Observing Human-Object Interactions: Using Spatial and Functional Compatibility for Recognition

Abhinav Gupta, *Member*, *IEEE*, Aniruddha Kembhavi, *Member*, *IEEE*, and
Larry S. Davis, *Fellow*, *IEEE*

Abstract—Interpretation of images and videos containing humans interacting with different objects is a daunting task. It involves understanding scene/event, analyzing human movements, recognizing manipulable objects, and observing the effect of the human movement on those objects. While each of these perceptual tasks can be conducted independently, recognition rate improves when interactions between them are considered. Motivated by psychological studies of human perception, we present a Bayesian approach which integrates various perceptual tasks involved in understanding human-object interactions. Previous approaches to object and action recognition rely on static shape/appearance feature matching and motion analysis, respectively. Our approach goes beyond these traditional approaches and applies spatial and functional constraints on each of the perceptual elements for coherent semantic interpretation. Such constraints allow us to recognize objects and actions when the appearances are not discriminative enough. We also demonstrate the use of such constraints in recognition of actions from static images without using any motion information.

Index Terms—Action recognition, object recognition, functional recognition.

1 Introduction

Understanding human-object interactions require integrating various perceptual elements. We present a Bayesian approach for the interpretation of human-object interactions, that integrates information from perceptual tasks such as scene analysis, human motion/pose estimation,[1] manipulable object detection, and "object reaction" determination.[2] While each of these tasks can be conducted independently, recognition rates improve when we integrate information from different perceptual analysis and also consider spatial and functional constraints.

Integrating information from different perceptual analyses enables us to form a coherent semantic interpretation of human-object interactions. Such an interpretation not only supports recognizing the interactions, but also the objects involved in those interactions and the effect of those interactions on those objects.

Interactions between different perceptual analyses allow us to recognize actions and objects when appearances are not discriminative enough. Consider two objects, such as the spray bottle and a drinking bottle shown in Fig. 1. These objects are similar in appearance and shape, but have different functionality. Due to their functional dissimilarity, people's interaction with these objects provides context for their recognition. Similarly, two similar human movements/poses can serve different purposes depending on the context in which they occur. For example, the poses of the humans shown in Fig. 2 are similar, but, due to the difference in context, the first action is inferred to be running and the second action to be kicking.

Another important element in the interpretation of human-object interactions is the effect of manipulation on objects. When interaction movements are too subtle to observe using computer vision, the effects of these movements can provide information on functional properties of the object. For example, when lighting a flashlight, recognizing the pressing of a button might be very difficult. However, the resulting illumination change can be used to infer the manipulation.

We present two computational models for the interpretation of human-object interactions in videos and static images, respectively. Our approach combines action recognition and object recognition in an integrated framework, and allows us to apply spatial and functional constraints for recognition. The significance of our paper is threefold: 1) Human actions and object reactions are used to locate and recognize objects which might be difficult to locate or recognize otherwise. 2) Object context and object reactions are used to recognize

[1] Recognition of action in static images is based on "implied" motion. "Implied" motion refers to the dynamic information implicit in the static image [26]. The inference of action from static images depends on implied motion, which itself depends on the phase of the action [27], [53]. This indicates that human pose provides important cues for action recognition in static images.

[2] Object reaction is the effect of manipulation of an object by human actor.

- *A. Gupta and L.S. Davis are with the Department of Computer Science, AV Williams Bldg., University of Maryland-College Park, College Park, MD 20742. E-mail: {agupta, lsd}@cs.umd.edu.*
- *A. Kembhavi is with the Department of Electrical and Computer Engineering, 3364 AV Williams Bldg., University of Maryland-College Park, College Park, MD 20742. E-mail: anikem@umd.edu.*

Manuscript received 14 Aug. 2008; revised 22 Dec. 2008; accepted 7 Apr. 2009; published online 9 July 2009.
Recommended for acceptance by Q. Ji, A. Torralba, T. Huang, E. Sudderth, and J. Luo.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMISI-2008-08-0503.
Digital Object Identifier no. 10.1109/TPAMI.2009.83.

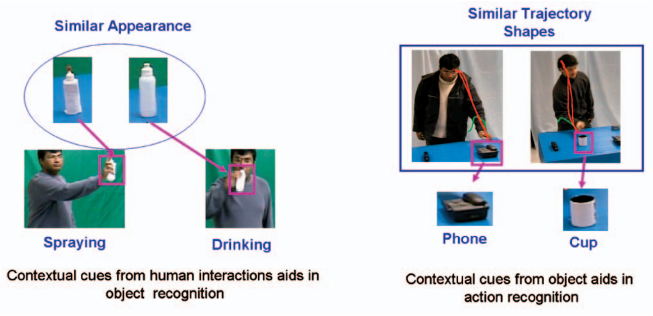

Similar Appearance
Spraying
Drinking
Contextual cues from human interactions aids in object recognition
Similar Trajectory Shapes
Phone
Cup
Contextual cues from object aids in action recognition

Fig. 1. Importance of interaction context in recognition of object and vice versa. While the objects might be difficult to recognize using shape features alone, when interaction context is applied the object is easy to recognize. Similarly, two actions might have similar dynamics and trajectories. It is difficult to differentiate between two actions based on the shape of trajectories. However, when cues from object are used in conjunction with cues from human dynamics, it is easy to differentiate between two actions.

actions which might otherwise be too similar to distinguish or too difficult to observe. In some cases, such as in recognition of actions from static images, there is no dynamic information; however, contextual information can be used in such cases for recognition. 3) We provide an approach for recognition of actions from "static" images. The extraction of "dynamic information" from static images has been well studied in the fields of psychology and neuroscience, but has not been investigated by the computer vision community.

2 Related Work

2.1 Psychological Studies

Our work is motivated by psychological studies of human information processing. With the discovery of mirror neurons in monkeys, there has been a renewed interest in studying the relationships between object recognition, action understanding, and action execution [38], [15], [16]. With the same neurons involved in execution and perception, a link between object recognition and action understanding has been established [38] in humans. Gallese et al. [15] 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. In another study, Frey et al. [25] showed that human inferior frontal cortex responds to static pictures of human-object interactions. The response was only observed in the presence of congruent poses and objects, suggesting that human poses are evaluated in the context of objects. On the other hand, the importance of action in perceiving and recognizing objects (especially manipulable objects like tools) has been shown [8].

Recent studies in experimental psychology have also confirmed the role of object recognition in action understanding and vice versa. Helbig et al. [23] show the role of action priming in object recognition and how recognition rates improve with action priming. Recognition rates of target objects were higher when the priming object was used in a similar action as the target object. In another study, Bub and Masson [7] investigated the role of object priming in static 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 image of a related hand gesture. In a recent study, Bach et al. [2] showed that when actions involving objects are perceived, spatial and functional relations provide context in which these actions are judged. These studies suggest that humans perceive implied motion from static poses under object and scene context.

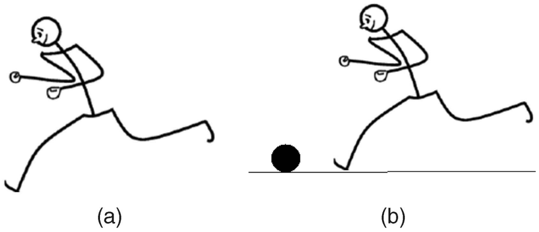


Fig. 2. Action recognition from static images requires contextual information. The same poses can have different meanings based on the context. (a) Running and (b) kicking.

While most of this work suggests interactions between object and action perception in humans, they have not examined the nature of the interaction between action and object recognition. Vaina and Jaulent [54] 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.

2.2 Computational Approaches

There has been a very large body of work carried out in both, object recognition and action recognition. Most approaches, however, address one or both of these problems, independent of the other.

Computational approaches for object recognition typically use local static features, based on shape and textural appearance [9], [34], [21]. Berg and Malik [3] proposed the 'geometric blur' feature that is robust under affine distortions. Bosch et al. [6] proposed the Pyramidal Histogram of Oriented Gradients (PHOG) feature and the Pyramidal Histogram of Visual Words (PHOW) feature to represent local image shape and its spatial layout. Wu and Nevatia [58] proposed a set of silhouette-oriented features, called edgelet features, which were learned in a boosting framework to detect humans. Such approaches work well for detecting articulated/rigid objects, but encounter difficulties in recognizing manipulable objects due to the lack of discriminative power in these features. Todorovic and Ahuja [51] model object categories as characteristic configurations of parts that are themselves simpler subcategories, allowing them to cope better with nonrigid objects. However, like all appearance-based approaches, they still cannot deal with the many real-world objects that are similar in appearance but dissimilar in functionality. Functional properties of objects have also been used for object recognition. Functional capabilities of objects are derived from shape [45], [48], physics, and motion [11]. These approaches are limited by the lack of generic models that can map static shape to function. 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 [35], [19] and object-scene relationships [49], [36]. Torralba and Sinha used low-level image cues [52] for providing context based on depth and viewpoint cues. Hoiem et al. [24] presented a unified approach for simultaneous estimation of object locations and scene geometry. Rabinovich et al. [43] proposed incorporating semantic object context as a postprocessing step to any object category recognition system using a conditional random field (CRF) framework.

There are a wide range of approaches to human action recognition [46], [32], [22]. Analyzing human dynamics from image sequences of actions is a common theme to many of these approaches [5], [61], [44], [50]. While human dynamics provides 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 and Jaulent [54] suggested that action comprehension requires understanding the goal of an action. The properties necessary for achieving the goal were called Action Requirements and are related to the compatibility of an object with human movements such as grasps.

Compared to the large body of work carried out in human action recognition from video sequences, there has been little work on recognition from single images. Wang et al. [56] presented an approach for discovery of action classes from static images using the shape of humans described by shape context histograms. Li et al. [29] tackled a different, but related, problem of event recognition from static images. They presented an approach to combine scene categorization and object recognition for performing event classification such as badminton and tennis. The problem of action recognition from static images is one level lower in the action hierarchy and corresponds to "verb" recognition in the hierarchy suggested by Nagel [37].

