with these objects were: drinking from a cup, spraying from a spray bottle, answering a phone call, making a phone call, pouring from a cup and lighting the flashlight. In addition to the four objects on which the detector was trained, the scene contained other objects, like a stapler, to confuse the object detector.

Object Classification: Among the objects used, it is hard to discriminate the spray bottle, flashlight and cup because all three are cylindrical (See Figures 11(a),(b)). Furthermore, the spray bottle detector also fired for the handset of the cordless phone (See Figure 11(d)). Our approach was also able to detect and classify object of interest even in cluttered scenes (See Figure 11(c)). Figures 7(a) and 7(b) shows the likelihood confusion matrix for both the original object detector and the object detector in the human-object interaction framework. Using interaction context, the recognition rate of objects at the end of reach locations improved from 78.33% to 96.67%[4].



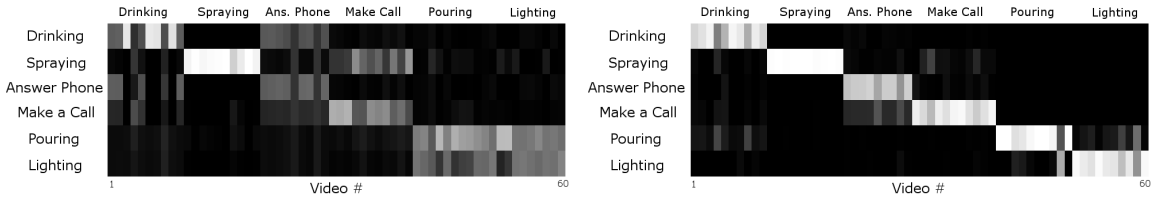
(a) HOG Detector (b) Using Framework

Figure 7. Object Likelihood Confusion Matrix: The i^{th} row depicts the expected likelihood values when i^{th} type of object is present.

Action Recognition: Of the six activities, it is very hard to discriminate between pouring and lighting on the basis of hand trajectories(See Figure 11(a) and (b)). While differentiating drinking from phone answering should be easy due to the differences in endpoint locations, there was still substantial confusion between the two due to errors in computation of hand trajectories. Figure 8(a) shows the likelihoods of actions that were obtained for all the videos using hand-dynamics alone. Figure 8(b) shows the confusion matrix when action recognition was conducted using our framework. The overall recognition rate increased from 76.67% to 93.34% when action was recognized using the contextual information from objects and object reactions.

Segmentation Errors: Apart from errors in classification, we also evaluated our framework with respect to segmentation of reach and manipulation motion. The segmentation error was the difference between the actual frame number and the computed frame number for the end of a reach motion. We obtained the ground truth for the data using manual labelling. Figure 9 shows the histogram of segmentation errors in the videos of the test dataset. It can be

[4]The recognition rate depicts the correct classification of localized object into one of the five classes: background, cup, spray-bottle, phone and flashlight



(a) HMM based Action Recognition

(b) HMM based recognition in Interaction Context

Figure 8. Comparison of Action Likelihoods without and with contextual information. Each Column represents the normalized likelihood values for six possible actions.

seen that 90% of detections were within 3 frames of actual end-frames of reach motion.

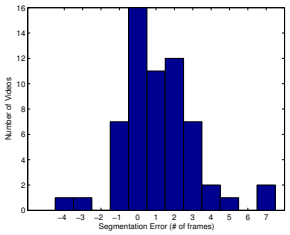


Figure 9. Segmentation Error Histogram

Object Properties: The first and second-order derivatives of the velocity profiles at the start and end of ballistic motion segments during manipulation are used as a feature set for classification of 'heavy/light' objects. The object used in the experiment was box (heavy/light) and the interaction was displacing the box from one end of table to another. Figure 10 shows the two derivates plotted for the training dataset. We achieved a classification accuracy of 89.58% for a linear classifier(LDA) using leave one-out cross-validation approach.

