Computer Vision and Image Understanding 115 (2011) 81–90

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Visual object-action recognition: Inferring object affordances from human demonstration ☆

Hedvig Kjellström *, Javier Romero, Danica Kragić

Computational Vision and Active Perception Lab, Centre for Autonomous Systems School of Computer Science and Communication, KTH, SE-100 44 Stockholm, Sweden

ARTICLE INFO

Article history:
Received 12 March 2010
Accepted 14 August 2010
Available online 20 August 2010

Keywords:
Object recognition
Action recognition
Contextual recognition
Object affordances
Learning from demonstration

ABSTRACT

This paper investigates object categorization according to function, i.e., learning the *affordances* of objects from human demonstration. Object affordances (functionality) are inferred from observations of humans using the objects in different types of actions. The intended application is learning from demonstration, in which a robot learns to employ objects in household tasks, from observing a human performing the same tasks with the objects. We present a method for categorizing manipulated objects and human manipulation actions in context of each other. The method is able to simultaneously segment and classify human hand actions, and detect and classify the objects involved in the action. This can serve as an initial step in a learning from demonstration method. Experiments show that the contextual information improves the classification of both objects and actions.

1. Introduction

In recent years, a tremendous research effort has been made in the area of visual object categorization [8], leading to methods with impressive performance on very difficult images. Object classes are typically semantic and appearance-based; common examples are cups, toys, bikes, cars, and trees.

For certain classes of applications, e.g., in Robotics, it is however more meaningful to categorize objects according to their *function* [31,39,40]. Both a chair, a sofa, and a stool might be categorized as "sittable", and a cup might be categorized both as "drinkable" and "pourable" (Fig. 1).

To a certain extent, functional object properties can be extracted from visual information. However, there are functional properties that can not be observed visually from a single image, such as temperature, flexibility and weight. We propose to learn functional properties of objects from video sequences where a human perform actions involving the objects.

The application we are focusing on is robot learning from demonstration [1], also denoted imitation learning [34]. With imitation we here do not mean blind reproduction of the movements of all body parts; rather, we mean observing an action and its effect on the world, and performing an action that has the same effect [24].

We here formulate the problem of learning from demonstration as one of learning the *affordances* of objects. Introduced as a concept by Gibson [12], affordances are properties of the environment that *afford* a certain action to be performed by a human or an animal. Here we study affordances of objects involved in human manipulation actions.

An affordance is an intrinsic property of an object, allowing an action to be performed with the object. The affordance also depends on the embodiment of the agent performing the action. For example, a human can use a knife to chop an onion, while a dog can not. Hence, the knife *affords* onion chopping to a human but not to a dog.

From this we can conclude that the learning of object affordances is facilitated if the agent (robot) learning the affordances has an embodiment similar to a human: Two arms with approximately the same reaching range and at the same height as human arms, and human-like hands, which can manipulate objects in the same way as human hands. This is however not an absolute requirement; there are methods for mapping motions and object manipulation actions between different embodiments [1]. This paper does not further treat robotic manipulation; we instead concentrate on the learning of object affordances from human demonstration.

Manipulation actions, i.e. hand actions for picking-up objects, doing something with them and putting them down again, is an important class of hand activity not well studied in computer vision. An important cue to the class of a manipulation action is the object handled; for example, seeing a human bring a cup towards his/her face brings us to believe that he/she is drinking, without actually seeing the fluid. Similarly, a strong cue to the class

☆ An early version of this article appears in [16].

* Corresponding author.
E-mail addresses: hedvig@kth.se (H. Kjellström), jrgn@kth.se (J. Romero), dani@kth.se (D. Kragić).
URLs: http://www.csc.kth.se/~hedvig (H. Kjellström), http://www.csc.kth.se/~jrgn (J. Romero), http://www.csc.kth.se/~danik (D. Kragić).

1077-3142/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2010.08.002

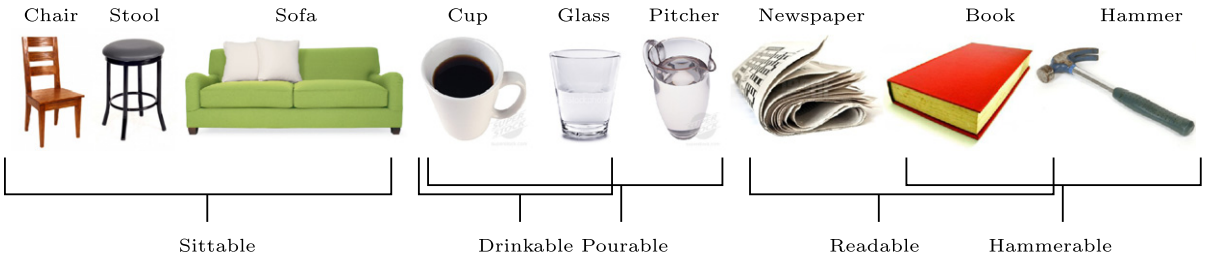


Fig. 1. Representing objects in terms of functionality and affordances. Top: Semantic, appearance-based categories. Bottom: Functional, affordance based categories.

of the object involved is the action; for example, a cup is to some extent defined as something you drink from. Therefore, it is beneficial to *simultaneously* recognize manipulation actions and manipulated objects.

Only one-hand actions are considered here, although this is not a limitation to the method in a formal sense. From a video sequence of the action, the human hand position and articulation in 3D are reconstructed and tracked using an example based method [32]. The action state space in each frame is the hand orientation and velocity as well as the finger joint angles, representing the hand shape.

Objects in this application are "graspable", i.e., fairly rigid, so shape is a good object descriptor. Objects are therefore detected in the neighborhood of the hand using a sliding window approach with a histogram of gradients (HOG) [7,11,36] representation and an SVM classifier [6]. Section 3 further describes the feature extraction.

There are implicit, complex interdependencies in the object and action data. The object detection is affected by occlusion and shading from the human hand. Similarly, the hand shape depends on the size, shape and surface structure of the object in the hand. These dependencies are difficult to model, which leads us to use a discriminative sequential classifier, conditional random fields (CRF) [17], that does not model the data generation process.

On a semantic level, there are also action-object dependencies of the type `drink-cup`, `drink-glass`, `write-pen`, `draw-pen` and so on, which can be explicitly modeled within the CRF framework. The action-object dependence is modeled on a per-frame basis using a factorial CRF (FCRF) [41]. This is detailed in Section 4.

A manipulation action is here thought of as beginning with the picking-up of an object and ending with the putting-down of that object – also referred to as "manipulation segments" [48]. However, two such actions with the same object might also follow each other directly, without any putting-down and picking-up events in between. The FCRF enables simultaneous object-action classification and temporal action segmentation, removing the need for special tags (e.g., grasping or reaching motions) in the beginning and end of each action [14,35]. This is further discussed in Section 5.

The concept of contextual object-action recognition, as well as the recognition method chosen in this paper, are experimentally evaluated in Section 6. From the experiments it can be concluded that both action and object recognition benefit from the contextual information.

2. Related work

Visual recognition, especially object recognition [8], is a vast area of research and can be regarded as one of the core problems in Computer Vision. We do not make an attempt to review the whole field, but focus on learning of object affordances and contextual recognition.

2.1. Learning of object affordances and learning from demonstration

The concept of affordances [12] has come in focus lately within the Cognitive Vision and Robotics communities. While many other papers on affordances, e.g. [3,33,40], concentrate on robotic grasping, we here focus on more composite, higher-level actions, which typically involve grasping as a sub-component.

