Simultaneous Visual Recognition of Manipulation Actions and Manipulated Objects

Hedvig Kjellström Javier Romero David Martínez Danica Kragić

Computational Vision and Active Perception Lab
School of Computer Science and Communication
KTH, SE-100 44 Stockholm, Sweden
hedvig,jrgn,davidmm,dani@kth.se

Abstract. The visual analysis of human manipulation actions is of interest for e.g. human-robot interaction applications where a robot learns how to perform a task by watching a human. In this paper, a method for *classifying manipulation actions in the context of the objects manipulated*, and *classifying objects in the context of the actions used to manipulate them* is presented. Hand and object features are extracted from the video sequence using a segmentation based approach. A shape based representation is used for both the hand and the object. Experiments show this representation suitable for representing generic shape classes. The action-object correlation over time is then modeled using conditional random fields. Experimental comparison show great improvement in classification rate when the action-object correlation is taken into account, compared to separate classification of manipulation actions and manipulated objects.

1 Introduction

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. The analysis of human manipulation is of interest for work-flow optimization, automated surveillance, and programming by demonstration (PbD) applications, in which a robot learns how to perform a task by watching a human do the same task.

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 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.

A manipulation action is here defined as beginning with the picking-up of an object and ending with the putting-down of that object. Only one-hand actions are considered, although this is not a limitation to the method in a formal sense. In a video sequence of the action, the human head and hand are segmented and

tracked using skin color, and objects are segmented as being in the neighborhood of the hand and moving with it.

The action state space in each frame is the image position of the hand relative to the face, and the shape of the hand, represented with a gradient orientation histogram pyramid [1, 2]. Section 5 shows the inclusion of hand shape in the state space to greatly improve action recognition compared to only hand position.

Objects in this application are "graspable" i.e. fairly rigid, so shape is a good object descriptor. We use pyramids of gradient orientation histograms for representation of object shape as well as hand shape. Experiments in Section 5 show this representation to lead to a state-of-the-art classification performance on the NORB dataset [3] which contains objects of this type. Specific for our application is that the classification method has access to several views of the object over the course of the action, something that improves the recognition. Section 3 describes the feature extraction.

There are implicit, complex interdependencies in the object and action data. The sequence of object viewpoints, as well as occlusion from the hand depend on the action; i.e. what the hand is doing with the object. 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) [4], 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 can be modeled on a per-frame basis using a factorial CRF (FCRF) [5]. However, it might be the case that the dependencies between particular frames are weaker than the dependence beween the action and the object as whole. To model sequence-level dependence, we introduce a CRF stucture which we call *connected hierarchic CRF:s (CHCRF)*. This is detailed in Section 4.

Experiments in Section 5 show three things. Firstly, CRF structures with many degrees of freedom, such as structures with hidden nodes or large data connectivity, perform worse than simple structures when the amount of training data is limited. Secondly, the correlated action-object recognition outperform separate classification, and CHCRF:s perform better than FCRF:s on the action-object classification task. Last, the information on actions implicit in the object data is redundant to the information on objects in the action data.

The primary contribution of this paper is the idea of recognizing manipulation actions and manipulated objects in context of each other, while secondary contributions are the definition of the CHCRF and the representation of object shape using pyramids of gradient orientation histograms.

2 Related Work

Actions. In the last few years, considerable research effort has been spent on the analysis of human motion from video [6]. For the purpose of detecting atomic

actions in video, Laptev and Pérez [7] use a boosted classifier of spatiotemporal volumes of optical flow. This approach is robust to significant changes in scale, appearance and viewpoint. Our method differs from [7] in that it is possible to model actions with a longer extension in time, possibly with a sub-structure. Furthermore, since our analysis involves the manipulated object, the hand needs to be located, putting different constraints on our feature extraction.

The analysis of hand motion is most often applied to gesture recognition for human-computer interfaces or sign language recognition [8]. These applications are often characterized by low or no occlusion of the hands from other objects, and a well defined and visually disparate set of hand poses; in the sign language case the gestures are designed to be easily separable to simplify fast communication. In contrast, the manipulation actions which we investigate suffer from large intra-class variability and sometimes occlusion of parts of the hand from the manipulated objects.