Attempts have been made before, to model the contextual relationship between object and action recognition. Wilson and Bobick [57] introduced parametric Hidden Markov Model (PHMM) for human action recognition. They indirectly model the effect of object properties on human actions. Davis et al. [10] presented an approach to estimate the weight of a bag carried by a person using cues from the dynamics of a walking person. Gupta et al. [17] presented an approach to estimate human pose using the contextual features from the objects being used in an activity. Moore et al. [33] conducted 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. Kuniyoshi and Shimozaki [28] describe a neural network for the recognition of "true" actions. The requirements for a "true" action included spatial and temporal relationships between object and movement patterns. Peursum et al. [41] 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. All of the above work, models only one of the possible interactions between two perceptual elements. Either they try to model the dependence of object recognition on human actions or vice versa. This assumes that one of the problems can be solved independent of the other, and the information from one can be used to aid in recognition of the other.

Our previous preliminary work [18] modeled the two-way interactions between human actions and object perception. We presented a Bayesian model for simultaneous recognition of human actions and manipulable objects. Following our work, several recent papers have modeled the action-object cycle. Wu et al. [60] recognized activities based on detecting and analyzing the sequence of objects manipulated by the user, using a dynamic Bayesian network model. They combined information from RFID and video data to jointly infer the most likely activity and objects in the scene. Filipovych et al. [14] proposed a probabilistic graphical model of primitive actor-object interactions that combines information about the interactions' dynamics, and actor-object static appearances and spatial configurations. However, none of these approaches can be easily extended to action recognition from static images.

3 Video Interpretation Framework

We first describe a computational model for interpretation of human-object interaction videos. We identify three classes of human movements involved in interactions with manipulable objects. These movements are 1) reaching for an object of interest, 2) grasping the object, and 3) manipulating the object. These movements are ordered in time. The reach movement is followed by grasping which precedes manipulation. In our model, we ignore 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.

3.1 Overview

We present a graphical model for modeling human-object interactions. The nodes in the model correspond to the perceptual analyses corresponding to the recognition of objects, reach motions, manipulation motions, and object reactions. The edges in the graphical model represent the interactions/dependencies between different nodes.

Reach movements enable object localization since there is a high probability of an object being present at the endpoint of a reach motion. Similarly, object recognition disables false positives in reach motion detection, since there should be an object present at the endpoint of a reach motion (see Fig. 3). Reach motions also help to identify the possible segments of video corresponding to manipulation of the object, since manipulation motion is preceded by reach motion. Manipulation movements provide contextual information about the type of object being acted on and object class provides contextual information on possible interactions with them, depending on affordances and function. Therefore, a joint estimation of the two perceptual elements provides better estimates as compared to the case when the two are estimated independently (see Fig. 4).

The object reaction to a human action, such as pouring liquid from a carafe into a cup or pressing a button that activates a device, provides contextual information about

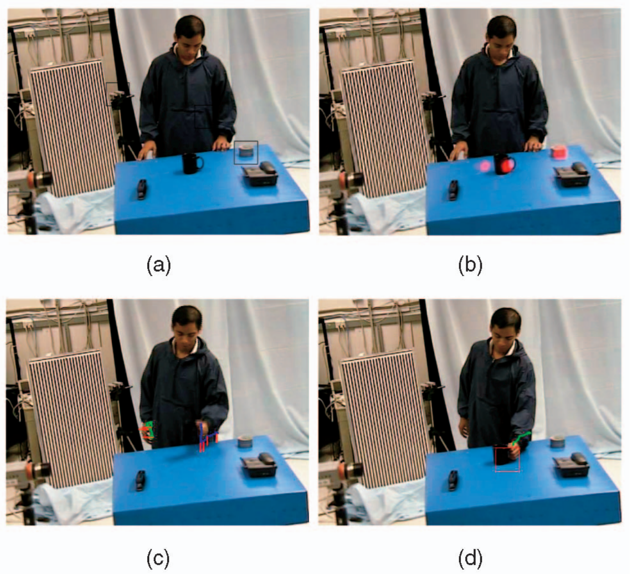


Fig. 3. Importance of contextual information involved in reach motions and object perception. (a) Object Detectors tend to miss some objects completely (original detector). (b) Lowering the detection threshold can lead to false positives in detection. The likelihood of a pixel being the center of the cup is shown by intensity of red (likelihood $P(O|e_O)$). (c) Reach Motion Segmentation also suffers from false positives (reach $P(M_r|e_r)$). The trajectories are shown in green and blue with possible endpoints of reach motion shown in red. (d) Joint probability distribution reduces the false positives in reach motion and false negatives in object detection ($P(O, M_r|e_O, e_r)$).

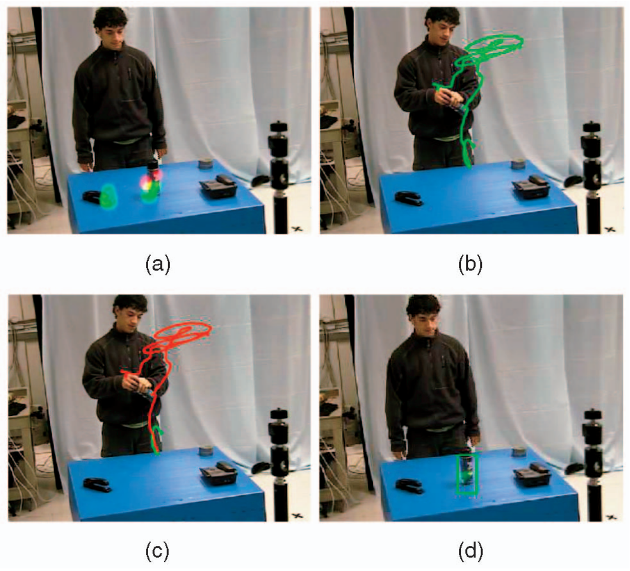


Fig. 4. Importance of contextual information from interaction motion in object class resolution. In this experiment, object detectors for cups and spray were used. (a) The likelihood value of a pixel being the center of cup and spray bottle is shown by intensity of red and green, respectively (likelihood $P(O|e_O)$). (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 (belief: $Bel(O)$). 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.

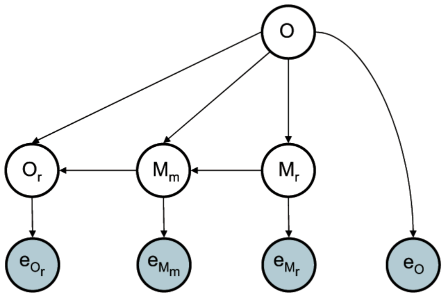


Fig. 5. Underlying Graphical Model for Human-Object Interaction. The observed and hidden nodes are shown in gray and white, respectively.

the object class and the manipulation motion. Our approach combines all these types of evidences 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.

3.2 Our Bayesian Model

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 Fig. 5.

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. Using Bayes rule and conditional independence relations, the joint probability distribution can be decomposed as[3]

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

We use loopy belief propagation algorithm [40] for inference over the graphical model. In the next few sections we discuss how to compute each of these terms. Section 3.3 discusses how to compute the object likelihoods $P(O|e_O)$. In Section 3.4.1 we explain the computation of reach motion likelihood, $P(M_r|e_r)$, and the contextual term $P(M_r|O)$. This is followed by a discussion on computation of manipulation motion likelihood, $P(M_m|e_m)$, and the term $P(M_m|M_r, O)$ in Section 3.4.2. In Section 3.5, we discuss the object reaction likelihood $P(O_r|e_{or})$ and the prior term, $P(O_r|O, M_m)$.

3.3 Object Perception

The object node in the graphical model represents the random variable O. We want to estimate the likelihood of the type of object and the location of the object. While our approach is independent of the likelihood model, we employ a variant of the histogram of oriented gradient (HOG) approach from [9], [62].[4] Our implementation uses a

[3] All of 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.

[4] We use linear gradient voting with nine orientation bins in 0-180; 12×12 pixel blocks of four 6×6 pixel cells.

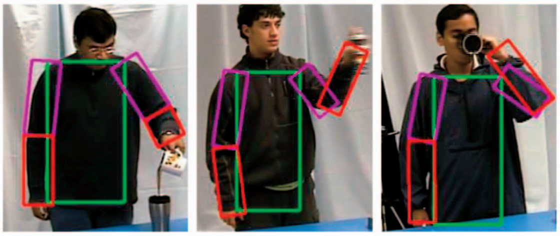


Fig. 6. Results of upper body pose estimation algorithm.

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.[5] 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 approximated by

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

3.4 Motion Analysis

We need to estimate the likelihoods of reach motion and manipulation motion. Our likelihood model is based on hand trajectories and, therefore, requires estimation of endpoints (hands in case of upper body pose estimation) in each frame. While one can use independent models for tracking the two hands, this could lead to identity exchange and lost tracks during occlusions. Instead, we pose the problem as upper body pose estimation. We implemented a variant of [12] for estimating the 2D pose of the upper body. In our implementation, we use an edge [20] and silhouette-based likelihood representation for body parts. We also use detection results of hands based on shape and appearance features and a temporal tracking framework where smoothness constraints are employed to provide priors. Fig. 6 shows the results of the algorithm on few poses.

3.4.1 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). We want to estimate the likelihood of reach motion ($M_r = (t_s^r, t_e^r, l_r)$) given the hand trajectories. An approach for detecting reach motion was presented in [42]. It is based on psychological studies which indicate that the hand movements corresponding to ballistic motion such as reach/strike have distinct "bell" shaped velocity profiles [31], [47] (see Fig. 7). There is an initial impulse accelerating the hand/foot toward the target, followed by a decelerating impulse to stop the movement. There is no mid-course correction. 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.

