Grasp Type Revisited: A Modern Perspective on A Classical Feature for Vision

Yezhou Yang[1], Cornelia Fermüller[1], Yi Li[2], and Yiannis Aloimonos[1]

[1]Computer Vision Lab, University of Maryland, College Park [2]NICTA and ANU

Abstract

The grasp type provides crucial information about human action. However, recognizing the grasp type from unconstrained scenes is challenging because of the large variations in appearance, occlusions and geometric distortions. In this paper, first we present a convolutional neural network to classify functional hand grasp types. Experiments on a public static scene hand data set validate good performance of the presented method. Then we present two applications utilizing grasp type classification: (a) inference of human action intention and (b) fine level manipulation action segmentation. Experiments on both tasks demonstrate the usefulness of grasp type as a cognitive feature for computer vision. This study shows that the grasp type is a powerful symbolic representation for action understanding, and thus opens new avenues for future research.

1. Introduction

The grasp type contains fine-grain information about human action. Consider the two scenes in Fig. 1 from the VOC challenge. Current computer vision systems can easily detect that there is one bicycle and one cyclist (human being) in the image. Through human pose estimation, the system can further confirm that these two cyclists are riding the bike. But humans can tell that the cyclist on the left side literally is not "riding" the bicycle since his hands are posing in a "Rest or Extension" grasp next to the handlebar while the cyclist on the right side is racing because his hands firmly hold the handlebar with a "Power Cylindrical" grasp. In other words, the recognition of grasp type is essential for a more detailed analysis of human action, beyond the processes of current state-of-the-art vision systems.

Moreover, recognizing grasp type can help an intelligent system predict the human action intention. Consider an intelligent agent looking at the two scenes in Fig. 2(a) and (b). Current state-of-the-art computer vision techniques can accurately recognize many visual aspects from both of these scenes, such as the fact that there must be a human being standing in the outdoor garden scene, with a knife in his/her hand. However, we human beings will react dramatically different when experiencing the two different scenes, because of our ability to recognize immediately the different ways the person is handling the knife, i.e., the grasp type. We can effectively infer the possible activity the man is going to do based on his way of grasping the knife. After seeing scene Fig. 2(a), we could believe this man is going to cut something hard, or even might be malicious, since he is "Power Hook" grasping the knife. After seeing scene Fig. 2(b), we may react with a movement to acquire the knife (shown in Fig. 2(c)) since the man is "Precision Lumbrical" grasping the knife indicating a passing action. From this example we can see that the grasp type is a strong cue for us to infer the human action intention.

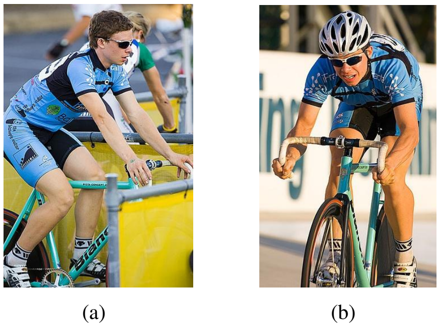
(a) (b)

Figure 1. (a) Rest or Extension on the handlebar vs. (b) Firmly power cylindrical grasping the handlebar.

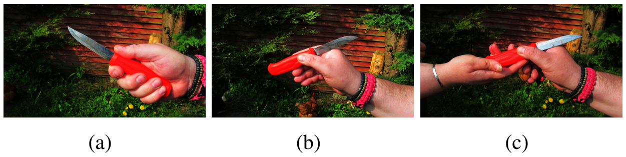
(a) (b) (c)

Figure 2. (a) Power Hook Grasp a knife vs. (b) Precision Lumbrical Grasp a knife. (c) A natural reaction when seeing scene (b) is to open the hand to receive the knife.