Feature extraction for hand action classification often means tracking the hand in 2D [9] or 3D [10, 11]. However, to be able to handle low video framerates we prefer to use a segmentation-based method for human pose recovery not relying on time-incremental estimation [12–14].

Objects. Object recognition 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 contextual object recognition and the representation of shape in object recognition.

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 [15], or to automatically label or cluster a large set of unlabeled images with captions given a smaller set of labeled images with captions [16].

In [17–19], 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. CRF:s have also been used [19, 20] to automatically learn sub-structure; the relations between different parts of the object or between different objects and the scene.

Earlier work on contextual object recognition [21, 22] has focused on functional object recognition; objects are then classified in terms of their function. This is similar in spirit to our contextual recognition; object classes are here defined by how the objects are used (in which action context they appear), and classes of manipulation actions are defined by the class (or classes) of objects that are involved in the action.

Modeling shape is difficult; an important tradeoff is the sparseness of the representation (from silhouettes [13], via edge maps [14, 23], to maps of gradient orientation) versus the robustness towards differences in lighting and fine object texture. We use pyramids of localized histograms of gradient orientation, a representation robust to small position and fine texture differences, while containing more texture information than e.g. edge maps.

Actions and objects. Moore et al. [24] provide a Bayesian framework for recognizing scenes, objects in it, and actions being performed on the objects. A system such as this could be the framework for our method: While their system keeps track the hands of a human, objects handled by the human, and which actions are performed on which objects, our method is concerned with classifying a single action and object.

Wu et al. [25] continue along this line, learning a dynamic Bayesian network model that represent *temporal sequences* of actions and objects involved in the actions. The features used for classification are RFID tags attached to the objects and the human hand. Again, our method could be incorporated in such a system, replacing the RFID-based classificiation with simultaneous classification of objects and actions from video.

Earlier on, Mann and Jepson [26] use a force-dynamic bottom-up approach to describe the interaction between hands and objects in scenes. In contrast, we use a statistical formulation, since the underlying process generating the video sequences in our case is far too complex to be modeled deterministically.

3 Features for Classification

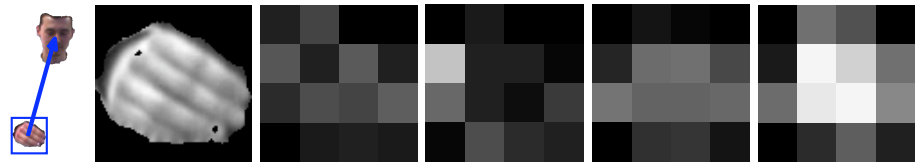
Extraction of image features could be done in a variety of ways [6] depending on the purpose of the feature extraction. Features representing the human motion and appearance as well as object motion and appearance are of interest. As opposed to many other action recognition applications, like video annotation [7], it is here necessary to obtain the location of the human hand to find the manipulated object.

Considering the low framerate, large motion blur and low resolution of the hands of our video sequences, and that articulated hand tracking is a difficult problem [10, 11], we use a segmentation based approach. The hand and face of the human is localized using skin color segmentation [27] and hand and face masks are extracted from the skin mask using connected components detection or with an $\alpha\beta$-filter when hand and face blobs merge.

Other cues than skin color, e.g. combinations of spatial or spatiotemporal filters, can of course also be exploited for the localization of hands and face.

The object involved in the manipulation action is in the human's right hand. To focus the attention of the object classification onto only that object, an object segmentation mask is also obtained, right of the hand in the image (based on the assumption that the object is in that area if grasped by the human). While the position of the area is automatically obtained from the hand position, the area shape is selected to fit the object in the first frame of the sequence, and is then held constant throughout the sequence.

Manipulation actions are here defined as beginning with a pick-up event and ending with a put-down event. With a fully automatic object segmentation, it is possible in each time-step to detect whether there is an object in the hand or not, so that the temporal segmentation can be done automatically.