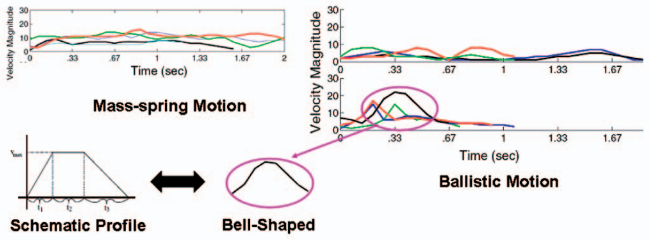


Fig. 7. Plot on the left shows velocity profiles of some mass-spring motions and the figure on the right shows some ballistic hand movements. The velocity remains low and constant during mass-spring movements. It reduces to zero only at the end of the movement. On the other hand, hand movements corresponding to ballistic motion such as reach/strike have distinct "bell" shapes.

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.

3.4.2 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. While one can use any gesture recognition approaches based on hand trajectories to estimate the likelihood, we use a simple discrete HMM-based approach to estimate it.

We need to first compute a discrete representation of the manipulation motion. Toward this end, we obtain a temporal segmentation of the trajectory based on a limb propulsion model. An approach for such a segmentation was presented in [42]. There are two models for limb propulsion in human movements: ballistic and mass-spring models [47]. Ballistic movements, discussed previously, involve impulsive propulsion of the limbs (acceleration toward the target followed by deceleration to stop the movement). In the mass-spring model, the limb is modeled as a mass connected to a spring. Therefore, the force is applied over a period of time.

To obtain the temporal segmentation of a velocity profile, it is observed that the endpoints of each ballistic segment

[5] Our experiments indicate that in many cases locations rejected by a classifier in the cascade are true object locations and selected by our framework.

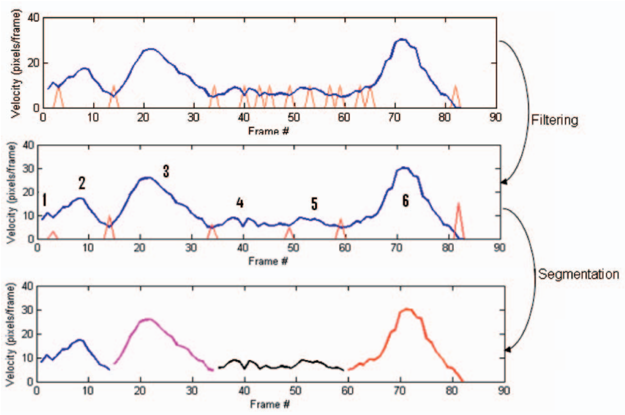


Fig. 8. Segmentation Procedure: The first graph shows the local minima of velocity profile. These local minima are classified into possible endpoints of each segment. This is followed by a maximum likelihood approach to obtain the segmentation of the velocity profile.

correspond to a local minima in the velocity profile. However, due to noise all local minimas are not the endpoints of atomic segments. Therefore, the segmentation problem is treated as that of classifying the points of local minima as being segmentation boundaries or not. The classification is based on features such as accelerating impulse and its duration. Given confidence values for each time instant to be a starting, ending, or negligible movement, we compute the most likely segmentation of the velocity profile using Maximum Likelihood (see Fig. 8).

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 HMMs.

In addition to computing the likelihood, we need to compute the term $P(M_m|M_r,O)$. Manipulation motion is defined as a three-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.[6] Hence, we decompose the prior term as

$$P(M_m|M_r,O)=P\left(t_s^m,t_e^m|M_r\right)P(T_m|O). \tag{2}$$

Assuming that 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 distribution $\mathcal{N}(t_s^m-t_e^r,0,\sigma^t)$, where σ^t is the observed variance in the training data set. $P(T_m=mtype|O=obj)$ is computed based on the number of occurrences of manipulation $mtype$ on object obj in our training data set.

3.5 Object Reactions

Object reaction is defined as the effect of manipulation on the object. In many cases, manipulation movements might be too subtle to observe using computer vision approaches. For example, in the case of a flashlight, the manipulation involved is pressing a button. While the manipulation motion is hard to detect, the effect of such manipulation (the lighting of the flashlight) 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.

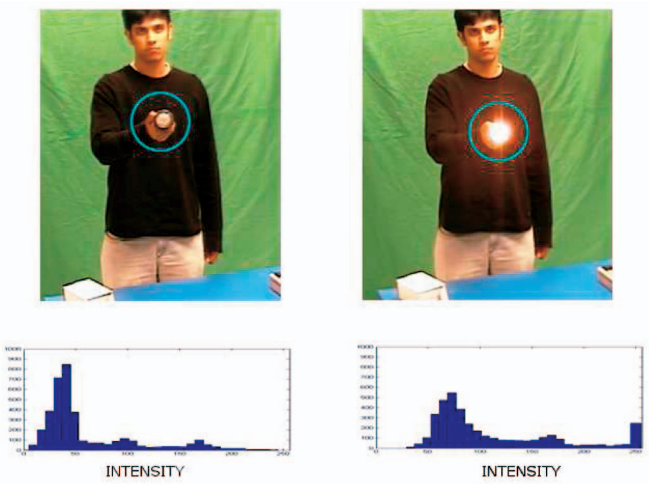


Fig. 9. Using appearance histograms around hand to estimate $P(O_r|e_{or})$. In the case above, illumination change due to flashlight causes the change in intensity histogram.

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. [30] presented an approach for understanding observations of interacting objects using Newtonian mechanics. This 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 (eight bins each in RGB space) around the hand location as a simple representation for reaction type classification (see Fig. 9). Such a representation is useful in recognizing reactions in which the appearance of the object at the 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 two-tuple, $O_r=(T_{or},t_{react})$. Using the independence of the two variables:

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

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, we 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

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

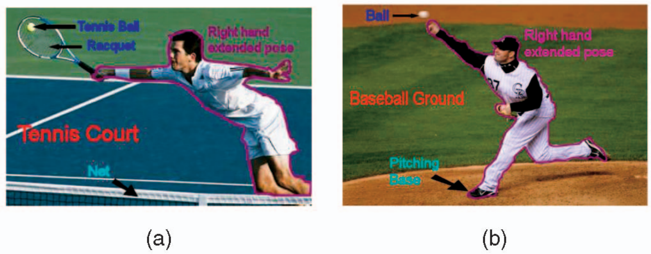


Fig. 10. Examples depicting the reasoning process in action inference from static images. The labels in red are the result of a scene categorization process, cyan labels and blue labels represent scene and manipulable objects, respectively, and the magenta label is the result of a pose estimation algorithm. For understanding actions from static images, information is combined from all components. While the pose is similar in both scenes, the presence of the racket and tennis ball, along with the tennis court environment suggests that the first picture is a "tennis-forehand" while the second is baseball pitching due to the presence of the pitching area and the baseball field. (a) Tennis-forehand. (b) Baseball pitching.

$\mathcal{N}(r_r, \mu_r, \sigma_r)$ over the reaction-time ratio, where μ_r and σ_r are the mean and variance of reaction-time ratios in the training data set.

4 Recognizing Interactions from Static Images

While action recognition requires motion information, in the case of static images, contextual information can be used in conjunction with human pose to infer action. Figs. 10a and 10b show examples of reasoning involved in inference of actions from a static image. In both cases, pose alone does not provide sufficient information for identifying the action. However, when considered in the context of the scene and the objects being manipulated, the pose become informative of the goals and the action.

Relevant objects in the scene generally bear both a semantic[7] and spatial relationship with humans and their poses. For example, in a defensive stance of a cricket batsman, the bat is facing down and is generally below or in level with the person's centroid. Similarly, the location of the cricket ball is also constrained by the person's location and pose (see Fig. 11). We describe how to apply spatial constraints on locations of objects in the action recognition framework. By combining action recognition from poses with object detection and scene analysis, we also improve the performance of standard object detection algorithms.

We first present an overview of the approach in Section 4.1. Section 4.2 describes our Bayesian model for recognition of actions and objects in static images. This is followed by a description of individual likelihood models and interactions between different perceptual elements in subsequent sections.

4.1 Overview

Studies on human-object perception suggest that people divide objects into two broad categories: scene and manipulable objects. These objects differ in the way inferences are made about them. Chao and Martin [8] showed that when humans see manipulable objects, there is cortical activity in the region that corresponds to action execution. Such responses are absent when scene objects, such as grass and house, are observed. Motivated by such studies, we treat the two classes differently in terms of the role they play in inferring human location and pose and represent them by two different types of nodes in the Bayesian model.

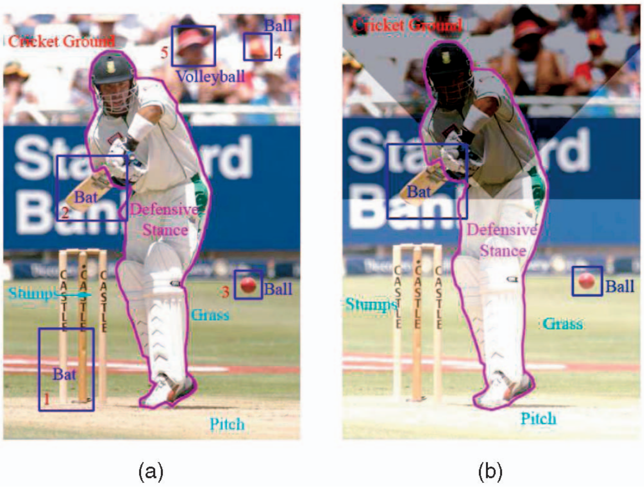


Fig. 11. Detection of manipulable objects can be improved using spatial constraints from human action. The ball detector detects two possible cricket balls. In the case of defensive batting, the probability of possible locations of the ball is shown by the shaded regions. Hence, the region below the centroid, where the ball is more likely to be present, is brighter. The ball denoted in box 4 lies in a darker region, indicating it is less likely to be a cricket ball due to its location with respect to the human. For objects such as bats, another important spatial constraint is connectedness. A segment of the bat should be connected to a segment of the human; therefore, false positives, such as object 1, can be rejected. (a) Without spatial constraints. (b) With spatial constraints.