The embodied/cognitive vision approach to affordance learning consists of an agent acting upon objects in the environment and observing the reaction. In [10], a robot pushes, pokes, pulls and grasps objects with its end-effector, thereby learning about rolling, sliding, etc. Montesano et al. [24] notes that an affordance can be described by the three interdependent entities of action, object, and effect. A robot first learns a set of affordances by exploration of the environment using preprogrammed basic motor skills. It can then imitate a human action, not by mimicking the action itself, but rather observing the effect and then selecting its own action that will have the same effect on the current object. The difference to our imitation learning is that we also learn the object affordances themselves from human demonstration.

To a certain degree, affordances can be observed in images. In three recent works, [3,33,40], relations between visual cues and grasping affordances are learned from training data. In [40], object grasping areas are extracted from short videos of humans interacting with the objects, while in [3,33] a large set of 2D object views are labeled with grasping points. Early work on functional object recognition [31,39] can be seen as a first step towards recognizing affordances from images. Objects are there modeled in terms of their functional parts, such as handle and hammer-head [31], or by reasoning about shape in association to function [39].

The robot can also learn through visually observing another agent – for example, a human – making use of object affordances. This is the approach we take in this article. A similar idea is also exploited in [45]. However, while they study whole-body activities such as `sitting-on-chair` and `walking-through-door`, we focus on manipulation actions, involving the human hands and arms.

Affordances relate to the concept of *task oriented vision* [15,22]. According to this notion, a Computer Vision system should be designed with a specific task in mind. This is put in contrast to Marr's [20] general purpose vision paradigm. The intended task will affect what aspects of the world are perceived and processed, as well as the design of the whole system. The inspiration comes from human vision; psychophysical experiments [44] indicate that humans indeed only perceive the aspects of the world relevant to the task at hand.

Ikeuchi and Hebert [15] exemplify task oriented vision by comparing two systems designed to solve two different grasping tasks. Miura and Ikeuchi [22] point out that knowledge of the task should be used to ensure that only relevant information is extracted. Although the rapid development of computational power has made this issue less critical today, it is still valid. In our learning from human demonstration method, the robot only includes objects near the human hand in the action-object analysis, rather than trying to model all objects in the scene.

2.2. Contextual recognition

There has been a large recent interest in contextual recognition within the Computer Vision community.

One type of contextual information for object detection and recognition is text. The caption of an image says something about what objects can be expected in it. When labeling images according to object content, any captions should therefore be taken into account. Caption-guided object detection can be used to segment the image into object regions and associate them with object labels [4], or to automatically label or cluster a large set of unlabeled images with captions given a smaller set of labeled images with captions [29]. Prepositions and comparative adjectives can also be used to discover spatial relations between objects in the image [13].

In [42,27,43], the scene itself, the "gist" of the image, is used to guide object recognition. The scene itself is a strong prior cue as to which objects can be expected and where they are most likely to be found. Similarly, in [21], actions and events are recognized in movies in context of the scene. Events can even be recognized from single images [19], if object and scene context is exploited.

Object recognition can also be guided by observations of human interaction with the objects. Moore et al. [25] provide a Bayesian framework for recognizing objects based on contextual information from other objects, human actions being performed on the object, and the scene. In [28], human actions are used to infer object class. Reversely, recognition of manipulation actions can be guided by knowledge about the objects involved. Wu et al. [47] represent kitchen activity solely in terms of the sequence of objects in contact with the hand during the activity. These approaches all relate to the work presented here, with the addition that we perform simultaneous recognition of actions and objects in context of each other.

The idea of simultaneous object-action recognition has been exploited before. In [9], primitive actor-object interactions such as grasp cup, touch spoon, are learned from video. We model more high-level actions, which might involve grasping, touching, etc. Gupta et al. [14] use an approach similar to ours to recognize actions and objects in context of each other. The main difference, apart from our affordance framework, is that they segment manipulation action by detection of reaching motion. We instead incorporate the temporal segmentation into the recognition using a conditional random field. This enables us to recognize actions following each other, without any special delimiter actions such as reaching, putting-down or picking-up. Furthermore, they focus on upper-body or whole body actions while we study hand manipulation actions.

3. Features for classification

For our purposes, extraction of object and action features could be done in a variety of ways [8,23] depending on the purpose of the feature extraction. As opposed to many other action recognition applications, it is here necessary to obtain the location of the human hand to find the object or objects involved. Furthermore, it should be possible to recreate the recognized action with a robot, which means that the hand position, orientation and articulated pose should be retrievable from the action representation. This is further discussed in Section 3.2 below.

3.1. Object features

Different actions involve different number of objects. For example, the action `pour` involves two containers, one to pour from and one to pour to, while the action `sit` involves one piece of furniture. (We do not here separate between tools and other objects; this is further discussed in Section 7.) The object state o_t therefore encodes both the number and the classes of objects involved in the action at time t.

The object state is approximated by a vector x_t^o where each element is the detection probabilities for each object class respectively. At this preprocessing stage, objects are categorized according to appearance into 6 semantic categories of the type shown at the top row of Fig. 1: `book`, `magazine`, `hammer`, `box`, `cup`, and `pitcher`. Section 5 describes how these object classes are grouped according to observed human use.

All objects of the known range of classes in the neighborhood of the hand are detected. We use sliding window detectors, one for each object class. The detector for a certain object class searches over image position, scale and height/width ratio in the image plane, in the vicinity of the human hand (see Section 3.2). The search limits for the sliding window detector in terms of window size, aspect ratio and offset from the human hand are learned from training data. Each window is classified as object or background using a two-class support vector machine (SVM) [6]. Fig. 2 shows example detections of the 6 different object classes.

A representation similar to histograms of oriented gradients (HOG) [7] is used in the SVM classification. Gradient orientation $\mathbf{\Phi} \in [0,\pi)$ is computed from an image window $\mathbf{W}$ as $\mathbf{\Phi} = \arctan\left(\frac{\partial \mathbf{W}}{\partial y} / \frac{\partial \mathbf{W}}{\partial x}\right)$ where x denotes downward (vertical) direction and y rightward (horizontal) direction in the image window. From $\mathbf{\Phi}$, a pyramid with L levels of histograms with different spatial resolutions are created; on each level l, the gradient orientation image is divided into $2^{L-l} \times 2^{L-l}$ equal partitions. A histogram with B bins is computed from each partition. In the SVM classification, a window $\mathbf{W}$ is represented by the vector w which is the concatenation of all histograms at all levels in the pyramid. The length of w is thus $B\sum_{l=1}^{L} 2^{2(L-l)}$. In our experiments in Section 6 we use $B = 4$ and $L = 4$.

In each classification step in the sliding window detection, each object class is treated separately. For each object class, a two-class SVM with an RBF kernel is trained with a set of feature vectors $\{w^{\text{fg}}\}$ containing image bounding boxes with objects, and a set $\{w^{\text{bg}}\}$ containing randomly chosen image windows. This basic object classifier is suitable for deformable objects, where the range of object appearances can not easily be parameterized; the range of appearances spans a manifold with complex shape in the feature space. This manifold is non-parametrically represented using the SVM.[1]

A sequence of object detections is denoted $\mathbf{x}^o = \{x_t^o\}, t = 1, \ldots, T$, where x_t^o is a vector of detection probabilities for the 6 known object classes. Similarly, $\mathbf{o} = \{o_t\}, t = 1, \ldots, T$ denotes the corresponding object state, where o_t is an integer value, indicating which combination of objects is involved in the action at time t.