These are two examples demonstrating how important it is for us to be able to recognize grasp types. The grasp type is an essential component in the characterization of human actions of manipulation ([26]). From the viewpoint of processing videos, the grasp contains information about the action itself, and it can be used for prediction or as a feature for recognition. It also contains information about the beginning and end of action segments, and thus it can be used to segment videos in time. If we are to perform the action with an intelligent agent, such as a humanoid robot, knowledge about how to grasp the object is necessary so the robot can arrange its effectors. For example, consider a humanoid with one parallel gripper and one vacuum gripper. When a power grasp is desired, the robot should select the vacuum gripper for a stable grasp, but when a precision grasp is desired, the parallel gripper is a better choice. Thus, knowing the grasp type provides information for the robot to plan the configuration of its effectors, or even the type of effector to use ([27]).

Here we present a study centered around human grasp type recognition and its applications in computer vision. The goal of this research is to provide intelligent systems with the capability to recognize the human grasp type from unconstrained static or dynamic scenes. To be specific, our system takes in an unconstrained image patch around the human hand, and outputs which category of grasp type is used (examples are shown in Fig. 3). In the rest of the paper, we show that this capability 1) is very useful for predicting human action intention and 2) helps to further understand human action by introducing a finer layer of granularity. Further experiments on two publicly available dataset empirically support that we can 1) infer human action intention in static scenes and 2) segment videos of human manipulation actions into finer segments based on the grasp type evolution. Additionally, we provide a labeled grasp type image data set and a human intention data set for further research.

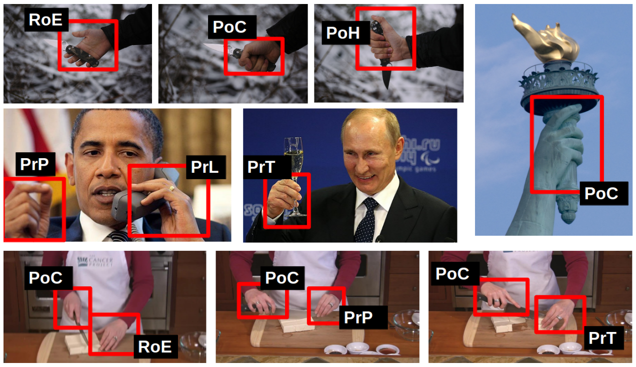

Figure 3. Sample outputs. PoC: Power Cylindrical; PoS: Power Spherical; PoH: Power Hook; PrP: Precision Pinch; PrT: Precision Tripod; PrL: Precision Lumbrical; RoE: Rest or Extension

2. Related Work

Human hand related: One way to recognize grasp type is through model based hand detection and tracking [17]. Based on the estimated articulated hand model, a set of biologically plausible features such as the arches formed by fingers [22] were used to infer the grasp type involved [26]. These approaches normally use RGB Depth data and require a calibration phase, which is not applicable or is too fragile for real world situations. Also a lot of research has been devoted to hand pose or gesture recognition with promising experimental results [15, 28]. The goal of these works is to recognize poses such as “POINT”, “STOP” or “YES” and “NO”, not considering the interaction with objects. When it comes to recognizing grasp type from unconstrained visual input, inevitably our system has to deal with the additional challenges introduced by the interaction with unknown objects. Later in the paper we will show that the large variation in the scenery will not allow traditional feature extraction and learning mechanism to work robustly on public available hand patch testing beds.

The robotics community has been studying perception and control problems of grasping for decades [20]. Recently, several learning based systems were reported that infer contact points or how to grasp an object from its appearance [19, 12]. However, the desired grasping type could be different for the same target object, when used for different action goals. The acquisition of grasp information from natural static or dynamic scenes is still considered very difficult because of the large variation in appearance and the occlusions of the hand from objects during manipulation.