Our Bayesian model consists of four types of nodes, corresponding to scene/event, scene objects, manipulable objects, and human. The scene node corresponds to the place where the action is being performed, such as a cricket ground or a tennis court. The scene object nodes correspond to objects which do not have causal dependency on the human actor and are mostly fixed in the scene, such as the net in the tennis court. Manipulable objects correspond to the instruments of the game such as a ball or a racket.

The interactions between these nodes are based on semantic and spatial constraints. The type of objects that occur in an image depends on the scene in which the action takes place. For example, it is more likely for a pitch to occur in a cricket ground than a tennis court. Therefore, there exist semantic relationships between scene and scene objects.

The type of action corresponding to a pose depends on the type of scene and the scene objects present. The type of action also depends on the location of the human with respect to the scene objects. For example, a pose with one hand up in a tennis court can either be a serve or a smash. However, if the human is located at the baseline it will more likely be a serve; otherwise, if he is near the net it will more likely be a smash. While considering that such spatial relationships are important, in this paper, we consider only the semantic relationships between actions and the scene and scene objects. Since we are not modeling spatial relationships between scene objects and human actions,

[7] By semantic relationships we refer to those relationships that are captured by co-occurrence statistics.

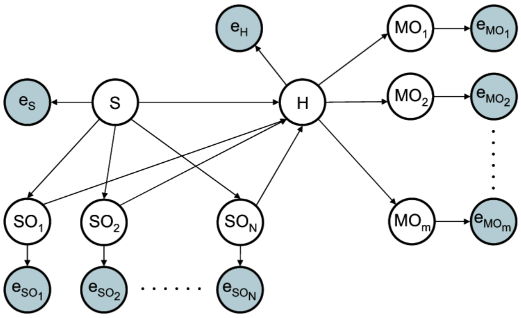


Fig. 12. Graphical model. The observed and hidden nodes are shown in blue and white, respectively.

we only consider the presence/absence of scene objects. Therefore, each scene object node (representing a class such as cricket-stumps) is characterized by a binary variable indicating the presence/absence of that scene object class.

For manipulable objects, there exist both spatial and semantic constraints between people and the objects. The type of manipulable objects in the image depends on the type of action being performed. Also, the location of the manipulable objects is constrained by the location of the human, the type of action and the types of manipulable objects. For example, the location of a tennis ball is constrained by the type of action (in the case of a forehand the ball is located to the side of a person while in the case of a serve it appears above). Spatial constraints also depend on the type of object; objects such as a tennis racket should be connected to the person while objects such as a ball generally have no such connectivity relationships. We describe an approach to represent such relationships in our Bayesian network.

4.2 Our Bayesian Model

The graphical model used for the scene interpretation framework is shown in Fig. 12. We simultaneously estimate the scene type, scene objects, human action, and manipulable object probabilities. Let S represent the scene variable, $SO_1 \dots SO_N$ represent the N type of scene objects, H represent the human, and $MO_1..MO_M$ represent the M possible manipulable objects. If $e = \{e_S, e_{SO_1}..e_{SO_N}, e_H, e_{MO_1}..e_{MO_N}\}$ represents the evidential variables or the observations, our goal is to estimate $P(S, H, SO_1..SO_N, MO_1..MO_M|e)$. This can be decomposed as

$$\prod_j P(MO_j|H)P(MO_j|e_{MO_j})P(H|S, SO_1..SO_N)P(H|e_H)\dots$$
$$\dots\prod_i P(SO_i|S)P(SO_i|e_{SO_i})P(S|e_S). \tag{4}$$

We use the loopy belief propagation algorithm [40] for inference over the graphical model.

4.3 Scene Perception

A scene is mainly characterized as a place in which we can move [39]. In this paper, the scene corresponds to the place where an action is being performed such as tennis court and croquet field. Each image is associated with a probability of belonging to one of the scene classes. Several experimental studies have shown that when humans view a scene, they extract functional and categorical information from the scene, whereas they tend to ignore information regarding specific objects and their locations. In accordance, Oliva and Torralba [39] bypass the segmentation and processing of individual objects in their scene classification framework. Rather than looking at a scene as a configuration of objects, they propose to consider a scene like an individual object, with a unitary shape. They show that scenes belonging to the same category share a similar and stable spatial structure that can be extracted at once, without segmenting the image. A set of holistic spatial properties of the scene, together referred to as a Spatial Envelope, are used, which include naturalness, openness, roughness, ruggedness, and expansion. We use their approach to compute the concatenated feature vector for every image in the data set. Using the training feature vectors we train a Support Vector Machine (SVM) for the classification task. For a test image, the SVM returns a score d_S which represents the distance of the test point from the separating hyperplane. Based on this distance, we estimate the probability $P(S|e_S)$ as

$$P(S|e_S) = \frac{1}{Z_{Scene}} exp(-\alpha_{Scene} d_S), \tag{5}$$

where α_{Scene} is the scaling parameter and Z_{Scene} is the normalization factor.

4.4 Scene Objects

Each scene object node corresponds to a class of scene objects and is represented by the probability of presence of that object class across the image. We uniformly sample points across the image and extract a patch around each point (for experiments, grid points are sampled at 25 pixels each in the x, y direction and the patch size of 50×50 is used). We classify each patch as belonging to one of the N scene object classes, using an adaboost-based classifier [55] based on features such as HOG, histograms of each color channel (eight bins each in color channel), and histograms of edge distance map values within the neighborhood. We compute $P(SO_i|S)$ based on the conditional probability tables learned using the co-occurrence relationships in the training data set.

4.5 Human in Action

Every detected person in the image is characterized by the action (A) he is performing, and location given by a bounding box (l^H). For action classification, we detect humans and employ the pose information. A similar approach has been proposed in a recent paper [13]. In our experiments, we detect humans using an approach similar to [59]. Since the observed image shape of a human, changes significantly with articulation, viewpoint, and illumination, it is infeasible to train a single human detector for all shapes. Instead, we first cluster the observed shapes from our training data, and train multiple human detectors, one for each shape cluster. Our human detectors closely match those proposed by [9]. Given a bounding box around a detected human, we segment the human using *GrabCut* [4], an efficient tool for foreground segmentation. Once we have a possible human segmentation, we extract shape context features (5 radial bins and 12 orientation bins) from the silhouette of the human. We then cluster shape context features [1] from the training database to build a dictionary of “shape context words.” A detected human in an image is then characterized by the histogram of shape context words. The number of words/clusters determines the dimensionality of our pose feature vector. We then use the K-Nearest Neighbor approach for classification, providing $P(H|e_H)$. Given a test sample, we determine the K-nearest neighbors in the training data. Each of the K neighbors vote for the class it belongs to with a weight based on its distance from the test sample. The final scores obtained for each class determine the likelihoods for each pose category, $P(H|e_H)$. For the experiments used in the paper, we use $K = 5$.

We also need to compute $P(H|S, SO_1..SO_N)$. Assuming conditional independence between scene object categories given human action, we rewrite as

$$P(H|S, SO_1..SO_N) = \prod_i^N P(H|S, SO_i). \quad (6)$$

Each of these can be computed using co-occurrence statistics of human action-scene-scene object combinations, independently for every scene object class.

4.6 Manipulable Objects

Each detected manipulable object in the image has the following attributes: an associated class id (c_i^m) and location parameters given by a bounding box (l_i^m) around the object. We use the object detector described in Section 3.3. Using this approach, however, we are unable to distinguish between objects that have the same shape but a different dominant color; for example, a cricket ball (often red or white in color) as opposed to a tennis ball (often yellow in color). Thus, we build appearance models of manipulable objects using nonparametric Kernel Density Estimation (KDE) to also perform an appearance-based classification. We sample pixels from training images of the manipulable objects and build a 3D model in the RGB space.

$$p_{Model}(r, g, b) = \frac{1}{N} \sum_{i=1}^{N} K_{\sigma_r}(r - r_i) K_{\sigma_g}(g - g_i) K_{\sigma_b}(b - b_i). \quad (7)$$

Given a test image, we first use the shape-based classifier to detect potential object candidates. Within each candidate window, we sample pixels and build a density estimate using KDE. This test density is compared to the color model of every object category using the Kullback-Leibler distance. This provides the final manipulable object detection probabilities based on appearance given by $P(MO_i|e^{ap}_{MO_i})$. Therefore, the probability $P(MO_i|e_{MO_i})$ is given by

$$P(MO_i|e_{MO_i}) = P\big(MO_i|e^{sh}_{MO_i}\big) P\big(MO_i|e^{ap}_{MO_i}\big), \quad (8)$$

where e^{sh} refers to shape and e^{ap} refers to appearance evidence. We also need to compute $P(MO_i|H)$. Human actions and locations provide both semantic and spatial constraints on manipulable objects. The spatial constraints given human locations are with respect to the type of manipulable object and type of action being performed. We model two kinds of spatial constraints: 1) Connectivity—Certain manipulable objects like a tennis racket or a cricket bat should be connected to the human in action. 2) Positional and Directional Constraints—These location constraints are evaluated with respect to the centroid of the human that is acting on them. The conditional probability densities are based on the type of action being performed. For example, given a tennis-serve action, it is more likely that the ball is above the player, while, if the action is forehand, it is more likely to the side of the player. We model positional relations in terms of the displacement vector of the object centroid from the centroid of the human body. Thus, we obtain

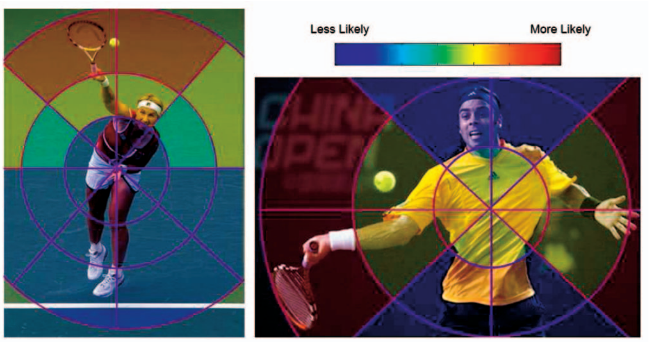


Fig. 13. Spatial constraints between locations of manipulable objects and humans for different poses. In an idealized scenario, for a forehand pose, the ball is more likely to be seen on the side; for a tennis-serve, it is more likely to be seen above the human. We use two radial bins and eight orientation bins to specify position of manipulable object with respect to the human body.

$$P\big(MO_i = (c_i^m, l_i^m)|H = (A, l^H)\big) = P\big(l_i^m|c_i^m, A, l^H\big) P\big(c_i^m|A\big). \quad (9)$$

The first term refers to the spatial constraints and can be learned by discretizing the space around the human as shown in Fig. 13. From the training images, we learn the condition probability tables of the region in which the manipulable object lies given the type of manipulable object and the type of action. The second term is the semantic constraint and is modeled from co-occurrence statistics of human action-manipulable objects combinations from training data.