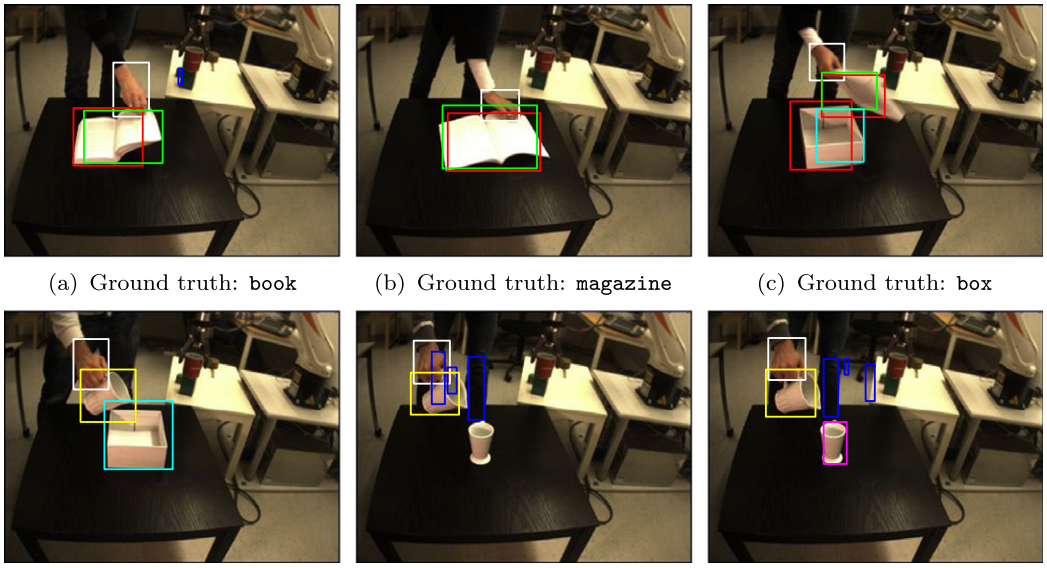
3.2. Action features

A human manipulation action is to a very large degree described by the articulated motion of the hand. We therefore use the hand pose reconstruction method in [32] to reconstruct and track the articulated motion of the hand in 3D.

The method is example based. In each time step, the hand is first segmented from the image using skin color.[2] The appearance of the hand is compared to a large database (on the order of 10^5

[1] For certain types of objects, one can of course create more accurate classifiers [8]. For example, rigid object appearance can be parameterized according to the object orientation with respect to the camera. Such objects are often better classified with several specialized classifiers, trained on different object views. Moreover, human appearance can be parameterized according to orientation, type of activity and phase in the motion cycle. Specialized classifiers can then be trained for different orientations, activities and phases.

[2] In its current form, the method can not separate the hand of interest from faces and other hands visible in the image. A principled preprocessing approach is to maintain an estimate of human pose, and use this estimate to guide the search for the hand region of interest.



(a) Ground truth: book (b) Ground truth: magazine (c) Ground truth: box

(d) Ground truth: box+pitcher (e) Ground truth: cup+pitcher (f) Ground truth: cup+pitcher

Fig. 2. Detection of objects in human action sequences. (a and b) Books and Magazines are difficult to distinguish visually. (c) A Box (and its lid) look similar to a closed Book or a Magazine. (a, e and f) Objects with non-discriminatory appearance (here Hammer) are sometimes "hallucinated". (e) Objects (here Cup) are sometimes missed. (a–f) Color coding: ● = book, ● = magazine, ● = hammer, ● = box, ● = cup, ● = pitcher. *This figure is best viewed in color.*

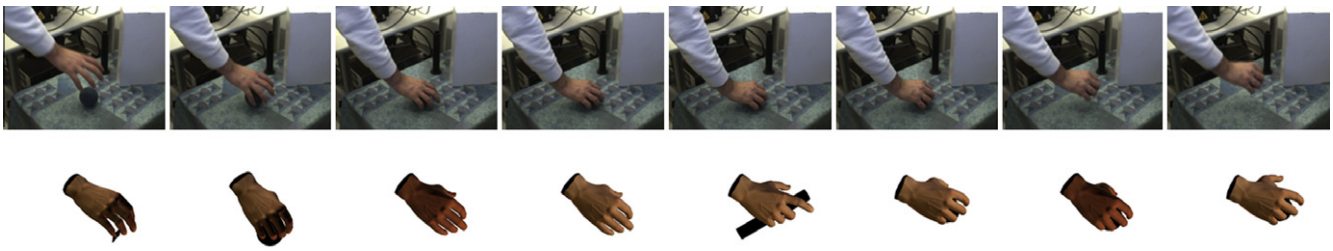


Fig. 3. Reconstruction and tracking of 3D articulated hand pose. Top: original frame. Bottom: reconstructed view (object not included in the pose description).

examples) of synthetic hand views tagged with articulated pose and orientation. Temporal pose consistency is enforced in the reconstruction. Moreover, typical occlusions from objects in the hand are modeled by including object occlusions in some examples in the database.

A reconstruction example can be seen in Fig. 3. The reconstruction is quite crude with angular errors of 10–20%. However, the method is functioning in real-time and is very robust to temporary tracking failure. Our intended application is to understand the relationship between objects and hand actions (e.g., grasps) performed on them on a semantic level. In other words, the object-action classification is intended for qualitative reasoning about which hand actions apply to which objects, rather than for learning precise hand motion from demonstration. Thus, speed and robustness in the hand reconstruction is more critical than accuracy, which makes the method described here suitable for our purposes.

To provide position invariance the global velocity of the hand is encoded, rather than the global position itself. A hand pose is thus defined by global velocity, global orientation, and joint angles. In a manner similar to the object feature extraction, the hand pose at time t is classified as being part of the actions `open`, `hammer`, `pour`, or as not involved in any particular action. A separate two-class SVM is trained for each type of action, rendering an `open/none` classifier, a `hammer/none` classifier, and a `pour/none` classifier.

In the following, a sequence of single-frame action classifications is denoted $\mathbf{x}^a = \{x_t^a\}, t = 1,\ldots,T$, where x_t^a is a vector of classification probabilities for the three action classes. The corresponding action state is denoted $\mathbf{a} = \{a_t\}$, $t = 1,\ldots,T$, where a_t is an integer value indicating action class.

3.3. Correlation between object and action features

The temporal classifier described in Section 4 models explicit semantic dependencies between manipulation actions and the manipulated objects. However, there are also dependencies on the feature level.

The shape of the hand encoded in x_t^a gives cues about the object as well, since humans grasp different types of objects differently, due to object function, shape, weight and surface properties. Similarly, the object detection results encoded in x_t^o is affected by the hand shape since the hand occludes the object in some cases. Furthermore, there are temporal dependencies: x_{t-1}^a and x_t^a are correlated as are x_{t-1}^o and x_t^o. This correlation within the data is *implicit* and difficult to model accurately, but should be taken into account when modeling the simultaneous action-object recognition.