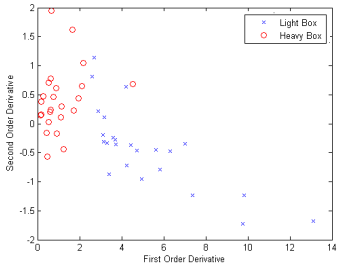


Figure 10. The first and second order derivatives of velocity at the start and end of ballistic motion.

4. Conclusion

Recent studies related to human information processing have confirmed the role of object recognition in action understanding and vice-versa. Motivated by such studies, we presented an approach to combine the inference process in object recognition and action understanding. The approach uses a probabilistic model to represent the elements of human-object interaction: object identity, reach motion, manipulation motion and object reaction. Using context from object type and object reaction, the model recognizes actions which are either too subtle to perceive or too similar to discriminate. Therefore, by enforcing global coherence between object type, action type and object reaction, we can improve the recognition performance of each element substantially.

5. Acknowledgement

The research was supported by Homeland Security Advanced Research Project Agency Award N0001405C0218.

References

[1] P. Bach, G. Knoblich, T. Gunter, A. Friederici, and W. Prinz. Action comprehension: Deriving spatial and functional relations. *J. Exp. Psych. Human Perception and Performance*, 31. 2

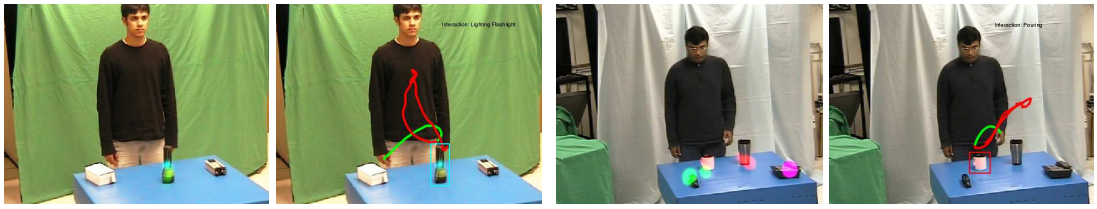
[2] A. Bobick and A. Wilson. A state-based approach to the representation and recognition of gesture. *IEEE PAMI*, 19(12):1325–1337, 1997. 2, 5

[3] D. Bub and M. Masson. Gestural knowledge evoked by objects as part of conceptual representations. *Aphasiology*, 20:1112–1124, 2006. 2

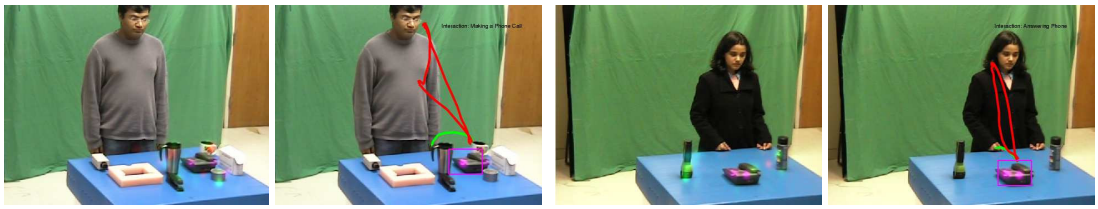
[4] L. L. Chao and A. Martin. Representation of manipulable man-made objects in dorsal stream. *NeuroImage*, 12:478–484, 2000. 2

[5] N. Dalal and B. Triggs. Histogram of oriented gradients for fast human detection. In *CVPR*, 2005. 2, 4

[6] J. Davis, H. Gao, and V. Kannappan. A three-mode expressive feature model of action effort. In *IEEE Worskshop on Motion and Video Computing*, 2002. 2



(a) HOG Detector (a) HOG in Framework (b) HOG Detector (b) HOG in Framework



(c) HOG Detector (c) HOG in Framework (d) HOG Detector (d) HOG in Framework

Figure 11. Results of object detection in the human-object interaction framework. The likelihoods of the centers of different objects are shown in different colors. The colors red, green, cyan and magenta show the likelihoods of cup, spray bottle, flashlight and phone respectively. (a) A flashlight is often confused as spray bottle by the HOG detector. However, when context from the framework is used there is no confusion. (b) Similarly a cup is often confused with a wide spray bottle. (c) Our detector can find and classify objects in clutter. (d) A spray bottle detector often fires at the handset of cordless phones due to the presence of parallel lines. However, such confusion can be removed using our framework.