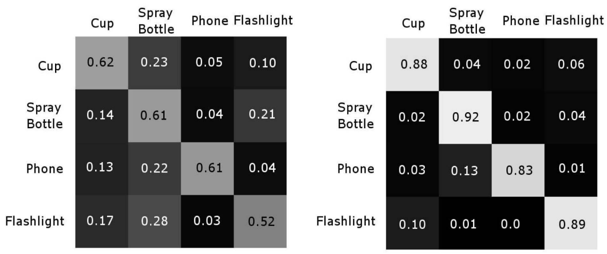
5 Experimental Evaluation

5.1 Video Interpretation

We evaluated our video interpretation framework on test data set[8] of 10 subjects performing six possible interactions with four different objects. The objects in the test data set 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.

Training. We used a fully supervised approach for training the Bayesian model for video interpretation. 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 object detector was done using images from training data sets obtained using Google image search (50 images for each object, negative images were used from INRIA and CALTECH data sets). HMM models were trained using a separate training data set

[8] The data sets used in all the experiments are available online and can be downloaded from http://www.umiacs.umd.edu/~agupta.



	Cup	Spray Bottle	Phone	Flashlight
Cup	0.62	0.23	0.05	0.10
Spray Bottle	0.14	0.61	0.04	0.21
Phone	0.13	0.22	0.61	0.04
Flashlight	0.17	0.28	0.03	0.52

(a)

	Cup	Spray Bottle	Phone	Flashlight
Cup	0.88	0.04	0.02	0.06
Spray Bottle	0.02	0.92	0.02	0.04
Phone	0.03	0.13	0.83	0.01
Flashlight	0.10	0.01	0.0	0.89

(b)

Fig. 14. Object likelihood confusion matrix: The ith row depicts the expected likelihood values when ith type of object is present. The right table shows the results of our whole framework, taking into account action, object reaction, and reach motion. (a) HOG detector. (b) Using whole framework.

of videos. The object reactions are learned using the supervised training scheme. In training videos, the frames for the object reaction were manually segmented and the appearance histograms around the hand were used to learn the appearance of object reaction. Additionally, our model requires co-occurrence statistics of object-interaction-reaction combinations, distance between grasp location and object center, and reaction-time ratios. We used a training data set of 30 videos of five actors performing different types of manipulations on the objects. Training was done in a fully supervised manner. All of the videos were manually labeled with object locations, hand locations and the type of objects, manipulation, and object reactions.

Object classification. Among the objects used, it is hard to discriminate the spray bottle, flashlight, and cup because all three are cylindrical (see Figs. 16a and 16b). Furthermore, the spray bottle detector also fired for the handset of the cordless phone (see Fig. 16d). Our approach was also able to detect and classify objects of interest even in cluttered scenes (see Fig. 16c). Fig. 18 shows some more object recognition results. Figs. 14a and 14b show 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 percent to 96.67 percent.[9]

Action recognition. Of the six activities, it is very hard to discriminate between pouring and lighting on the basis of hand trajectories (see Figs. 16a and 16b). 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. Fig. 15a shows the likelihoods of actions that were obtained for all the videos using hand-dynamics alone. Fig. 15b shows the confusion matrix when action recognition was conducted using our framework. The overall recognition rate increased from 76.67 percent to 93.34 percent when action was recognized using the contextual information from objects and object reactions. While the trajectories might be similar in many cases, the context from object provided cues to differentiate between confusing actions. Similarly, in the cases of lighting and pouring, contextual cues from object reaction helped in differentiating between those two actions.

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 labelings. Fig. 17 shows the histogram of segmentation errors in the videos of the test data set. It can be seen that 90 percent of detections were within three frames of actual end-frames of reach motion. The average length of the video sequence was approximately 110 frames.

5.2 Image Interpretation

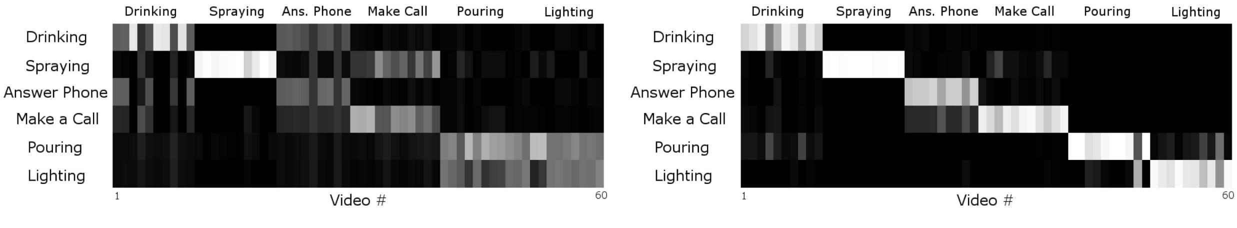
Data set. We evaluated our approach on a data set which had six possible actions: "tennis-forehand," "tennis-serve," "volleyball-smash," "cricket-defensive shot," "cricket-bowling," and "croquet-shot." The images for the first five classes were downloaded from the internet and for the sixth class, we used a publicly available data set [29]. A few images from the data set are shown in Fig. 19. The classes were selected so that they had significant confusion due to scene and pose. For example, the poses during "volleyball-smash" and "tennis-serve" are quite similar and the scenes in "tennis-forehand" and "tennis-serve" are exactly the same.

Training. We used a fully supervised approach for training the Bayesian model for image interpretation. We have to learn the parameters for individual likelihood functions and parameters of the conditional probabilities which model the interactions between different perceptual analyses. To learn parameters of individual likelihood functions, we trained individual detectors separately using training images from Google image search (50 images each for every object and 30 silhouettes each for the pose likelihood). Learning parameters corresponding to conditional probabilities requires a separate training data set of images. Our training data set consisted of 180 images (30 from each class).

Evaluation. We tested the performance of our algorithm on a data set of 120 test images (20 from each class). We compared the performance of our algorithm with the performance of models based on isolated components. Fig. 20 shows the confusion matrix obtained using the full model described in the paper. We also show some failure cases in the figure. Our approach gave some misclassifications when the scene involved is the same but actions are different such as bowling being classified as batting. This occurs whenever the pose classification algorithm gives a wrong action likelihood (mostly due to faulty segmentation by Grabcut) and the manipulable object detector fails to find any discriminating manipulable object.

Fig. 21a shows the performance of a pose-based classification algorithm. We used the pose component of our model to obtain the confusion matrix. As expected, the performance of pose-only model is very low due to similar poses being shared by different actions. For example, there is high confusion between "tennis-serve" and "bowling," since both actions share a high arm pose. Similarly, we see confusion between "bowling" and "volleyball." The confusion between

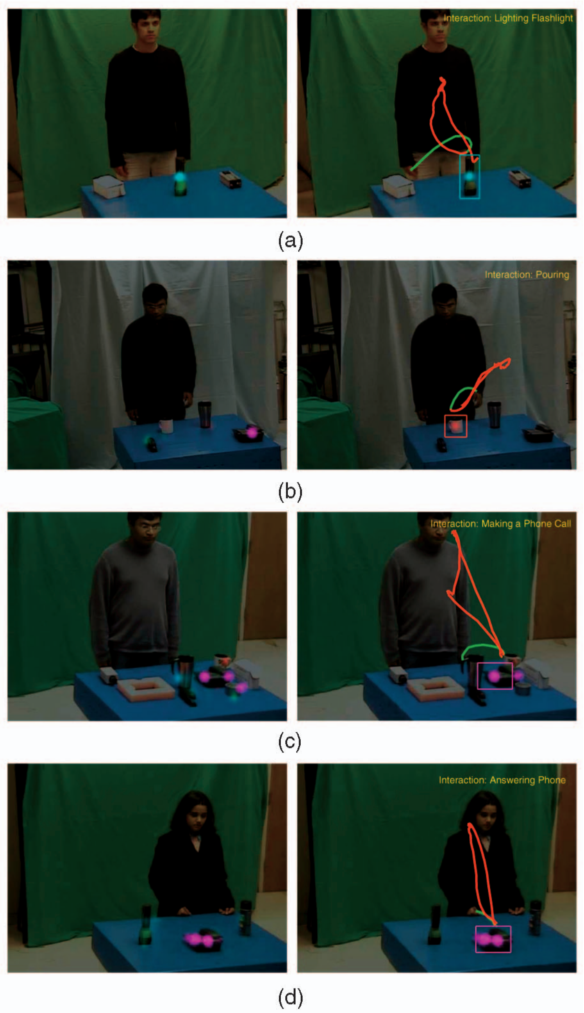
[9] The recognition rate depicts the correct classification of localized object into one of the five classes: background, cup, spray bottle, phone, and flashlight.