4. Classification of object-Action data

Since we can expect complex dependencies within our action data $\mathbf{x}^a$ and object data $\mathbf{x}^o$ over time, a discriminative classifier which does not model the data generation process is preferable over a generative sequential classifier like a hidden Markov model (HMM) [30]. We thus employ conditional random fields (CRFs) [17]

which are undirected graphical models that represent a set of state variables $\mathbf{y}$, distributed according to a graph $\mathcal{G}$, and conditioned on a set of measurements $\mathbf{x}$. CRFs have previously been used to model human activity, e.g. in [38].

Let $C = \{\{\mathbf{y}_c, \mathbf{x}_c\}\}$ be the set of cliques in $\mathcal{G}$. Then,

$$P(\mathbf{y}|\mathbf{x};\theta) = \frac{1}{Z(\mathbf{x})}\prod_{c\in C}\Phi(\mathbf{y}_c,\mathbf{x}_c;\theta_c), \tag{1}$$

where Φ is a potential function parameterized by θ as

$$\Phi(\mathbf{y}_c,\mathbf{x}_c;\theta_c) = e^{\sum_k \theta_{c,k} f_k(\mathbf{y}_c,\mathbf{x}_c)} \tag{2}$$

and $Z(\mathbf{x}) = \sum_{\mathbf{y}}\prod_{c\in C}\Phi(\mathbf{y}_c,\mathbf{x}_c;\theta_c)$ is a normalizing factor. The feature functions $\{f_k\}$ are given, and training the CRF means setting the weights θ, e.g., using belief propagation [17].

4.1. Linear-chain CRF

For linear-chain data (for example a sequence of object or action features and labels), $\mathbf{y} = \{y_t\}$ and $\mathbf{x} = \{x_t\}$, $t = 1,\ldots,T$ as shown in Fig. 4a. This means that the cliques are the edges of the model, which gives

$$P(\mathbf{y}|\mathbf{x};\theta) = \frac{1}{Z(\mathbf{x})}\prod_{t=2}^{T}\Phi(y_{t-1},y_t,\mathbf{x};\theta_t) \tag{3}$$

with a potential function

$$\Phi(y_{t-1},y_t,\mathbf{x};\theta_t) = e^{\sum_k \theta_{t,k} f_k(y_{t-1},y_t,\mathbf{x})}. \tag{4}$$

Each state y_t can depend on the whole observation sequence $\mathbf{x}$ – or any subpart of it, e.g. the sequence $\{x_{t-\mathscr{C}},\ldots,x_{t+\mathscr{C}}\}$, $\mathscr{C}$ being the *connectivity* of the model.

4.2. Factorial CRF

In Section 3.3 we argue that there are correlations between action observations $\mathbf{x}^a$ and object observations $\mathbf{x}^o$ implicit in the data. We make use of this correlation on the data level by not imposing a simplified model on the data generation process and instead using a discriminative classifier, CRF. However, there is also an *explicit*, semantic correlation between actions and objects on the label level, as discussed in the introduction. This correlation can be modeled using a factorial CRF (FCRF) [41]. Fig. 4b shows an FCRF with two states, action class a_t and object class o_t, for three time steps $t = 1, 2, 3$. The cliques in this model are the within-chain edges $\{a_{t-1}, a_t\}$ and $\{o_{t-1}, o_t\}$, and the between-chain edges $\{a_t, o_t\}$. The probability of $\mathbf{a}$ and $\mathbf{o}$ is thus defined as

$$P(\mathbf{a},\mathbf{o}|\mathbf{x};\theta) = \frac{1}{Z(\mathbf{x})}\prod_{t=1}^{T}\Phi(a_t,o_t,\mathbf{x};\theta_t) \times \prod_{t=2}^{T}\Phi(a_{t-1},a_t,\mathbf{x};\theta_{a,t})\Phi(o_{t-1},o_t,\mathbf{x};\theta_{o,t}). \tag{5}$$

The weights θ are obtained during training, e.g., using loopy belief propagation [41].

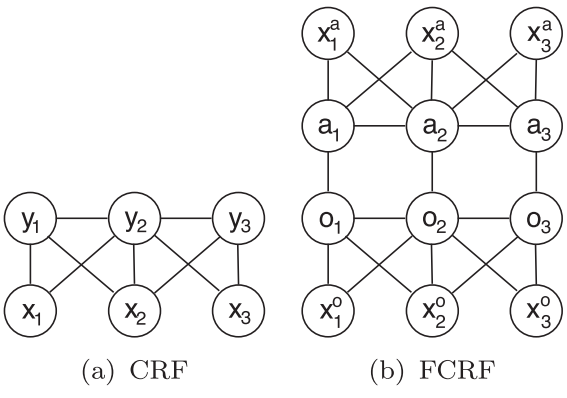


Fig. 4. CRF structures. (a) Linear-chain CRF [17], used for action or object recognition. (b) Factorial CRF [41], used for simultaneous object-action recognition.

5. Object-action recognition using CRF:s

Using the approach described above, the actions and objects in a stereo sequence of human activity can be both temporally segmented and classified, using a CRF in a sliding window manner over time.

An FRCF structure of length $T = 3$ is trained with $\mathbf{X}$ object-action patterns ($\mathbf{o}$, $\mathbf{x}^o$, $\mathbf{a}$, $\mathbf{x}^a$), also of length $T = 3$, involving $\mathbf{Y}$ different action classes and $\mathbf{Z}$ different object classes.

A new sequence ($\boldsymbol{\chi}^\omega$, $\boldsymbol{\chi}^\alpha$) of length τ can now be segmented and classified using this model. For each time step $t = 2,\ldots,\tau-1$, the pattern $(\chi^\omega_{t-1},\chi^\omega_t,\chi^\omega_{t+1},\chi^\alpha_{t-1},\chi^\alpha_t,\chi^\alpha_{t+1})$ renders the classification ω_t, α_t.

Objects can also be ordered into affordance categories using correlation information extracted from the training data ($\mathbf{o}$, $\mathbf{a}$). This is represented with a correlation matrix $\mathbf{C}$ where element C_{ij} indicates the degree to which object class i can be used to perform action j.

6. Experiments

The feature extraction and classifiers were implemented in Matlab, using the LibSVM toolbox [5] and the CRF toolbox by [26]. The object and action feature extraction methods described in Section 3 were first evaluated (Sections 6.1 and 6.2). We then evaluated the temporal object-action segmentation and classification. This is described in Section 6.3.

6.1. Evaluation of object classifier

HOG-like features has previously been shown [7] to be good image representations, since it allows for high intra-class variability (differences between class instances, lighting and pose variation, etc.) while being discriminant with respect to inter-class variability. To verify this, we evaluated the HOG- and SVM-based classifier which is the basis of the sliding window object detector described in Section 3.1.

We first experimented with the NORB dataset [18], which contains five different classes of rigid objects; `animals`, `humans`, `airplanes`, `trucks`, and `cars` with 10 instances of each, five for test and five for training (Fig. 5). The database contains stereo views of each object from 18 different azimuths and nine elevations in six different lighting conditions. Only the normalized-uniform part of the dataset, designed to test classification performance, was used.