(a) Pos (b) Image (c) $\pi/8$ (d) $3\pi/8$ (e) $5\pi/8$ (f) $7\pi/8$

Fig. 1. Features used for action classification. a) 2D position p_t of the centroid of the right hand segment relative to centroid of the face segment. b) Hand image $\mathbf{I}_t^a$. c–f) Gradient orientation histograms from $\mathbf{I}_t^a$, with $B = 4$ bins, on level $l = 1$ in the pyramid of $L = 3$ levels. c) Bin 1, orientation $\pi/8$. d) Bin 2, orientation $3\pi/8$. e) Bin 3, orientation $5\pi/8$. f) Bin 4, orientation $7\pi/8$.

Action features. We seek a representation of hand motion that captures both the global hand position and the articulated hand pose over time. The global position of the hand at time t is represented with the 2D position p_t^a of the centroid of the hand mask relative to the centroid of the face mask (Figure 1a).

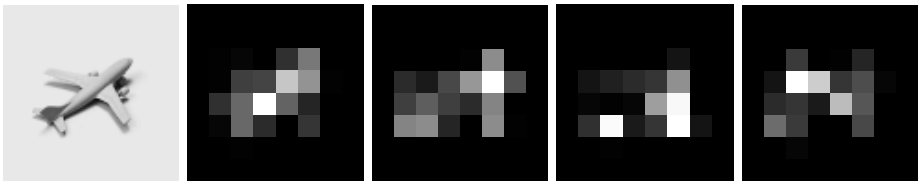
The local articulated hand pose is represented using gradient orientation histograms, frequently used for representation of human shape [1, 2]. Gradient orientation $\boldsymbol{\Phi}_t \in [0, \pi)$ is computed from the segmented hand image $\mathbf{I}_t^a$ (Figure 1b) as $\boldsymbol{\Phi}_t = \arctan(\frac{\partial \mathbf{I}_t^a}{\partial y} / \frac{\partial \mathbf{I}_t^a}{\partial x})$ where x denotes downward (vertical) direction and y rightward (horizontal) direction in the image.

From $\boldsymbol{\Phi}_t$, 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. Figures 1c–f show histograms at the lowest level of the pyramid.

The hand pose at time t is represented by the vector x_t^a which is the concatenation of the position p_t^a and all histograms at all levels in the pyramid. The length of x_t^a is thus $2 + B\sum_{l=1}^{L} 2^{2(L-l)}$. The performance of the classifier is quite insensitive to choices of $B \in [3, 8]$ and $L \in [2, 4]$; in our experiments in Section 5 we use $B = 4$ and $L = 3$. Before concatenation, p_t^a is normalized so that the standard deviations of the two dimensions of x_t^a originating from p_t^a have the same standard deviation in the training set (Section 5) as the remaining dimensions. The sequence of poses over the sequence is $\mathbf{x}^a = \{x_t^a\}, t = 1, \ldots, T$.

Object features. The objects considered in this application are all "graspable", i.e. more or less rigid. For example, a cup is graspable, but water is not. Shape can therefore be expected to be a good object class descriptor, while local descriptors like SIFT features [28] are unsuitable for our purposes.

Contrary to in many other object recognition applications, e.g. labeling of images according to object content, there is no search for object location involved. For manipulated objects, the position of the hand grasping the object gives an indication of the expected object location, and the recognition problem becomes one of classifying the given region. However, there might be deviations in position and orientation of the object within this region, as well as devia-



(a) Image (b) $\pi/8$ (c) $3\pi/8$ (d) $5\pi/8$ (e) $7\pi/8$

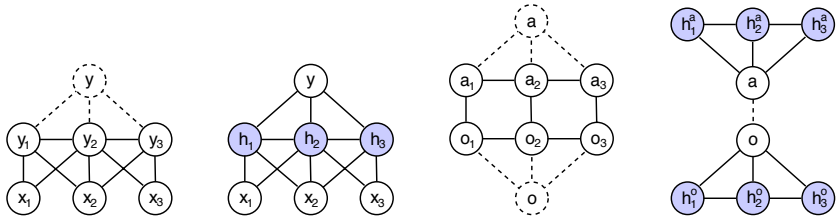
Fig. 2. Features used for object classification. a) Object image $\mathbf{I}_t^o$. b–e) Gradient orientation histograms from $\mathbf{I}_t^o$, with $B = 4$ bins, on level $l = 1$ in the pyramid of $L = 4$ levels. b) Bin 1, orientation $\pi/8$. c) Bin 2, orientation $3\pi/8$. d) Bin 3, orientation $5\pi/8$. e) Bin 4, orientation $7\pi/8$.

tions in size, shape and color of different instances within each object class. The descriptor must therefore be insensitive to these intra-class variations, but still capture inter-class variations.