Vision beyond appearance: The very small number of works in computer vision, which aim to reason beyond appearance models, are also related to this paper. [24] proposed that beyond state-of-the-art computer vision techniques, we could possibly infer implicit information (such as functional objects) from video, and they call them “Dark Matter” and “Dark Energy”. [25] used stochastic tracking and graph-cut based segmentation to infer manipulation consequences beyond appearance. [9] used a ranking SVM to predict the persuasive motivation (or the intention) of the photographer who captured an image. More recently, [18] seeks to infer the motivation of the person in the image by mining knowledge stored in a large corpus using natural language processing techniques. Different from these fairly general investigations about reasoning beyond appearance, our paper seeks to infer human action intention from a unique and specific point of view: the grasp type.

Convolutional neural networks: The recent development of deep neural networks based approaches revolutionized visual recognition research. Different from the traditional hand-crafted features [13, 4], a multi-layer neural network architecture efficiently captures sophisticated hierarchies describing the raw data [1], which has shown superior performance on standard object recognition benchmarks [10, 2] while utilizing minimal domain knowledge.

The work presented in this paper shows that with the recent developments of deep neural networks, we can learn a model to recognize grasp type from unconstrained visual inputs with robustness. We believe we are among the first to apply deep learning on grasp type recognition.

3. Our Approach

First, we briefly summarize the basic concepts of Convolutional Neural Networks (CNN), and then we present our implementations for grasp type recognition, human action intention prediction and fine level manipulation action segmentation using the change of grasp type over time.

3.1. Human Grasp Types

A number of grasping taxonomies have been proposed in several areas of research, including robotics, developmental medicine, and biomechanics, each focusing on different aspects of action. In a recent survey, Feix et al. [6] reported 45 grasp types in the literature, of which only 33 were found valid. In this work, we use a categorization into seven grasp types. First we distinguish, according to the most commonly used classification (based on functionality), into power and precision grasps [7]. Power grasping is used when the object needs to be held firmly in order to apply force, such as "grasping a knife to cut"; precision grasping is used in order to do fine grain actions that require accuracy, such as "pinch a needle". We then further distinguish among the power grasps, whether they are cylindrical, spherical, or hook. Similarly, we distinguish the precision grasps into pinch, tripodal and lumbrical. Additionally, we also consider a Rest or Extension position (no grasping performed). Fig. 4 illustrates the grasp categories.

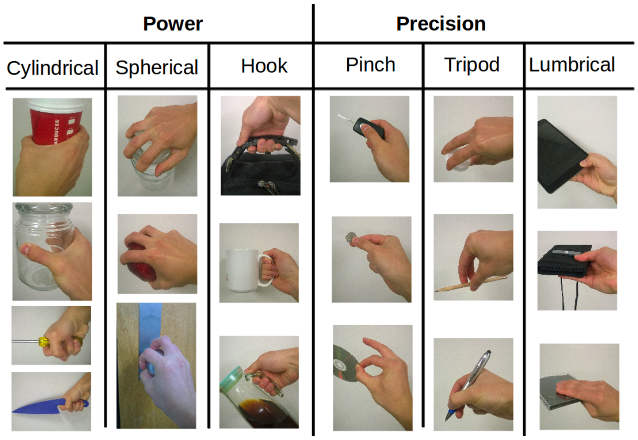

Figure 4. The grasp types considered. Grasps which can not be categorized into the six types here are considered as the "Rest and Extension" (no grasping performed).

Humans, when looking at a photograph, can more or less tell what kind of grasp the person in the picture is using. The question becomes, whether using the current state-of-the-art computer vision technique, whether we can develop a system that learns the pattern from human labeled data and recognizes grasp type from a patch around each hand? In the following section, we present our take and show that a grasp type recognition model with decent robustness can be learned using Convolutional Neural Network (CNN) techniques.