To evaluate the suitability of the HOG-like feature representation (Section 3.1) for modeling shape *categories*, a five-class SVM was trained with features extracted from the NORB training images. Table 1 left shows the results compared to others. Our classifier reached the same classification accuracy as a state-of-the-art method for object categorization [18], which indicates that our representation captures the specifics of a shape class, while allowing a significant variability among instances of that class. In comparison, training an SVM on the raw image downsampled to a size of 32×32 led to twice the classification error (a surprisingly good result, as noted in [18], given that the task is object categori-

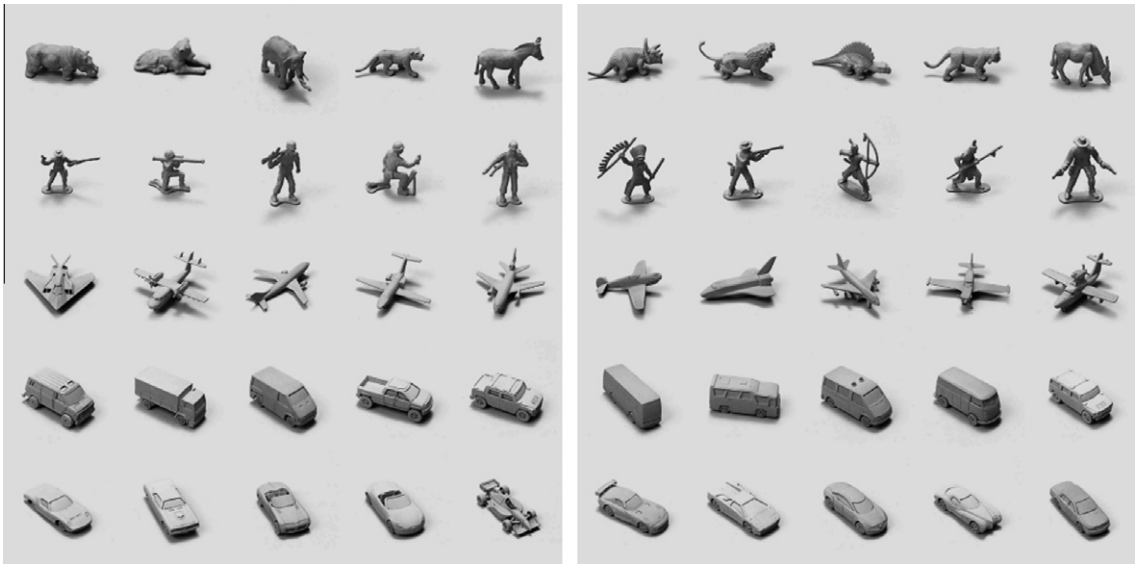


Fig. 5. The 50 instances in the normalized-uniform NORB dataset [18] (for one lighting condition, elevation, azimuth each). Training data left, test data right.

Table 1
Results on the normalized-uniform NORB dataset, percent error. (a) Classification error percentage compared to results from [18], marked (LC). (b) Generalization test; robustness to different amount of jitter in test data (training data unaltered).

	Mono	Stereo		Brightness			Shift		
(a) Classification error			*(b) Robustness to jitter*						
Hist+SVM	6.4	6.2		±0	±10	±20	±30	±3	±6	±9
Raw+SVM	12.6 (LC)	–	Hist+SVM	6.4	6.4	7.1	8	10.3	18.1	29.2
Conv Net	–	6.6 (LC)	Raw+SVM	12.6 (LC)	15.8	18	21	20.8	35.1	48.6

zation, not instance recognition). Furthermore, we note that the incorporation of stereo does not add much to the accuracy.

Certain robustness towards differences in color and lighting, as well as small position errors of the object detection window, is also desirable. In [18], this was tested by adding "jitter", i.e. small transformations to both the training and test set. However, this arguably tested how the methods performed with a larger test set, rather than how they could handle noise that was not seen before (not present in the training data). Therefore, we did a variant of this experiment where we added jitter to *only the test set* (Table 1 right). First, the overall brightness of each test image was varied. Our feature representation was very robust to this noise, which is expected since it relies solely on the gradient orientations and not on their value. In comparison, the raw image classification error grew much quicker. Then, the test images were shifted vertically and horizontally in a random manner. The feature representation was more sensitive to this noise, but less so than the raw image representation.

We then proceed to evaluate the performance of our classifier for classification of the six object categories described in Section 3.1. We collected a dataset consisting of 4–6 different instances of each class (Fig. 6a), with 330 views of each instance. Each object instance was grasped and moved by a human in some views, and the deformable objects were deformed (opened, pages flipped, etc.) in other views. In each view, a quadratic bounding box of the object was manually marked in the image. Our feature representation used for training and classification was then computed over this window.

Training and testing were carried out in a jackknife manner, where one instance at a time of each class were removed from the dataset during training, and used for testing. (Thus, the same instance was never used for both training and testing.) For each training-test data division, a six-class SVM was trained with the feature representations. Fig. 6b shows the confusion matrix representing the mean classification result.

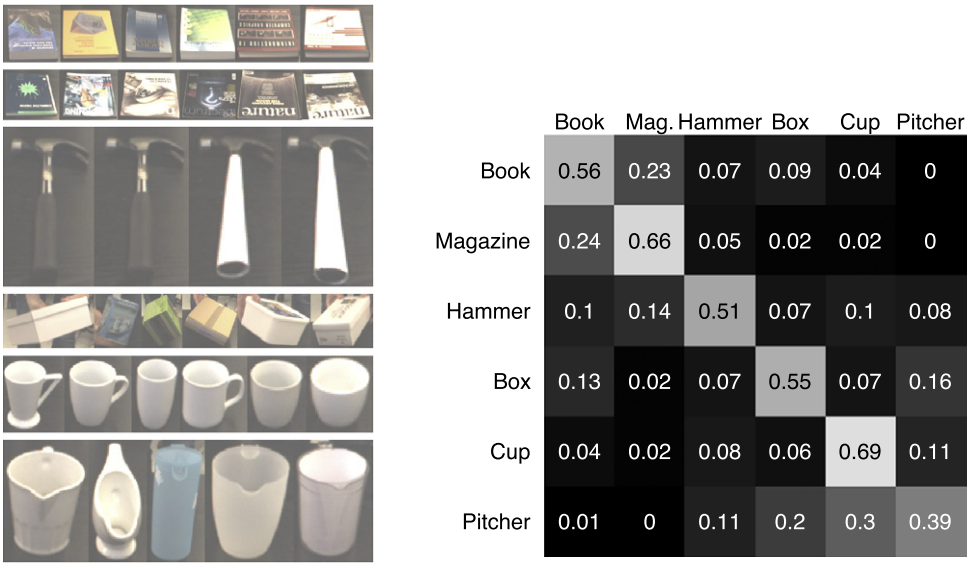
The experiments with the NORB dataset above indicated that our representation allows certain intra-class variation, while being rich enough to make inter-class discrimination possible. The confusion between `book` and `magazine`, between `book` and `box`, and between `box` and `pitcher` is therefore probably intrinsic to the classes themselves. Books and magazines look very similar both closed and with pages flipped (23% and 24% misclassification) – the main difference is the thickness of the volume. Boxes look like closed books (13% misclassification), but opened books and books with pages flipped do not look like boxes (9% misclassification). Hammers are hard to model with this classifier (7–14% misclassifications against other classes) since most bounding boxes around hammers contains very much background. Pitchers are often misclassified as boxes (20% misclassification) since many of them look quadratic from a side view, and as cups (30% misclassification) since the shape of those object classes are very similar, with handles and openings at the top.

The classes `book` and `magazine`, and `pitcher` and `cup`, are good examples of object classes that are hard to distinguish from appearance only.