We select the same gradient orientation histogram representation as for the hand shape, a description that captures the shape of the object, with a certain insensitivity to absolute greylevels and small displacements of object parts (Figure 2). Note that this representation is not invariant to e.g. in-plane rotations; this is deliberate, since global orientation is indicative of object class in our application. The object at time t is represented by x_t^o, the concatenation of all histograms at all levels in the pyramid. In Section 5 this representation is evaluated on the problem of recognizing generic object categories, and it is found to be robust to intra-class variability in shape, orientation, position, rotation and lighting conditions, while maintaining a good inter-class discriminability.

Another factor specific to this object recognition application is that the data consist of not only one, but a sequence of object views. In most cases, parts of the object are also occluded by the human hand grasping it. The change in orientation of the object with respect to the camera during the sequence and the occlusion from the hand are descriptive of the object, since they reflect the way this object class is used by the human; they can be termed "typical view sequences" and "typical occlusions". Thus, the classifier should take the whole sequence of object views into account. Each measurement is therefore described by a sequence of descriptors $\mathbf{x}^o = \{x_t^o\}, t = 1, \ldots, T$.

Correlation between action and object features. As discussed in the introduction, the shape of the hand encoded in $\mathbf{x}^a$ 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 view change in $\mathbf{x}^o$ over the course of the sequence is correlated with the type of action performed with the object. This representation of the correlation between manipulation actions and manipulated objects is *implicit* and difficult to model accurately, but should be taken into account when modeling the simultaneous action-object recognition.



(a) CRF (b) HCRF (c) FCRF (d) CHCRF

Fig. 3. Different CRF structures used for action-object recognition with pre-segmented data. Dotted edges and nodes indicate that the weights θ associated with these are constrained during the training of the CRF. a) Linear-chain CRF [4]. b) Hidden CRF [20]. c) Factorial CRF [5], data layer not shown. d) Connected hierarchic CRF:s, data layer not shown.

4 Discriminative Classification using CRF:s

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) [29]. We thus employ conditional random fields (CRF:s) [4] 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 $\mathbf{x}$. 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 using belief propagation [4].

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 Figure 3a. 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 $\{x_{t-\mathcal{C}}, \ldots, x_{t+\mathcal{C}}\}$, $\mathcal{C}$ being the *connectivity* of the model.

A CRF returns an individual label for each time-step, which means that the time-sequential data $\mathbf{x}$ to be classified do not have to be segmented prior to classification. However, if a segmentation is readily available, as in our case, the robustness of the classification is increased by assuming that for all labels within the segment, $y_t = y \,\forall\, t \in [1, T]$. With pre-segmentation, the model thus becomes $P(y|\mathbf{x};\theta) \equiv P(\mathbf{y} = [y, \ldots, y]|\mathbf{x};\theta)$. This is illustrated by the dotted layer in Figure 3a. We will formalize this below.

The introduction of a hidden layer, where each hidden label represents the classification of a sub-part of the sequence (Figure 3b), has been shown [20] to improve the recognition rate in situations where there is such a sub-structure present in the data. This is indeed the case in our object and action recognition problems. With an observable label y and a set of hidden labels $\mathbf{h}$ that form a time-chain, the probabilistic model of a hidden CRF (HCRF) becomes

$$P(y|\mathbf{x};\theta) = \frac{1}{Z(\mathbf{x})} \sum_{\mathbf{h}} \prod_{t=1}^{T} \Phi(y, h_t, \mathbf{x}; \theta_t^h) \prod_{t=2}^{T} \Phi(y, h_{t-1}, h_t, \mathbf{x}; \theta_t^{hh}) \tag{5}$$

with potential functions

$$\Phi(y, h_t, \mathbf{x}; \theta_t^h) = e^{\sum_k \theta_{t,k}^h f_k(y, h_t, \mathbf{x})},\ \Phi(y, h_{t-1}, h_t, \mathbf{x}; \theta_t^{hh}) = e^{\sum_k \theta_{t,k}^{hh} f_k(y, h_{t-1}, h_t, \mathbf{x})}. \tag{6}$$