3.2. CNN for Grasp Type Recognition

Convolutional Neural Network (CNN) is a multilayer learning framework, which may consist of an input layer, a few convolutional layers and an output layer. The goal of CNN is to learn a hierarchy of feature representations. Response maps in each layer are convolved with a number of filters and further down-sampled by pooling operations. These pooling operations aggregate values in a smaller region by down-sampling functions including max, min, and average sampling. In this work we adopt the softmax loss function which is given by:

$$L(t, y) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{C} t_k^n log(\frac{e^{y_k^n}}{\sum_{m=1}^{C} e^{y_m^n}}) \qquad (1)$$

where t_k^n is the n-th training example's k-th ground truth output, and y_k^n is the value of the k-th output layer unit in response to the n-th input training sample. N is the number of training samples, and since we consider 7 grasp type categories, $C = 7$. The learning in CNN is based on Stochastic Gradient Descent (SGD), which includes two main operations: Forward and Back Propagation. The learning rate is dynamically lowered as training progresses. Please refer to [11] for details.

We used a five layer CNN (including the input layer and one fully-connected perception layer for regression output). The first convolutional layer has 32 filters of size 5×5 with max pooling, the second convolutional layer has 32 filters of size 5×5 with average pooling, and the third convolutional layer has 64 filters of size 5×5 with average pooling, respectively. Convolutional layer convolves its input with a bank of filters, then applies point-wise non-linearity and max or average pooling operation.

The final fully-connected perception layer has 7 regression outputs. Fully-connected perception layer applies linear filters to its input, then applies point-wise non-linearity. Our system considers 7 grasp type classes.

For testing, we pass each target hand patch to the trained CNN model, and obtain an output of size 7×1: $P_{GraspType}$. In the action intention and segmentation experiments we use the classification for both hands to obtain $P_{GraspType1}$ for the left hand, and $P_{GraspType2}$ for the right hand, respectively. To have a fully automatic fine level manipulation segmentation approach, we need to localize the input hand patches from videos and then recognize grasp types using CNN. We use the hand detection method of [14] to detect hands in the first frame, and then apply a meanshift algorithm based tracking method [3] on both hands to continuously extract the image patch around each hand.

3.3. Human Action Intention

Our ability to interpret other people's actions hinges crucially on predicting their intentionality. Even 18-month-old infants behave altruistically when they observe an adult accidentally dropping a marker on the floor but out of his reach, and they can predict his intention to pick up the marker [23]. From the point view of machine learning for intelligent systems and human-robot collaboration, due to the differences in the embodiment of humans and robots, a direct mapping of action signals is problematic. One solution is that the robot predicts the intent of the observed human activity and implements the same intention using its own sensorimotor apparatus [21].

Previous studies showed that there are several key factors that affect the grasp type [16]. One crucial deciding factor for the selection of the grasp type to use is the intended activity. We choose here a categorization into three human action intentions, closely related to the functional classification discussed above (Fig. 4). The first category reflects the intention to apply force onto the physical world, such as for example "cut down a tree with an ax", and we refer to it as "Force-oriented". The second category reflects fine-grained activity where sensitivity and dexterity are needed, such as "tie shoelaces", and we refer to it as "Skill-oriented". The third category has no intention of specific action, such as "showcasing and posing", and we call it "Casual". Fig. 5 illustrates the action intention categories by showing one typical example of each. We should note that the three categories: "force-oriented", "skill-oriented" and "casual" are closely related to the three functional categories "power" "precision", and "rest", respectively (Fig. 4). We used a different labeling, because we encounter a larger variety of hand poses in the static images used for intention classification than in the videos of human manipulation activities used for functional categorization.

We investigate the causal relation between human grasp type and action intention by training a classifier using grasp types of both hands as input, and the category of action intention as output. As shown next, our experiment demonstrates a strong link. We want to point out that certainly a finer categorization is possible. For example, "Force oriented" intention can be further divided into sub classes such as "Selfish" or "Altruistic" and so on. However, such a classification would require other dynamic observations. Here we show that from the grasp type in a single image a classification into basic intentions (shown in Fig. 5) is possible.