(a)

(b)

Fig. 15. Comparison of action likelihoods without and with contextual information. Each column represents the normalized likelihood values for six possible actions. (a) HMM-based action recognition. (b) HMM-based recognition in interaction context.


Interaction: Lighting Flashlight
Interaction: Pouring
Interaction: Making a Phone Call
Interaction: Answering Phone

(a)

(b)

(c)

(d)

Fig. 16. 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. Also note that the users do not have to wear long-sleeved clothing for the hand tracking to work, since the pose estimation framework uses a likelihood model based on edges and background subtraction in addition to skin color for searching the correct pose.

"volleyball-smash" and "tennis-forehand" is mainly due to incorrect segmentations by grabcut.

The comparison between overall performance of our approach and the individual components is shown in Fig. 21b. The performance of our approach was 78.86 percent as compared to 57.5 percent by the pose-only model and 65.83 percent by the scene-only model.

Figs. 22 and 23 show some examples of correct classification by our algorithm. In both cases, our approach rejects false positives because the belief in the objects falls below the detection threshold when combined with other elements like pose and scene information. For example, in Fig. 22, the false positives of bats are rejected as they fail to satisfy spatial constraints. Also, in both cases, detections related to objects incongruent with scene and action information are also rejected.

Influence of parameters. We evaluated our system with respect to the parameters of each component of our system. We varied the parameter α_{Scene} used to obtain the scene classification probabilities (Section 4.3). Fig. 24a shows that action recognition accuracy increases with increasing α_{Scene}, but flattens out after a value of 5. The discriminative power of the scene component lowers with decreasing α_{Scene} and, therefore, we observe a lower system performance. In our experiments, we use $\alpha_{Scene} = 5$.

Oliva and Torralba [39] use the Windowed Discriminant Spectral Template (WDST) which describes how the spectral components at different spatial locations contribute to a spatial envelope property, and sample it at regular intervals to obtain a discrete representation. One of the components of their method, w_{Scene}, determines the coarseness of this sampling interval. We varied the coarseness of the sampling where smaller w_{Scene} refers to coarser sampling. Fig. 24b shows our performance accuracy with respect to w_{Scene}. Our action recognition accuracy reduces

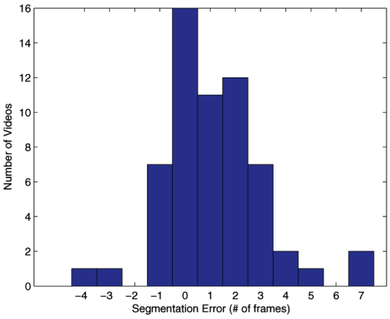


Fig. 17. Segmentation error histogram.

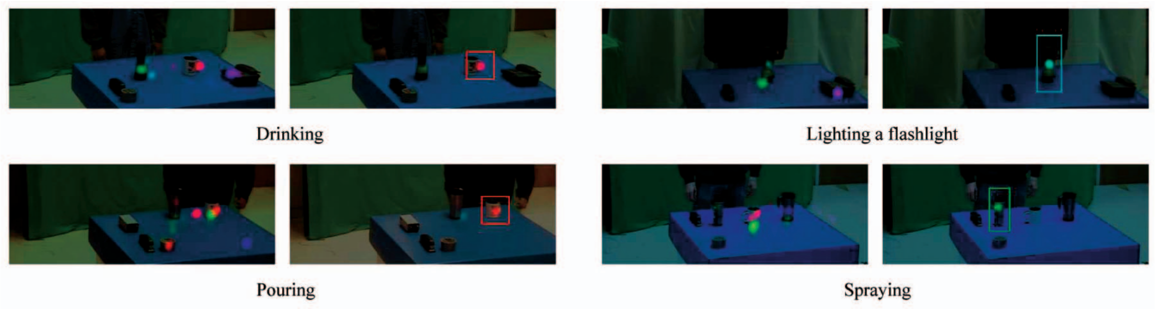


Fig. 18. Object recognition using contextual cues from reach, manipulation, and object reaction. As before, the colors red, green, cyan, and magenta show the likelihoods of cup, spray bottle, flashlight, and phone, respectively. The activities in the four cases above are drinking, pouring, lighting, and spraying, respectively.

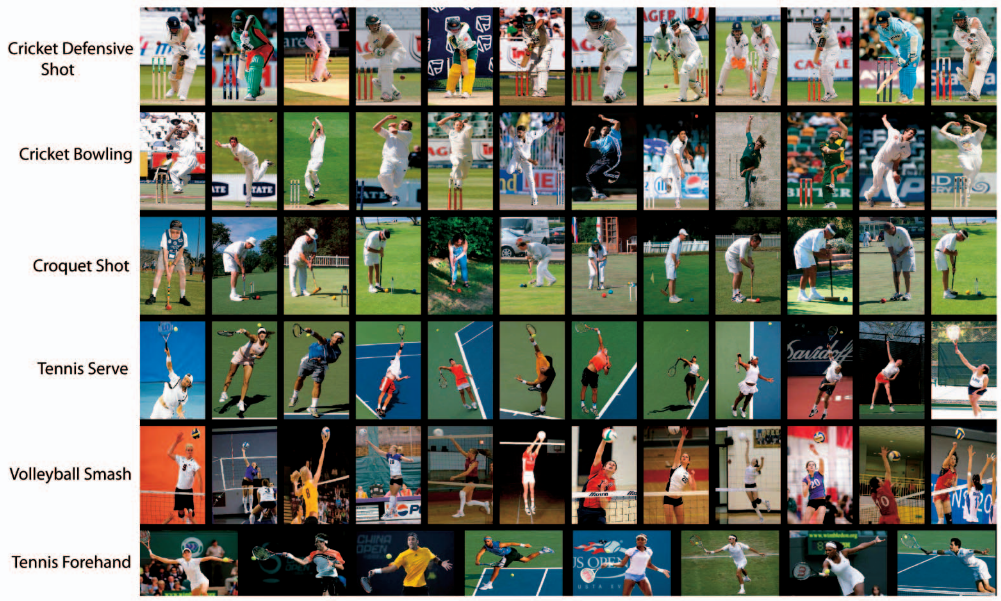
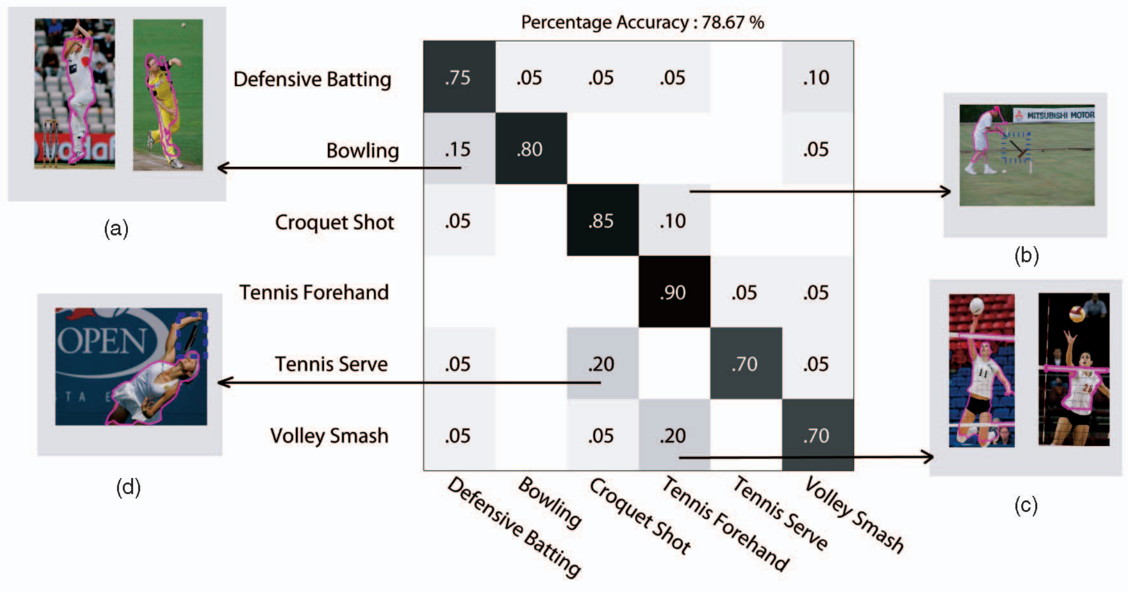


Fig. 19. Our data set.



	Defensive Batting	Bowling	Croquet Shot	Tennis Forehand	Tennis Serve	Volley Smash
Defensive Batting	.75	.05	.05	.05		.10
Bowling	.15	.80				.05
Croquet Shot	.05		.85	.10		
Tennis Forehand				.90	.05	.05
Tennis Serve	.05		.20		.70	.05
Volley Smash	.05		.05	.20		.70

Fig. 20. Confusion matrix (full model): The figure shows the confusion matrix obtained using the full model. We also show some failure cases in the adjoining boxes. (a) The scene in these cases is classified correctly as cricket ground; however, due to faulty segmentations, the hands of the bowler are missed and the pose is misclassified as batting. (b) The pose is again misclassified as that of forehand due to some extra regions added to human segment. The missed detection (shown in dotted blue) of croquet bat also contributes to the misclassification. (c) In both the cases the segmentation fails, leading to inclusion of net with the human segment. (d) Apart from the error in the pose module, the racket is also missed and the ball is not present in the scene.

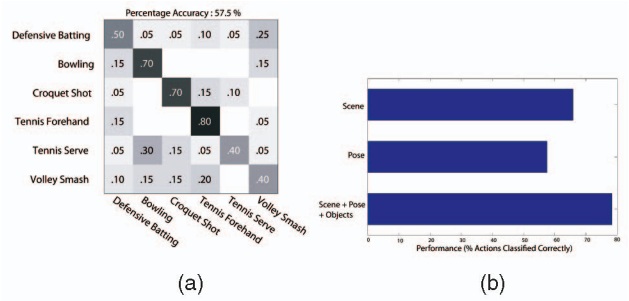


Fig. 21. (a) Confusion matrix (pose only): The confusion matrix is that only pose information is used for action classification. (b) Comparative performance of our approach with individual components.

for a very coarse sampling of the WDST, but is stable at finer scales. We use $w_{Scene} = 4$ for the experiments.