Both inference and parameter estimation can be done using exact methods, provided that there are no loops in the hidden layer [20]. However, the introduction of hidden parameters leads to a non-convex optimization problem, which means that the parameter estimation procedure requires more data to converge, and might also reach local optima. Note also that an HCRF requires pre-segmented data; for continuous classification, other structures with hidden layers have been presented, like the latent-dynamic CRF (LDCRF) [30]. CRF and HCRF performance for pre-segmented data is compared in the experiments in Section 5.

The CRF in Figure 3a is in fact a special case of the HCRF in Figure 3b, where the weights θ are restricted so that *A*) $\mathbf{h}$ are not hidden, $h_t = y \,\forall\, t \in [1, T]$; and *B*) equal weight is given to each timestep, $\theta_{t_1}^h = \theta_{t_2}^h \,\forall\, t_1, t_2 \in [1, T]$.

"Early fusion": factorial CRF. In Section 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 in two ways, which we denote "early" and "late fusion". Early fusion corresponds to modeling the correlation on a per-frame basis, i.e. the correlations between the labels a_t and o_t for each frame of the action, using a factorial CRF (FCRF) [5]. Figure 3c shows an FCRF with two states,

action class a_t and object class o_t, in each time-step t. The conditional dependence on data is omitted in the figure for visibility. 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

$$P(\mathbf{a}, \mathbf{o}|\mathbf{x}; \theta) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \Phi(a_t, o_t, \mathbf{x}; \theta_t) \prod_{t=2}^{T} \Phi(a_{t-1}, a_t, \mathbf{x}; \theta_{a,t}) \Phi(o_{t-1}, o_t, \mathbf{x}; \theta_{o,t}) \ . \tag{7}$$

The weights θ are obtained during training using loopy belief propagation [5]. As for the linear-chain CRF, pre-segmentation means that $a_t = a, o_t = o \forall\, t \in [1, T]$ and that the distribution over the two dotted layers in the FCRF can be expressed as $P(a, o|\mathbf{x}; \theta) \equiv P(\mathbf{a} = [a, \ldots, a], \mathbf{o} = [o, \ldots, o]|\mathbf{x}; \theta)$.

"Late fusion": connected hierarchic CRF:s. Late fusion corresponds to modeling the correlation on a sequence level. The assumption is here that it is the action label a that is correlated with the object label o, not the labels a_t and o_t of a particular frame. The structure of a CRF for "late fusion", called *connected hierarchic CRF:s (CHCRF)* is shown in Figure 3d. In the most general case the two linear-chain layers are hidden, and the probability of an action a and an object o conditioned on the data $\mathbf{x}$ is

$$\begin{aligned} P(a, o|\mathbf{x}; \theta) = \frac{1}{Z(\mathbf{x})} \Phi(a, o, \mathbf{x}; \theta^{ao}) & \sum_{\mathbf{h}^a} \prod_{t=1}^{T} \Phi(a, h_t^a, \mathbf{x}; \theta_t^a) \prod_{t=2}^{T} \Phi(a, h_{t-1}^a, h_t^a, \mathbf{x}; \theta_t^{aa}) \\ & \sum_{\mathbf{h}^o} \prod_{t=1}^{T} \Phi(o, h_t^o, \mathbf{x}; \theta_t^o) \prod_{t=2}^{T} \Phi(a, h_{t-1}^o, h_t^o, \mathbf{x}; \theta_t^{oo}) \\ = \frac{\Phi(a, o, \mathbf{x}; \theta^{ao})}{\sum_{a,o} \Phi(a, o, \mathbf{x}; \theta^{ao})} & P(a|\mathbf{x}; \theta^a) P(o|\mathbf{x}; \theta^o) \end{aligned} \tag{8}$$

in analog with Eq (5). To make the training more efficient, the data dependency in the first term is omitted, becoming $\mathcal{K}(a, o) = \frac{\Phi(a,o;\theta^{ao})}{\sum_{a,o} \Phi(a,o;\theta^{ao})}$, the *co-occurrence rate* of action label a and object label o in the training data. The individual probabilities $P(a|\mathbf{x}; \theta^a)$ and $P(o|\mathbf{x}; \theta^o)$ are estimated as in Eq (3), Eq (5), or using any classification method that returns probability estimates. The parameters of the action and object classifiers are learned separately.