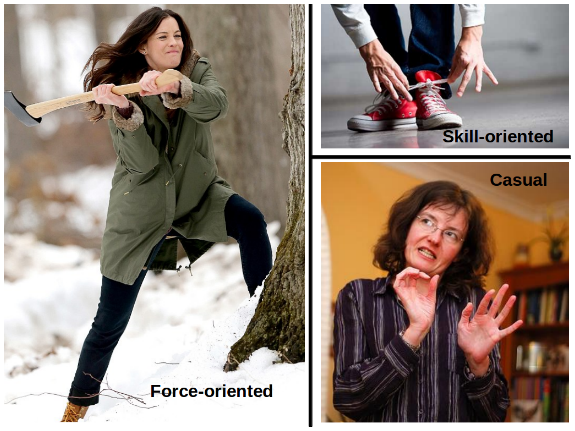

Figure 5. Human action intention categories.

3.4. From Grasp Type to Action Intention

Our hypothesis is that the grasp type is a strong indicator of human action intention. In order to validate this, we train an additional classifier layer. The procedure is as follows. For each training image, we first pass the target hand patches (left hand and right hand, if present) of the main character in the image to the trained CNN model, and we obtain two belief distributions: $P_{GraspType1}$ and $P_{GraspType2}$. We concatenate these two distributions and use them as our feature vector for training. We train a support vector machine (SVM) classifier f, which takes as input the grasp type belief distributions and derives as output an action intention distribution P_{Int} of size 3×1:

$$P_{Int} = f(P_{GraspType1}, P_{GraspType2}|\theta), \tag{2}$$

where θ are the model parameters learned from labeled pairs. Fig. 6 shows a diagram of the approach. We need to point out that in the human action intention recognition we use belief distributions instead of final class labels of the two hands as input feature vectors. Thus, a certain category of grasp type does not directly indicate a certain action intention in our model. A further experiment using detected grasp type labels of both hands (the grasp type with the highest belief score) to infer action intention achieves a slightly worse performance, which confirms our claim here.

3.5. Grasp Type Evolution

In manipulation actions involving tools and objects, the details of the small sub actions contain rich semantics. Current computer vision methods do not consider them. Consider a typical kitchen action, as shown in Fig. 7. In most approaches the whole sequence would be denoted as "sprinkle the steak", and the whole segment would be considered an atomic part for recognition or analysis. However, within this around 15 second long action, there are several finer segments. The gentleman first "Pinch" grasps the salt to

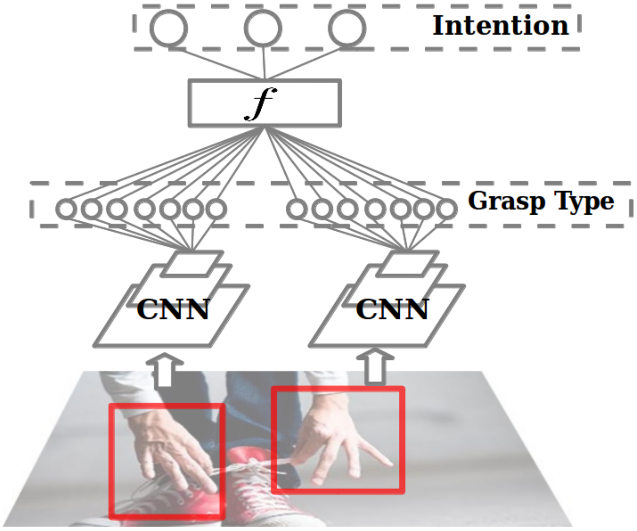

Figure 6. Inference of human action intention from grasp type recognition.

sprinkle the beef, then he "Extends" to point at the oil bottle, and later he "Power Spherical" grasps a pepper bottle to further sprinkle black pepper onto the beef. Here we can see that the dynamic changes of grasp type characterize the start and end of these finer actions.