Our object detection module detects multiple objects in the scene and passes the top few detections onto the Bayesian framework. We evaluated our system accuracy with regards to the number of manipulable object detections passed to the Bayesian framework. For lower number of detections, the Bayesian framework has lower performance due to missing true detections. For higher number of detections, the Bayesian framework has lower performance due to the confusion from false positives. This effect is more pronounced for lower α_{Scene} values where the scene component has lower discriminativeness (see Fig. 24c).

Finally, we evaluated our system with respect to the dimensionality of the pose feature vector. This dimensionality is determined by the number of "shape context words" formed in the shape dictionary. Fig. 24d shows the accuracy

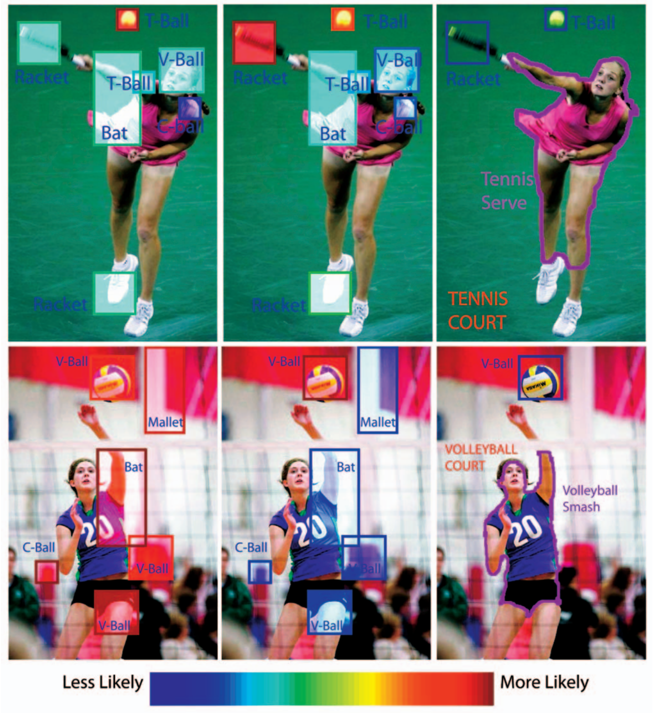


Fig. 23. Some other examples: In the first case, the tennis racket was detected with a lower likelihood as compared to other objects. After combining information from scene and action, the belief in the tennis racket increases since the action and the scene are tennis-serve and tennis court, respectively. In the second case, our approach rejects false positives of objects such as a mallet and bat. These objects are rejected, as they are not congruent to a volleyball-court and a volleyball-smash action. The false positives in volleyballs are also rejected as they fail to satisfy spatial constraints. Same abbreviations as in Fig. 22.

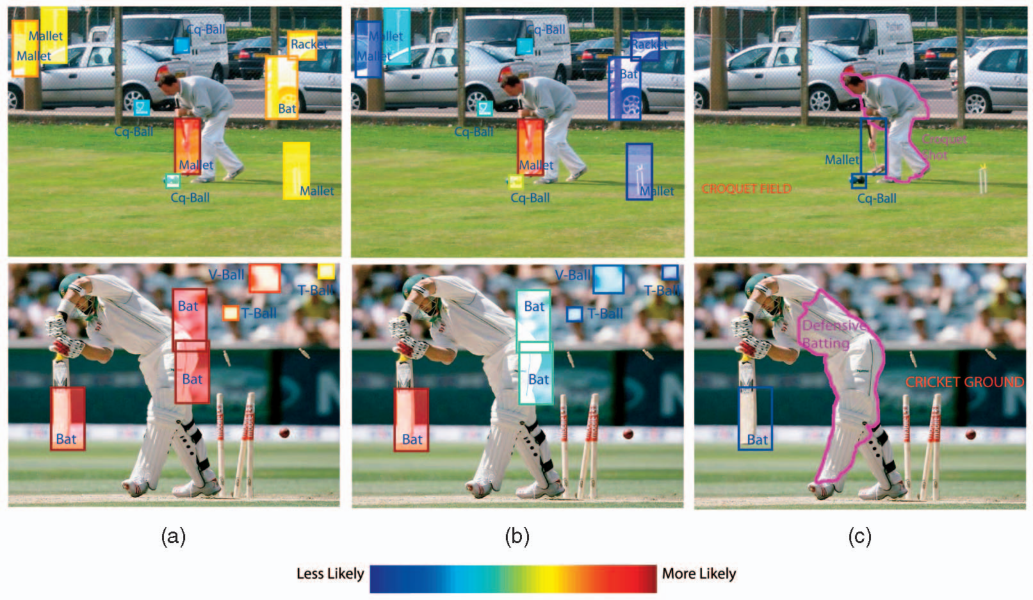
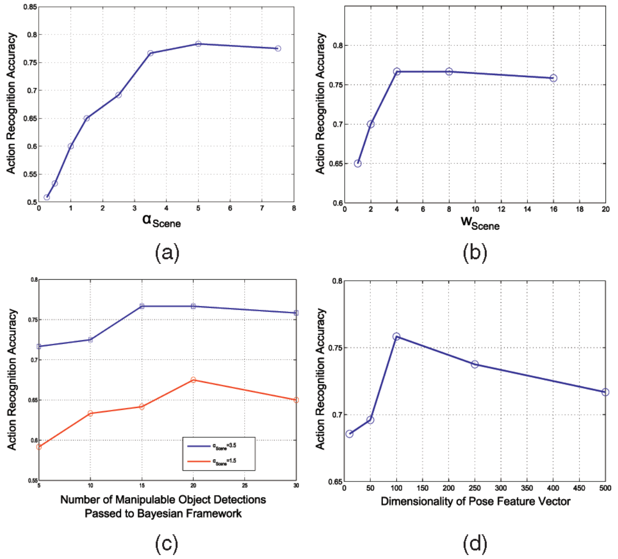


Fig. 22. Some illustrative examples showing the performance of the system. (a) The likelihood of various objects using independent detectors. The colors of the rectangles represent the likelihood probability (red meaning higher probability and blue meaning lower probability). (b) The posterior probabilities after the framework was applied. (c) The final result of our approach. In the first example, the detector detects four possible mallets and three possible croquet balls. After applying the spatial constraints, all the false positives are rejected as they fail to satisfy spatial constraints (the other mallets are not connected to a human body and the other balls are above the detected human centroid). In the second example, the false positives of bats are rejected as they fail to satisfy spatial constraints. Also, in both cases, detections related to objects incongruent with scene and action information are also rejected. (Note the abbreviations *T-Ball*, *C-Ball*, *V-Ball*, and *Cq-Ball* refer to tennis, cricket, volley, and croquet balls, respectively.)



(a)

(b)

(c)

(d)

Fig. 24. Influence of parameters on the system performance. (a) α_{Scene}. (b) w_{Scene}. (c) Number of manipulable objects. (d) Dimensionality of pose features.

of our system against the dimensionality of the pose feature vector. As expected, our performance reduces when using a very small number of words. In our experiments, we use a dictionary of 100 visual words resulting in a 100-dimensional pose feature vector.

6 Conclusion

Recent studies related to human information processing have confirmed the role of object recognition in action understanding and vice versa. Furthermore, neuropsychological studies have also shown that not only videos but also static images of humans in action evoke cortical responses in the brain's motor area, indicating that humans tend to perceive dynamic information from static images as well. Motivated by such studies, we present two Bayesian models for interpretation of human-object interactions from videos and static images, respectively.

Our approach combines the processes of scene, object, action, and object reaction recognition. Our Bayesian model incorporates semantic/functional and spatial context for both object and action recognition. Therefore, by enforcing global coherence between different perceptual elements, we can improve the recognition performance of each element substantially.

Acknowledgments

The authors would like to acknowledge VACE for supporting the research. The authors would like to thank Vitaladevuni Shiv and Mohamed Hussein for providing the code for reach motion detection and object detection, respectively. The authors would also like to thank the authors of [29] for providing the image data set for croquet-shot action in the paper, the subjects who helped in collection of human-object interaction data set (videos), and Swati Jarial for the help in formation of static image data set. A preliminary version of this paper appears in [18].

References

[1] A. Agarwal and B. Triggs, "3d Human Pose from Silhouettes by Relevance Vector Regression," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2004.

[2] P. Bach, G. Knoblich, T. Gunter, A. Friederici, and W. Prinz, "Action Comprehension: Deriving Spatial and Functional Relations," *J. Experimental Psychology Human Perception and Performance,* vol. 31, no. 3, pp. 465-479, 2005.

[3] A. Berg and J. Malik, "Geometric Blur for Template Matching," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2001.

[4] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr, "Interactive Image Segmentation Using an Adaptive GMMRF Model," *Proc. European Conf. Computer Vision,* 2004.

[5] A. Bobick and A. Wilson, "A State-Based Approach to the Representation and Recognition of Gesture," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 12, pp. 1325-1337, Dec. 1997.

[6] A. Bosch, A. Zisserman, and X. Muñoz, "Image Classification Using Random Forests and Ferns," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[7] D. Bub and M. Masson, "Gestural Knowledge Evoked by Objects As Part of Conceptual Representations," *Aphasiology,* vol. 20, pp. 1112-1124, 2006.

[8] L.L. Chao and A. Martin, "Representation of Manipulable Man-Made Objects in Dorsal Stream," *NeuroImage,* vol. 12, pp. 478-484, 2000.

[9] N. Dalal and B. Triggs, "Histogram of Oriented Gradients for Fast Human Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2005.

[10] J. Davis, H. Gao, and V. Kannappan, "A Three-Mode Expressive Feature Model of Action Effort," *Proc. IEEE Workshop Motion and Video Computing,* 2002.