6.2. Evaluation of action classifier

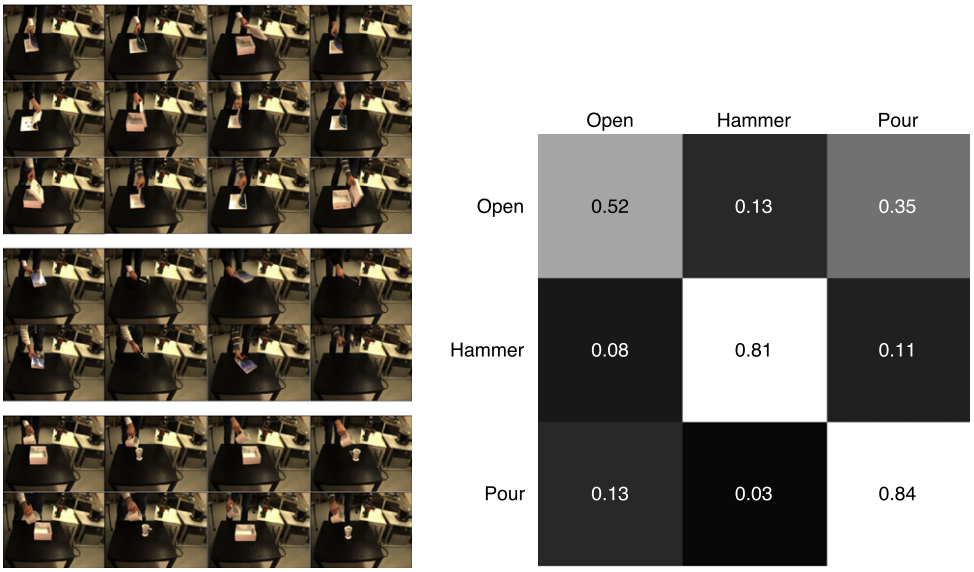
The action classification described in Section 3.2 was evaluated in a similar manner. A dataset consisting of 8–12 examples of each class of actions, performed with different objects and by four different individuals, was collected (Fig. 7a).

Training and testing were done similarly to the above, where the examples from one individual at a time were removed from



(a) Dataset (b) Confusion matrix

Fig. 6. Classification of objects of the six classes used in this paper. (a) The 33 instances used in the experiment. Three hundred and thirty views of each instance are provided. Each object is grasped and moved by a human in some views, and the deformable objects are deformed (opened, pages flipped, etc.) in other views. (b) Object classification confusion matrix (rows: true classes, columns: classification ratios).



(a) Dataset (b) Confusion matrix

Fig. 7. Classification of actions of the three classes used in this paper. (a) The 28 instances used in the experiment. Each instance is a sequence of hand velocities and articulated hand poses, extracted separately from each frame. Each frame is considered as a separate data point. (b) Action classification confusion matrix (rows: true classes, columns: classification ratios).

the training data. Training was done on examples from the three other individuals, and the resulting classifier was evaluated with the examples of the fourth individual. Fig. 7b shows the confusion matrix representing the mean classification result.

As with the object classification, there are certain confusions intrinsic to the actions themselves. Most notably, `open` actions are often misclassified as `pour` (35% misclassification), probably since the global hand velocity is similar in these action, and since different individuals configure their hands very differently while opening objects. On the other hand, pouring actions are much more distinct in terms of hand articulation and velocity, and are more seldom misclassified as opening (13% misclassification). Due to its rapid vertical velocity, `hammer` actions are easily recognizable.

The classes `open` and `pour` are good examples of action classes that are hard to distinguish without contextual information, e.g., from objects involved in the action.

6.3. Classifying actions and objects together

Experiments with the object and action feature extractors in Sections 6.1 and 6.2 showed certain confusions between classes, which appeared to be intrinsic to the object and action classes themselves. For example, books and magazines can not always be distinguished by appearance only. However, they afford slightly different ranges of actions, which means that the action observed in connection to the object can be used to constrain the object

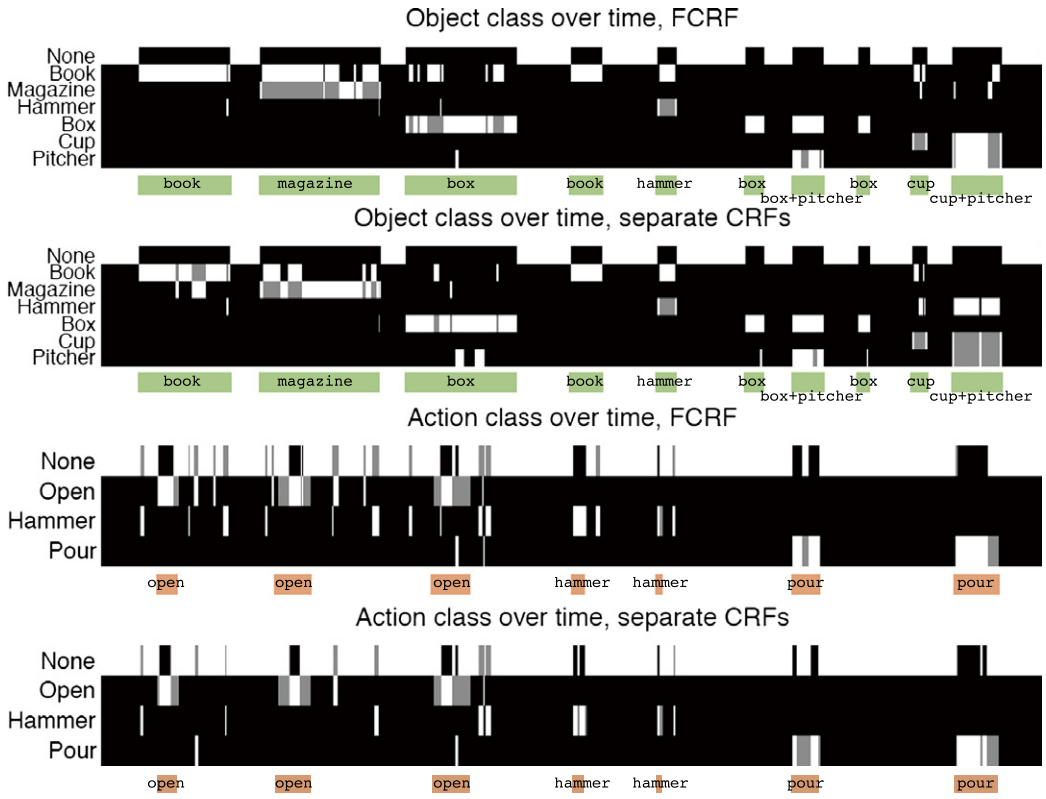


Fig. 8. Object-action classification over time. The depicted sequence contains seven object-action combinations: `open-book`, `open-magazine`, `open-box`, `hammer-with-book`, `hammer-with-hammer`, `pour-from-pitcher-into-box`, `pour-from-pitcher-into-cup`. Time on x axis, Classification on y axis. White block = (F)CRF classification during this time period. Grey block = classification ground truth during this time period. Ground truth object and action classifications are also indicated with blocks below each diagram.

classification. Similarly, action classification can be constrained by the objects observed in the vicinity of the hand performing the action.

In this experiment, we trained and evaluated an FCRF (Section 5) with the 6 object classes and the three action classes mentioned above, in the seven following combinations: `open-book`, `open-magazine`, `open-box`, `hammer-with- book`, `hammer-with-hammer`, `pour-from-pitcher-into-box`, and `pour-from-pitcher-into-cup`. Thus, an `open` action can only be observed together with either a `book`, a `magazine` or a `box`; a `hammer` action only together with a `book` or a `hammer`; a `pour` action only together with a `box` and `pitcher` or a `cup` and `pitcher`.