5 Experiments

Evaluation of the object features. The object feature representation was first evaluated on its own, without the CRF framework. For this we used the NORB dataset [3], which contains 5 different classes of rigid objects; "animals", "humans", "airplanes", "trucks", and "cars" with 10 instances of each, 5 for test and 5 for training.The database contains stereo views of each object from 18 different

azimuths and 9 elevations in 6 different lighting conditions. Only the normalized-uniform part of the dataset, designed to test classification performance, was used. (The other part of the dataset is designed to test object detection, a task we do not claim to address with this method.)

To evaluate the suitability of the feature representation for modeling shape *categories*, a support vector machine (SVM) [31] was trained with feature vectors x^o extracted from the NORB training images as described in Section 3. Table 1 left shows the results compared to others. Our object view representation together with a standard classifier reached the same classification accuracy as a state-of-the-art method for object categorization [3], which indicates that the gradient orientation histograms capture the specifics of a shape class, while allowing a significant variability among instances of that class. In comparison, training on the raw image downsampled to a size of 32×32 led to twice the classification error (a surprisingly good result, as noted in [3], given that the task is object categorization, 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 segmentation mask, is also desireable. In [3], 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.

Simultaneous action and object recognition. For the purpose of testing the simultaneous action-object recognition, the OAC (Object-Action-Complex) dataset was collected. The dataset consists of 50 instances each of three different action-object combinations; "look through binoculars", "drink from cup", and "pour from pitcher". The actions were performed by 7 men and 3 women, 5 times each. The classes were selected so that the object and action data are complementary:

Table 1. Results on the normalized-uniform NORB dataset, percent error. Left: Classification error percentage compared to methods presented in [3]. Right: Generalization test; robustness to different amount of jitter in test data (training data unaltered).

	Mono	**Stereo**
Hist + SVM	6.4	6.2
Raw + SVM	12.6 **[3]**	—
Conv Net 80	—	6.6 **[3]**

		Brightness			**Shift**		
	±0	**±10**	**±20**	**±30**	**±3**	**±6**	**±9**
Hist + SVM	6.4	6.4	7.1	8	10.3	18.1	29.2
Raw + SVM	12.6 **[3]**	15.8	18	21	20.8	35.1	48.6

(a) Instance of "look through binoculars"

(b) Instance of "drink from cup"

(c) Instance of "pour from pitcher"

Fig. 4. The three classes of the OAC dataset (for one person, instance each). Training and testing was performed in a jackknife manner, where the 15 sequences of one person at a time was used as a test data, the 135 sequences of the other 9 persons as training set.

two of the actions, "look through" and "drink from" are similar, while "cup" and "pitcher" are similar.

Only one instance of each object was used, so the full object representation generated a perfect classification performance with all parameter settings. To simulate the performance in the more general object category recognition case, all spatial information was removed, by using only $L = 1$ level in the gradient orientation histogram pyramid, and by normalizing the object segmentation window with respect to aspect ratio (i.e. scale all object segments to squares). The experiments below are not indicative of the object classification, but rather of the action classification and the benefits of combining object and action classification for manipulation action applications.

Table 2. Experiments with separate action and object (H)CRF classification with different connectivity $\mathcal{C}$, percent error on the format "median (max)" of 9 runs.