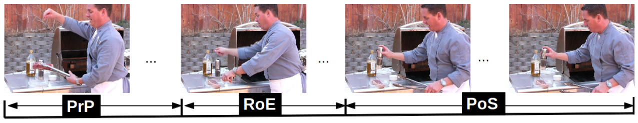

Figure 7. Grasp type evolution (right hand) in a manipulation action.

In order to see if grasp type evolution actually can help with a finer segmentation of manipulation actions, we first recognize the grasp type of both hands, frame by frame, and then output a segmentation at the points in time when any of the hands has a change in grasp type. We design a third experiment on a public cooking video dataset from Youtube for validation.

3.6. Finer segment action using grasp type evolution

We adopt a straightforward approach. Let's denote the sets of grasp types along the time-line of an action of length M as $G_l = \{G_l^1, G_l^2...G_l^M\}$ for the left hand and as $G_r = \{G_r^1, G_r^2...G_r^M\}$ for the right hand. Assuming that during a manipulation action the grasp type evolves gradually, we first apply a one dimensional mode filter to smooth temporally. Each grasp type detection at time t is replaced by its most common neighbor in the window of $[t-\delta/2, t+\delta/2]$, where δ is the window size.

Then, whenever at a time instance $t \in [1, M]$, if $G_l^t \neq G_l^{t+1}$ or $G_r^t \neq G_r^{t+1}$, our system outputs one segment at t, denoted as S_t. The set S_t yields a finer segmentation of the manipulation action clip.

4. Experiments

The theoretical framework we presented suggests three hypotheses that deserve empirical tests: (a) the CNN based grasp type recognition module can robustly classify input frame patches around hands into correct categories; (b) hand grasp type is a reliable cognitive feature to infer human action intention; (c) the evolution of hand grasp types is useful for fine-grain segmentation of human manipulation actions.

To test the three hypotheses empirically, we need to define a set of performance variables and how they relate to our predicted results. The first hypothesis relates to visual recognition, and we can empirically test it by comparing the detected grasp type labels with the ground truth ones using the precision and recall metrics. We further compare the method with a traditional hand-crafted feature based approaches to show the advantage of our approach. The second hypothesis relates to the inference of human action intention, and we can also empirically test it by comparing the predicted action intention with the ground truth ones on a testing set. The third hypothesis relates to manipulation action segmentation, and we can test it by comparing the computed key segment frames with the ground-truth ones. We used two publicly available datasets: (1) the Oxford hand dataset [14] and (2) a unconstrained cooking video dataset (YouCook) [5].

4.1. Grasp Type Recognition in Static Images

Dataset and Experimental protocol

The Oxford hand dataset is a comprehensive dataset of hand images collected from various different public image data set sources with a total of 13050 annotated hand instances. Hand patches larger than a fixed area of (a bounding box of 1500 sq. pixels) were considered sufficiently 'large' and were used for evaluation. This way we obtained 4672 hand patches from the training set and 660 hand patches from the testing set (VOC images). We then further augmented the dataset with new annotations. We categorized each patch into one of the seven classes by considering its functionality given the image context and its appearance following Fig. 4. We followed the training and testing protocol from the dataset.

For training the grasping type, the image patches were resized to 64×64 pixels. The training set contains 4672 image patches and was labeled with the seven grasping types. We used a GPU based CNN implementation [8] to train the neural network, following the structure described above (Sec. 3.2).

We compared our approach with traditional hand-crafted feature based approaches. One was the histogram of oriented gradients (HoG) + Bag of Words (BoW) + SVM clas-

Methods	PoC		PoS		PoH		PrP		PrT		PrL		RoE		Overall
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	Accu
HoG+BoW+SVM	.44	.46	0	NaN	0	NaN	0	0	0	NaN	0	NaN	.81	.41	.42
HoG+BoW+RF	.50	.40	0	NaN	0	NaN	.03	.17	0	0	0	0	.62	.36	.36
CNN	.59	.60	.38	.62	.38	.58	.62	.60	.56	.66	.36	.40	.69	.56	.59

Table 1. Precision (P) and Recall (R) for each grasp type category and overall accuracy.

sification, the other HoG + BoW + Random Forest. The number of orientations we selected for HoG was 32, and the number of dictionary entries for BoW was 100. The parameters for the baseline methods were tuned to have the best performance.