We collected a training set consisting of sequences in which three different individuals performed all seven object-action combinations, using the first object instance from each class in the evaluation set in Fig. 6a. Each frame of the sequences were labeled with object (`none`, `book`, `magazine`, `hammer`, `box`, `cup`, `pitcher`, `box+pitcher`, or `cup+pitcher`) and action (`none`, `open`, `hammer`, or `pour`) ground truth. Only the frames with hand detections were used for training.

The evaluation set consisted of a sequence where a fourth individual performed the same object-action combinations, using the same object instance. Only the frames with hand detections were used for evaluation. (The frames with no hand were automatically labeled as object `none`, action `none`.)

From each frame of each of the four sequences, object features were extracted as described in Section 3.1. The 6 background/object classifiers were trained with images of a fifth person handling the object instances in the same way as in the training and evaluation sequences. Action features were extracted as described in Section 3.2. The three background/action classifiers were trained on hand poses and velocities from the three training sequences.

An FCRF was trained with sequences of $T = 3$ consecutive frames of object and action features from the training data, in all 1471 training examples. This FCRF can be expected to learn

1. allowed object-action combinations,
2. allowed temporal action transitions (in this dataset, only `none`-action and action-`none`),
3. typical errors in the per-frame object feature extraction,
4. typical errors in the per-frame action feature extraction.

To provide a baseline, two individual CRFs were trained with sequences of $T = 3$ consecutive frames of object features and action features, respectively. The object CRF can be expected to learn aspect 3 above, and the action CRF can be expected to learn aspects 2 and 4; none of them capture aspect 1.

The object-action FCRF and the two individual action and object CRFs were used to classify the evaluation sequence. The classification result is shown in Fig. 8.

First, it can be noted that many frames of the sequence contains an object but action `none`; in other words, the human is doing something else with the object than opening, hammering or pouring. In these frames, the object classification in the FCRF (Fig. 8, row 1) is not supported by more information than the classification in the separate object CRF (Fig. 8, row 2), since all combinations of objects and action `none` are present in the training data. In the remainder of the analysis, we therefore focus on the 184 frames where there is an `open`, `hammer` or `pour` action taking place (red[3] blocks below the diagrams in Fig. 8, rows 3 and 4).

[3] For interpretation of color in Figs. 2 and 8, the reader is referred to the web version of this article.

From the results in Fig. 8 we can conclude that both object and action recognition are improved by the contextual information: For the frames with an action taking place, the separate CRFs have a correct object classification rate of 52% and a correct action classification rate of 46%. The FCRF, which takes contextual recognition into regard, has a correct object classification rate of 60% and a correct action classification rate of 58%.

However, from an application perspective we are not primarily interested whether each frame of an object-action combination are correctly classified; the main focus is instead on whether the action-object combination is detected and correctly classified at all. From this perspective, both the FCRF and the individual CRFs detected all seven object-action combinations, i.e., classified some frames of each object-action combination as something other than object `none`, action `none`.

The classification of the detected object-action combination is here defined as the majority vote among the classifications in the frames of the detection. The FCRF (Fig. 8, rows 1 and 3) detected the seven following object-action combinations: `open-book` (correct), `open-book` (incorrect but allowed),[4] `open-box` (correct), `hammer-with-book` (correct), `hammer-with-hammer` (correct), `pour-from-pitcher-into-box` (correct), `pour-from-pitcher-into- cup` (correct). This concords with the findings in the experiments with the object feature extraction above: Books are often detected as magazines and vice versa (see also Fig. 2a and b), and they both afford opening, which means that the contextual action information could not guide the object classification in the second object-action combination. The inclination to classify the magazine as `book` in Fig. 8, rows 1 and 2 could be due to coincidences in the training and evaluation data – there were more images of books than magazines in the training data with the same orientation as the magazine in the evaluation data.

The two individual CRFs (Fig. 8, rows 2 and 4) detected the seven following object-action combinations: `open-book` (correct), `open-book` (incorrect but allowed), `open-box` (correct), `hammer-with-book` (correct), `hammer-with- hammer` (correct), `pour-from-pitcher-into-box` (correct), `pour-from-hammer` (incorrect).[5] In the last combination, the object detection was inadequate by itself, but the contextual action information in the FCRF helped in inferring the correct object classes (Fig. 8, row 1). Furthermore, the actions are more accurately detected by the FCRF in the two last combinations (Fig. 8, row 3), than by the individual action CRF (Fig. 8, row 4). The reason is most certainly the contextual object information provided by the FCRF.

This shows that the FCRF is able to infer information about the object and action present in a frame, not immediately apparent from the present image information, from knowledge about which object-action combinations are commonly observed in other data.

The many spurious detections, particularly of `hammer` actions, would pose problems to a learning from demonstration system employing the classification method. One way to address this problem is to increase the number of time steps in the FCRF, e.g., to use five or seven time steps instead of three. However, this increases the number of parameters to learn; a larger set of example sequences is then required to train the FCRF. Another measure to take is to improve the feature extractors, so that the FCRF is fed with cleaner data. This also requires much larger and more diverse datasets; more individuals, more object instances, more action instances performed by each individual.

[4] This object-action combination is allowed since it is observed in the training data; books and magazines both afford opening.

[5] This object-action combination is not allowed; hammers do not afford pouring.

7. Conclusions

This paper investigated object categorization according to function, i.e., learning the affordances of objects from human demonstration. More specifically, we presented a method for classifying objects grasped and manipulated by a human in context of which actions the human was involved in, and at the same time, classifying human actions in context of the object involve in the action. An FCRF was trained with short sequences of simultaneously extracted object and action features, modeling both information about objects and action detection, and the likelihood of observing different object-action combinations.

Experiments with a dataset of combinations of three actions and 6 objects showed that the FCRF captured contextual dependencies which could be used to infer information about both actions and objects not present in the image data. This improved the classification of both actions and objects.

7.1. Future Work

In the applications of interest here, primarily learning from demonstration, the requirement of fully labeled training data is a limiting factor. Our intention is therefore to develop methods for *semi-supervised training* of the FCRF, e.g., using co-training [2] with an object and an action view.

A related avenue of research is the introduction of *grammatical structures* to describe human activity [37]. These structures can be considered as contextual information, guiding the classification of individual actions and objects. The structures can be learnt in a semi-supervised manner from a combination of the demonstrations themselves and the human demonstrator's utterances during the demonstration.

A slightly more philosophical question regards the different roles of objects in actions. E.g., the action-object combination `hammer-with-hammer- on-nail` contains two objects, where `hammer` is a tool and `nail` is not. Tools are tricky when reasoning about affordances [12] – when used, they can be regarded as part of the agent's body. (A robotic agent can very well have a hammer permanently attached to its body.) Thus, the division between agent, objects and scene is not clear [46]. At present, our method does not differ between tools and other objects in any principal way. However, this will be addressed in future work.

Acknowledgments

This research has been supported by the EU through PACO-PLUS, FP6-2004-IST-4-27657, and by the Swedish Foundation for Strategicv Research.