	CRF conn 0	**CRF conn 1**	**CRF conn 2**	**HCRF conn 0**	**Baseline**
Actions	5.3 (9.3)	6 (12.7)	10.7 (18.7)	14.7 (21.3)	17.3 (20) **(2D)**
Objects	8 (8.7)	11.3 (13.3)	20.7 (28)	24.7 (28.7)	8.7 **(1:st fr)**

Table 2 shows separate action and object classification with different parameter settings. The baseline for actions is a representation without spatial information – only 2D pos $\mathbf{p}^a$ over time, and for objects a representation without temporal information – only the first frame x_1^o. Two things can be noted. Firstly, while the object classification does well without the temporal information, the action is to a large degree determined by spatial information, i.e. the shape of the hand. Secondly, when comparing to similar experiments in [20] where HCRF:s outperformed CRF:s, we draw the conclusion that it depends on the amount of training data. With relatively little data, a model with fewer parameters will perform better. In the applications we are considering, the less training data needed, the better, since data collection takes time and the system should be adaptable to different environments.

When comparing late and early fusion (Table 3) we see that while late fusion (CHCRF) greatly improves the classification, there is only a marginal improvement in the classification with early fusion (FCRF). There are two possible interpretations which might both be true: *A*) the per-frame correlation of actions and objects is simply a bad model of reality; *B*) the larger set of parameters defining the FCRF leads to a worse performance with our relatively small training set. As a side note, the CHCRF object and action classification rates are identical since the co-occurrence matrix $\mathcal{K}$ (Eq (8)) is diagonal in this example. A non-unique action-object mapping (e.g. "drink"–"glass" and "drink"–"cup") would lead to differences in action and object classification rate.

How important is the implicit, data-level correlation compared to the explicit, semantic correlation modeled in the FCRF and CHCRF? To test this, late fusion was performed with full data, $\mathbf{x} = [\mathbf{x}^a, \mathbf{x}^o]$, correlation in either cue removed, $\mathbf{x} = [\mathbf{x}^a, x_1^o]$ or $[\mathbf{p}^a, \mathbf{x}^o]$, and all correlation removed, $\mathbf{x} = [\mathbf{p}^a, x_1^o]$. From Table 4, it seems that the information on objects in the action data $\mathbf{x}^a$ and the information on actions in the object data $\mathbf{x}^o$ is largely redundant; the classification performance is not affected to any greater extent by using *either* $\mathbf{p}^a$ or x_1^o, but if *all* correlation data are removed, the classification is seriously affected.

6 Conclusions

A method for simultaneous sequential recognition of manipulation actions and manipulated objects was presented, employing CRF:s trained with object and hand shape over the course of the action. Two different CRF structures for fusion of action and object classification were compared; FCRF:s, which model the correlation on a per-frame level, and CHCRF:s, a structure introduced in this

Table 3. Experiments comparing late (CHCRF) and early (FCRF) fusion, percent error on the format "median (max)" of 9 runs. Connectivity 0 everywhere.

	Connected Hiearchic CRF:s	**Factorial CRF**	**Baseline (sep CRF:s)**
Actions	3.3 (4.7)	5.3 (8)	5.3 (9.3)
Objects	3.3 (4.7)	6 (12)	8 (8.7)

paper, which correlate the action and object classifcation as whole. CHCRF:s outperformed FCRF:s on the task of correlated classification of action and object sequences, probably because the action and object data are not correlated on a per-frame basis.

CRF structures with many degrees of freedom, such as HCRF:s, or CRF:s with high data connectivity, were found to perform worse than simpler structures with relatively small amounts of training data, although they have higher descriptive power. Thus, in applications where the training data are limited, the complexity of the model should be selected so that the training procedure will converge with the amount of training data at hand. Moreover, the implicit information on actions in the object data and on objects in the action data was found to be redundant. Thus, removing the correlated data in one cue or the other can be done without affecting the overall classification, while removing the correlated data in both cues will have serious effects on the classification rate.

The representation of object shape using a pyramid of gradient orientation histograms was shown to give state-of-the-art classification results on the NORB dataset, indicating that the representation is robust to intra-class differences in quite general shape categories, while capturing inter-class differences.

In the future, we plan to incorporate the method for correlated action and object classification into a PbD framework such as [24, 25].