Experimental results

We achieved an average of 59% classification accuracy using the CNN based method. Table 1 shows the performance metrics of each grasp type category and the overall performance in comparison to baseline algorithms. It can be seen that the CNN based approach has a decent advantage. To provide a full picture of our CNN based classification model, we also show the confusion matrix in Fig. 8. Our system mainly confused “Power Cylindrical Grasp” with “Rest or Extension”. We believe that this is mostly because the fingers form natural curves when resting and this makes the hand look very similar to a cylindrical grasp with large diameter. Also our model does not perform well on “Precision Lumbrical” grasp due to the relatively small amount of training samples in this category. Fig. 9 shows some correct grasp type predictions shown by black boxes, and some failure examples denoted by red and blue bounding boxes. Blue boxes denote a correct prediction of the underlying high-level grasp type in either the “Power” or “Precision” category, but incorrect recognition in finer categories. Red box denotes a confusion between“Power” and “Precision” grasp. Intuitively, the blue marked errors should be penalized less than the red marked ones.

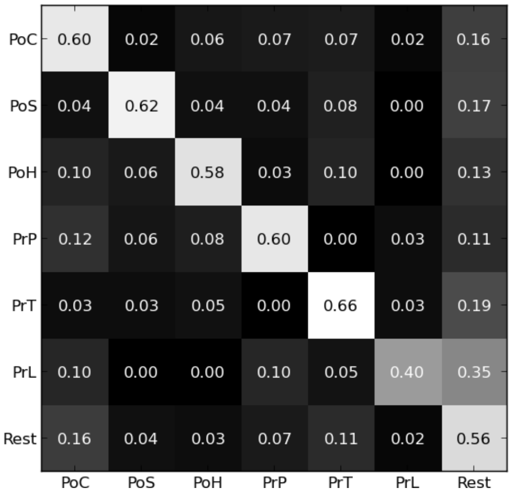

Figure 8. Category pairwise confusion matrix for grasp type classification using CNN.

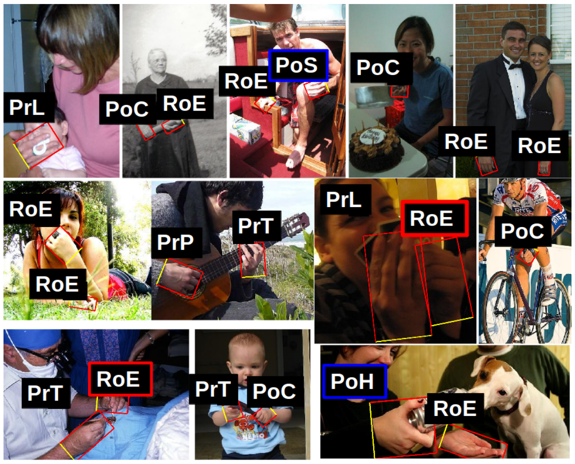

Figure 9. Examples of correct and false classification. PoC: Power Cylindrical; PoS: Power Spherical; PoH: Power Hook; PrP: Precision Pinch; PrT: Precision Tripod; PrL: Precision Lumbrical; RoE: Rest or Extension.