References

[1] A. Billard, S. Calinon, R. Dillman, S. Schaal, Robot programming by demonstration, in: B. Siciliano, O. Khatib (Eds.), Handbook of Robotics, Springer-Verlag, New York, NY, USA, 2008 (Chapter 59).
[2] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Conference on Computational Learning Theory, 1998.
[3] J. Bohg, D. Kragić, Grasping familiar objects using shape context, in: International Conference on Advanced Robotics, 2009.
[4] P. Carbonetto, N. de Freitas, K. Barnard, A statistical model for general contextual object recognition, in: European Conference on Computer Vision, vol. 1, 2004, pp. 350–362.
[5] C.-C. Chang, C.-J. Lin, LIBSVM: A Library for Support Vector Machines, 2001. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
[6] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, UK, 2000.
[7] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2005, pp. 886–893.

[8] L. Fei-Fei, R. Fergus, A. Torralba, Recognizing and Learning Object Categories: Short Course at ICCV, 2009. <http://people.csail.mit.edu/torralba/shortCourseRLOC>.
[9] R. Filipovych, E. Ribeiro, Recognizing primitive interactions by exploring actor-object states, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[10] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, G. Sandini, Learning about objects through action – initial steps towards artificial cognition, in: IEEE International Conference on Robotics and Automation, 2003.
[11] W.T. Freeman, M. Roth, Orientational histograms for hand gesture recognition, in: IEEE International Conference on Automatic Face and Gesture Recognition, 1995.
[12] J. Gibson, The Ecological Approach to Visual Perception, Lawrence Erlbaum, Associates, Hillsdale, NJ, USA, 1979.
[13] A. Gupta, L.S. Davis, Beyond nouns: exploiting prepositions and comparative adjectives for learning visual classifiers, in: European Conference on Computer Vision, 2008.
[14] A. Gupta, A. Kembhavi, L.S. Davis, Observing human-object interactions: using spatial and functional compatibility for recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (10) (2009) 1775–1789.
[15] K. Ikeuchi, M. Hebert, Task oriented vision, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, 1992, pp. 2187–2194.
[16] H. Kjellström, J. Romero, D. Martínez, D. Kragić, Simultaneous visual recognition of manipulation actions and manipulated objects, in: European Conference on Computer Vision, vol. 2, 2008, pp. 336–349.
[17] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: International Conference on Machine Learning, 2001.
[18] Y. LeCun, F.J. Huang, L. Bottou, Learning methods for generic object recognition with invariance to pose and lighting, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2004, pp. 97–104.
[19] L.-J. Li, L. Fei-Fei, What, where and who? Classifying events by scene and object recognition, in: IEEE International Conference on Computer Vision, 2007.
[20] D. Marr, Vision, Freeman, New York, NY, USA, 1982.
[21] M. Marszalek, I. Laptev, C. Schmid, Actions in context, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[22] J. Miura, K. Ikeuchi, Task-oriented generation of visual sensing strategies, in: IEEE International Conference on Computer Vision, 1995.
[23] T.B. Moeslund, A. Hilton, V. Krüger, A survey of advances in computer vision-based human motion capture and analysis, Computer Vision and Image Understanding 104 (2–3) (2006) 90–126.
[24] L. Montesano, M. Lopes, A. Bernardino, J. Santos-Victor, Learning object affordances: from sensory motor coordination to imitation, IEEE Transactions on Robotics 24 (1) (2008) 15–26.
[25] D.J. Moore, I.A. Essa, M.H. Hayes, Exploiting human actions and object context for recognition tasks, in: IEEE International Conference on Computer Vision, 1999.
[26] K. Murphy, A CRF implementation for general graphs, 2004. <http://people.cs.ubc.ca/~murphyk/Software/CRF/crf.html>.
[27] K. Murphy, A. Torralba, W.T. Freeman, Using the forest to see the trees: a graphical model relating features, objects, and scenes, in: Neural Information Processing Systems, 2003.
[28] P. Peursum, G. West, S. Venkatesh, Combining image regions and human activity for indirect object recognition in indoor wide-angle views, in: IEEE International Conference on Computer Vision, vol. 1, 2005, pp. 82–89.
[29] A. Quattoni, M. Collins, T. Darrell, Learning visual representations using images with captions, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[30] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–286.
[31] E. Rivlin, S.J. Dickinson, A. Rosenfeld, Recognition by functional parts, Computer Vision and Image Understanding 62 (2) (1995) 164–176.
[32] J. Romero, H. Kjellström, D. Kragić, Hands in action: real-time 3D reconstruction of hands in interaction with objects, in: IEEE International Conference on Robotics and Automation, 2010.
[33] A. Saxena, J. Driemeyer, A.Y. Ng, Robotic grasping of novel objects using vision, International Journal of Robotics Research 27 (2) (2008) 157–173.
[34] S. Schaal, Is imitation learning the route to humanoid robots?, Trends in Cognitive Sciences 3 (1999) 233–242
[35] I. Serrano Vicente, V. Kyrki, J. Kragić, Action recognition and understanding through motor primitives, Advanced Robotics 21 (13) (2007) 1473–1501.
[36] G. Shakhnarovich, P. Viola, T. Darrell, Fast pose estimation with parameter sensitive hashing, in: IEEE International Conference on Computer Vision, vol. 2, 2003, pp. 750–757.
[37] Y. Shi, A. Bobick, I. Essa, Learning temporal sequence model from partially labeled data, in: IEEE Conference on Computer Vision and Pattern Recognition, 2003, pp. 1631–1638.
[38] C. Sminchisescu, A. Kanujia, D. Metaxas, Conditional models for contextual human motion recognition, Computer Vision and Image Understanding 104 (2-3) (2006) 210–220.
[39] L. Stark, K. Bowyer, Generic Object Recognition using Form and Function, W. Sci. Ser. Machine Perception and Artificial Intelligence, vol. 10, 1996.
[40] M. Stark, P. Lies, M. Zillich, J. Wyatt, B. Schiele, Functional object class detection based on learned affordance cues, in: International Conference on Computer Vision Systems, 2008, pp. 435–444.
[41] C. Sutton, K. Rohanimanesh, A. McCallum, Dynamic conditional random fields: factorized probabilistic models for labeling and segmenting sequence data, In: International Conference on Machine Learning, 2004.
[42] A. Torralba, Contextual priming for object detection, International Journal of Computer Vision 53 (2) (2003) 169–191.
[43] A. Torralba, K. Murphy, W.T. Freeman, Contextual models for object detection using boosted random fields, in: Neural Information Processing Systems, 2004.
[44] J. Triesch, D.M. Ballard, M.M. Hayhoe, B.T. Sullivan, What you see is what you need, Journal of Vision (3) (2003) 86–94.
[45] M. Veloso, F. von Hundelshausen, P.E. Rybski, Learning visual object definitions by observing human activities, in: IEEE-RAS International Conference on Humanoid Robots, 2005.
[46] F. Wörgötter, A. Agostini, N. Krüger, N. Shylo, B. Porr, Cognitive agents – a procedural perspective relying on the predictability of object-action-complexes (OACs), Robotics and Autonomous Systems 57 (4) (2009) 420–432.
[47] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, J.M. Rehg, A scalable approach to activity recognition based on object use, in: IEEE International Conference on Computer Vision, 2007.
[48] R. Zöllner, M. Pardowitz, S. Knoop, R. Dillman, Towards cognitive robots: building hierarchical task representations of manipulations from human demonstration, in: IEEE International Conference on Robotics and Automation, 2005, pp. 1535–1540.