[7] Z. Duric, J. Fayman, and E. Rivlin. Function from motion. *IEEE PAMI*, 18(6):579–591, 1996. 2

[8] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2003. 5

[9] V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti. Action recognition in premotor cortex. *Brain*, 1996. 2

[10] G. Guerra and Y. Aloimonos. Discovering a language for human activity. In *AAAI Work. on Anticipation in Cognitive Systems*, 2005. 2

[11] H. B. Helbig, M. Graf, and M. Kiefer. The role of action representation in visual object. *Experimental Brain Research*, 174:221–228, 2006. 2

[12] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006. 2

[13] Y. Kuniyoshi and M. Shimozaki. A self-organizing neural model for context based action recognition. In *IEEE EMBS Conference on Neural Engineering*, 2003. 2

[14] R. Mann, A. Jepson, and J. Siskind. The computational perception of scene dynamics. *CVIU*, 65(2):113–128, 1997. 6

[15] A. D. Milner and M. A. Goodale. *The Visual Brain in Action*. Oxford University Press, 1995. 2

[16] D. Moore, I. Essa, and M. Hayes. Exploiting human action and object context for recognition tasks. In *ICCV*, 1999. 2

[17] H. Murase and N. S.K. Learning object models from appearance. In *National Conf. on Artificial Intelligence*, 1993. 2

[18] K. Murphy, A. Torralba, and W. Freeman. Graphical model for scenes and objects. In *NIPS*, 2003. 2

[19] R. Neal, C. Snyder, and P. Kroonenberg. Individual differences and segment interaction in throwing. *Human Movement Sci.*, 10, 1991. 3

[20] K. Nelissen, G. Luppino, W. Vanduffel, G. Rizzolatti, and G. Orban. Observing others: Multiple action representation in frontal lobe. *SCIENCE*, 310:332–336, 2005. 2

[21] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Network and Plausible Inference*. Morgan Kaufmann, 1988. 6

[22] P. Peursum, G. West, and S. Venkatesh. Combining image regions and human activity for indirect object recognition in indoor wide-angle views. In *ICCV*, 2005. 2

[23] V. Prasad, V. Kellokompu, and L. Davis. Ballistic hand movements. In *AMDO*, 2006. 5

[24] E. Rivlin, S. Dickinson, and A. Rosenfeld. Recognition by functional parts. In *CVPR*, 1994. 2

[25] L. Rizzolatti, G.and Fadiga, L. Fogassi, and V. Gallese. Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3:131–141, 1996. 2

[26] I. Smyth and M. Wing. *The Psychology of Human Movement*. The Psychology of Human Movement, 1984. 5

[27] L. Stark and K. Bowyer. Generic recognition through qualitative reasoning about 3d shape and object function. In *CVPR*, 1991. 2

[28] E. Sudderth, A. Torralba, W. Freeman, and A. Wilsky. Learning hierarchical models of scenes, objects and parts. In *ICCV*, 2005. 2

[29] A. Torralba and P. Sinha. Statistical context priming for object detection. In *ICCV*, 2001. 2

[30] L. Vaina, H. Goodglass, and L. Daltroy. Influence of object use from pantomimed actions by aphasics and patients with right hemisphere lesions. *Synthese*, 104:43–57, 1995. 2

[31] L. Vaina and M. Jaulent. Object structure and action requirements: A compatibility model for functional recognition. *Int. Journal of Intelligent Systems*, 6:313–336, 1991. 2

[32] A. Wilson and A. Bobick. Parametric hidden markov models for gesture recognition. *IEEE PAMI*, 1999. 2

[33] Q. Zhu, S. Avidan, M. Ye, and K. Cheng. Fast human detection using a cascade of histograms of oriented gradients. In *CVPR*, 2006. 4