[11] Z. Duric, J. Fayman, and E. Rivlin, "Function from Motion," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 6, pp. 579-591, June 1996.

[12] P. Felzenszwalb and D. Huttenlocher, "Pictorial Structures for Object Recognition," *Int'l J. Computer Vision,* vol. 61, pp. 55-79, 2005.

[13] V. Ferrari, M. Marin, and A. Zisserman, "Progressive Search Space Reduction for Human Pose Estimation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.

[14] R. Filipovych and E. Ribeiro, "Recognizing Primitive Interactions by Exploring Actor-Object States," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.

[15] V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti, "Action Recognition in Premotor Cortex," *Brain,* vol. 2, pp. 593-609, 1996.

[16] G. Guerra and Y. Aloimonos, "Discovering a Language for Human Activity," *Proc. Assoc. Advancement of Artificial Intelligence Workshop Anticipation in Cognitive Systems,* 2005.

[17] A. Gupta, T. Chen, F. Chen, D. Kimber, and L. Davis, "Context and Observation Driven Latent Variable Model for Human Pose Estimation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.

[18] A. Gupta and L. Davis, "Objects in Action: An Approach for Combining Action Understanding and Object Perception," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.

[19] A. Gupta and L. Davis, "Beyond Nouns: Exploiting Prepositions and Comparative Adjectives for Learning Visual Classifiers," *Proc. European Conf. Computer Vision,* 2008.

[20] A. Gupta, A. Mittal, and L. Davis, "Constraint Integration for Efficient Multiview Pose Estimation with Self-Occlusions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 3, pp. 493-506, Mar. 2008.

[21] A. Gupta, J. Shi, and L. Davis, "A 'Shape Aware' Model for Semi-Supervised Learning of Objects and Its Context," *Proc. Conf. Neural Information Processing Systems,* 2008.

[22] A. Gupta, P. Srinivasan, J. Shi, and L. Davis, "Understanding Videos, Constructing Plots—Learning a Visually Grounded Storyline Model from Annotated Videos," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2009.

[23] H.B. Helbig, M. Graf, and M. Kiefer, "The Role of Action Representation in Visual Object," *Experimental Brain Research,* vol. 174, pp. 221-228, 2006.

[24] D. Hoiem, A. Efros, and M. Hebert, "Putting Objects in Perspective," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2006.

[25] S.H. Johnson-Frey, F.R. Maloof, R. Newman-Norlund, C. Farrer, S. Inati, and S.T. Grafton, "Actions or Hand-Object Interactions? Human Inferior Frontal Cortex and Action Observation," *Neuron,* vol. 39, pp. 1053-1058, 2003.

[26] Z. Kourtzi, "But Still it Moves," *Trends in Cognitive Science,* vol. 8, pp. 47-49, 2004.

[27] Z. Kourtzi and N. Kanwisher, "Activation in Human MT/MST by Static Images with Implied Motion," *J. Cognitive Neuroscience,* vol. 12, pp. 48-55, 2000.

[28] Y. Kuniyoshi and M. Shimozaki, "A Self-Organizing Neural Model for Context Based Action Recognition," *Proc. IEEE Eng. Medicine and Biology Soc. Conf. Neural Eng.,* 2003.

[29] L.-J. Li and L. Fei-Fei, "What, Where and Who? Classifying Events by Scene and Object Recognition," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[30] R. Mann, A. Jepson, and J. Siskind, "The Computational Perception of Scene Dynamics," *Computational Vision and Image Understanding,* vol. 65, no. 2, pp. 113-128, 1997.

[31] R. Marteniuk, C. MacKenzie, M. Jeannerod, S. Athenes, and C. Dugas, "Constraints on Human Arm Movement Trajectories," *Canadian J. Psychology,* vol. 41, pp. 365-378, 1987.

[32] T.B. Moeslund, A. Hilton, and V. Kruger, "A Survey of Advances in Vision-Based Human Motion Capture and Analysis," *Computer Vision and Image Understanding,* vol. 2, pp. 90-126, 2006.

[33] D. Moore, I. Essa, and M. Hayes, "Exploiting Human Action and Object Context for Recognition Tasks," *Proc. IEEE Int'l Conf. Computer Vision,* 1999.

[34] H. Murase and S. Nayar, "Learning Object Models from Appearance," *Proc. Nat'l Conf. Artificial Intelligence,* 1993.

[35] K. Murphy, A. Torralba, and W. Freeman, "Graphical Model for Scenes and Objects," *Proc. Conf. Neural Information Processing Systems,* 2003.

[36] K. Murphy, A. Torralba, and W. Freeman, "Using the Forest to See the Trees: A Graphical Model Relating Features, Objects and Scenes," *Proc. Conf. Neural Information Processing Systems,* 2004.

[37] H. Nagel, "From Image Sequences towards Conceptual Descriptions," *Image and Vision Computing,* vol. 6, no. 2, pp. 59-74, 1988.

[38] K. Nelissen, G. Luppino, W. Vanduffel, G. Rizzolatti, and G. Orban, "Observing Others: Multiple Action Representation in Frontal Lobe," *Science,* vol. 310, pp. 332-336, 2005.

[39] A. Oliva and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," *Int'l J. Computer Vision,* vol. 42, pp. 145-175, 2001.

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

[41] P. Peursum, G. West, and S. Venkatesh, "Combining Image Regions and Human Activity for Indirect Object Recognition in Indoor Wide Angle Views," *Proc. IEEE Int'l Conf. Computer Vision,* 2005.

[42] V. Prasad, V. Kellokompu, and L. Davis, "Ballistic Hand Movements," *Proc. Conf. Articulated Motion and Deformable Objects,* 2006.

[43] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in Context," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[44] C. Rao, A. Yilmaz, and M. Shah, "View-Invariant Representation and Recognition of Actions," *Int'l J. Computer Vision,* vol. 2, pp. 203-226, 2002.

[45] E. Rivlin, S. Dickinson, and A. Rosenfeld, "Recognition by Functional Parts," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 1994.

[46] M. Shah and R. Jain, *Motion-Based Recognition.* Kluwer Academic, 1997.

[47] I. Smyth and M. Wing, *The Psychology of Human Movement.* Academic Press, 1984.

[48] L. Stark and K. Bowyer, "Generic Recognition through Qualitative Reasoning about 3D Shape and Object Function," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 1991.

[49] E. Sudderth, A. Torralba, W. Freeman, and A. Wilsky, "Learning Hierarchical Models of Scenes, Objects and Parts," *Proc. IEEE Int'l Conf. Computer Vision,* 2005.

[50] J. Sullivan and S. Carlsson, "Recognizing and Tracking Human Action," *Proc. European Conf. Computer Vision,* 2002.

[51] S. Todorovic and N. Ahuja, "Learning Subcategory Relevances for Category Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.

[52] A. Torralba and P. Sinha, "Statistical Context Priming for Object Detection," *Proc. IEEE Int'l Conf. Computer Vision,* 2001.

[53] C. Urgesi, V. Moro, M. Candidi, and S. Aglioti, "Mapping Implied Body Actions in the Human Motor System," *J. Neuroscience,* vol. 26, pp. 7942-7949, 2006.

[54] L. Vaina and M. Jaulent, "Object Structure and Action Requirements: A Compatibility Model for Functional Recognition," *Int'l J. Intelligent Systems,* vol. 6, pp. 313-336, 1991.

[55] A. Vezhnevets and V. Vezhnevets, "'Modest Adaboost'—Teaching Adaboost to Generalize Better," *Proc. Graphicon,* 2005.

[56] Y. Wang, H. Jiang, M. Drew, Z. Li, and G. Mori, "Unsupervised Discovery of Action Classes," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2006.

[57] A. Wilson and A. Bobick, "Parametric Hidden Markov Models for Gesture Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 9, pp. 884-900, Sept. 1999.

[58] B. Wu and R. Nevatia, "Detection of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detectors," *Proc. IEEE Int'l Conf. Computer Vision,* 2005.

[59] B. Wu and R. Nevatia, "Detection and Tracking of Multiple Humans with Extensive Pose Articulation," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[60] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg, "A Scalable Approach to Activity Recognition Based on Object Use," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[61] A. Yilmaz and M. Shah, "Actions Sketch: A Novel Action Representation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2005.

[62] Q. Zhu, S. Avidan, M. Ye, and K. Cheng, "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2006.

Abhinav Gupta received the MS degree in computer science from the University of Maryland in 2007 and the BTech degree in computer science and engineering from the Indian Institute of Technology, Kanpur, in 2004. He is a doctoral candidate in the Department of Computer Science at the University of Maryland, College Park. His research focuses on visually grounded semantic models and how language and vision can be exploited to learn such models. His other research interests include combining multiple cues, probabilistic graphical models, human body tracking, and camera networks. He has published numerous papers in prestigious journals and conferences on these topics. He has received several awards during his academic career including the University of Maryland Dean's Fellowship for excellence in research. He is a member of the IEEE.

Aniruddha Kembhavi received the bachelor's degree in electronics and telecommunications engineering from the Government College of Engineering, Pune, India, in 2004. He is currently working toward the PhD degree in the Computer Vision Laboratory at the University of Maryland, College Park. His current research focuses on the problem of variable and feature selection for object classification. His research interests include human detection, object classification, and machine learning. He is a member of the IEEE.

Larry S. Davis received the BA degree from Colgate University in 1970 and the MS and PhD degrees in computer science from the University of Maryland in 1974 and 1976, respectively. From 1977 to 1981, he was an assistant professor in the Department of Computer Science at the University of Texas, Austin. He returned to the University of Maryland as an associate professor in 1981. From 1985 to 1994, he was the director of the University of Maryland Institute for Advanced Computer Studies, where he is currently a professor, and is also a professor and the chair of the Computer Science Department. He was named a fellow of the IEEE in 1997.