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

References

1. Freeman, W.T., Roth, M.: Orientational histograms for hand gesture recognition. In: IEEE Int. Conf. Automatic Face and Gesture Recognition. (1995)
2. Shakhnarovich, G., Viola, P., Darrell, T.: Fast pose estimation with parameter sensitive hashing. In: ICCV. Volume 2. (2003) 750–757
3. LeCun, Y., Huang, F.J., Bottou, L.: Learning methods for generic object recognition with invariance to pose and lighting. In: CVPR. Volume 2. (2004) 97–104
4. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML. (2001)
5. Sutton, C., Rohanimanesh, K., McCallum, A.: Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. In: ICML. (2004)

Table 4. How important is the implicit, data-level correlation to the overall action-object classification? Classification where none, either, or both cues are stripped baseline versions, percent error on the format "median (max)" of 9 runs. Late (CHCRF) fusion, connectivity 0 everywhere.

	Actions	**Actions Baseline (only 2D pos)**
Objects	3.3 (4.7)	2.7 (8)
Objects Baseline (SVM 1:st fr)	4.7 (8.7)	14.7 (18)

6. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in computer vision-based human motion capture and analysis. CVIU **104** (2006) 90–126
7. Laptev, I., Pérez, P.: Retrieving actions in movies. In: ICCV. (2007)
8. Pavlovic, V.I., Sharma, R., Huang, T.S.: Visual interpretation of hand gestures for human-computer interaction: A review. PAMI **19** (1997) 677–695
9. Yang, M., Ahuja, N., Tabb, M.: Extraction of 2d motion trajectories and its application to hand gesture recognition. PAMI **24** (2002) 1061–1074
10. Stenger, B., Thayananthan, A., Torr, P.H.S., Cipolla, R.: Model-based hand tracking using a hierarchical bayesian filter. PAMI **28** (2006) 1372–1384
11. Sudderth, E., Mandel, M.I., Freeman, W.T., Willsky, A.S.: Visual hand tracking using non-parametric belief propagation. In: IEEE Workshop on Generative Model Based Vision. (2004)
12. Sminchisescu, C., Kanujia, A., Metaxas, D.: Conditional models for contextual human motion recognition. CVIU **104** (2006) 210–220
13. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. PAMI **24** (2002) 509–522
14. Sullivan, J., Carlsson, S.: Recognizing and tracking human action. In: ECCV. Volume 1. (2002) 629–644
15. Carbonetto, P., de Freitas, N., Barnard, K.: A statistical model for general contextual object recognition. In: ECCV. Volume 1. (2004) 350–362
16. Quattoni, A., Collins, M., Darrell, T.: Learning visual representations using images with captions. In: CVPR. (2007)
17. Torralba, A.: Contextual priming for object detection. IJCV **53** (2003) 169–191
18. Murphy, K., Torralba, A., Freeman, W.T.: Using the forest to see the trees: A graphical model relating features, objects, and scenes. In: NIPS. (2003)
19. Torralba, A., Murphy, K., Freeman, W.T.: Contextual models for object detection using boosted random fields. In: NIPS. (2004)
20. Quattoni, A., Wang, S., Morency, L., Collins, M., Darrell, T.: Hidden conditional random fields. PAMI **29** (2007) 1848–1852
21. Rivlin, E., Dickinson, S.J., Rosenfeld, A.: Recognition by functional parts. CVIU **62** (1995) 164–176
22. Stark, L., Bowyer, K.: Generic Object Recognition using Form and Function. World Sci. Ser. Machine Perception and Artificial Intelligence; Vol. 10 (1996)
23. Jurie, F., Schmid, C.: Scale-invariant shape features for recognition of object categories. In: CVPR. Volume 2. (2004) 90–96
24. Moore, D.J., Essa, I.A., Hayes, M.H.: Exploiting human actions and object context for recognition tasks. In: ICCV. (1999)
25. Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., Rehg, J.M.: A scalable approach to activity recognition based on object use. In: ICCV. (2007)
26. Mann, R., Jepson, A.: Towards the computational perception of action. In: CVPR. (1998)
27. Argyros, A.A., Lourakis, M.I.A.: Real time tracking of multiple skin-colored objects with a possibly moving camera. In: ECCV. Volume 3. (2004) 368–379
28. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
29. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77** (1989) 257–286
30. Morency, L.P., Quattoni, A., Darrell, T.: Latent-dynamic discriminative models for continuous gesture recognition. In: CVPR. (2007)
31. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001) Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.