4.2. Inference of Action Intention from Grasp Type

Dataset and Experimental protocol

A subset of 200 images from the Oxford hand dataset serves as testing bed for action intention classification. Since not every image in the test set contains an action intention that falls into one of the three major categories described above, the subset was selected with the following rules: (1) at least one hand of the main character can be seen from the image and (2) the main character has a clear action intention. For example, we can infer that the character from Fig. 10(a) is going to perform a skill-oriented actions that requires accuracy, while this is not clear from the character in Fig. 10(b) (pull the rope with force or just posing casually?). We labeled the 200 images into the three major action intention categories and used them as ground truth. The grasp type CNN model was used to extract a 14 dimension belief distribution as grasp type feature (which is due to data from both hands of the main character). A 5 folds cross validation protocol was adopted and we trained each fold using a linear SVM classifier.

Experimental results

We achieved an average 65% prediction accuracy. Table 2 reports precision and recall metrics for each category of action intention. We also run the same experiment using grasp type labels instead of belief distributions (GL+SVM).

(a)

(b)

Figure 10. Clear action intention vs. an ambigous one

Methods	F-O		S-O		C		Overall
	P	R	P	R	P	R	Accu
GL+SVM	.54	.35	.73	.59	.80	.89	.63
GT+SVM	.61	.35	.82	.71	.82	.83	.65

Table 2. Precision (P) and Recall (R) for each intention category and overall accuracy. GL: Grasp type Label; GT: Grasp Type belief distribution.

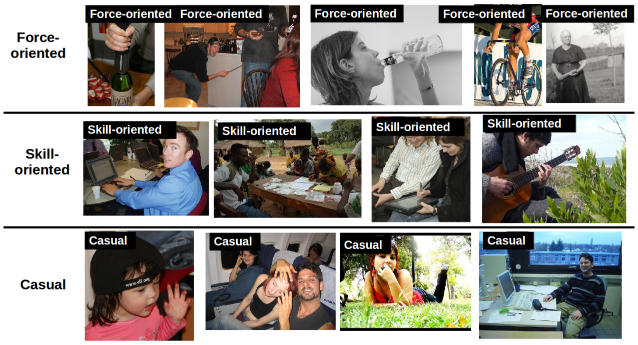

Figure 11. Correct examples of predicting action intention.

We can see that it achieves slightly worse performance than using belief distributions. Fig. 11 shows some interesting correct cases, and Fig. 12 shows several failure predictions. We believe that the failure cases are mostly due to the wrong grasp type recognition inherited from the previous section. Because of the small amount of pairs with ground truth, we were not able to train for comparison a converging CNN model, that would predict action intention directly from hand patches.

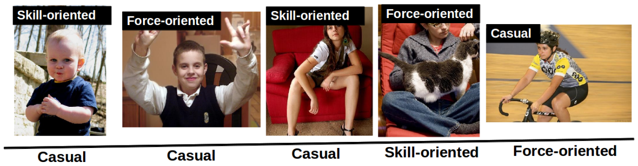

Figure 12. Failure cases of predicting action intention. The label at the bottom denotes the human labeling.

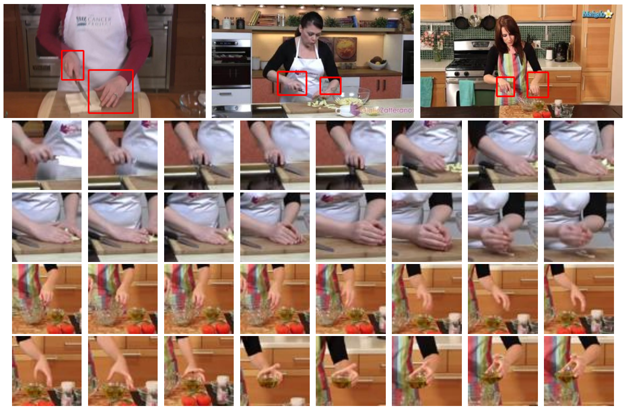

Figure 14. 1st row: sample hand localization on first frame using [14]. 2nd to 5th row: two sample sequences of hand patches extracted using meanshift tracking [3].

4.3. Manipulation Action Fine Level Segmentation using Grasp Type Evolution

In this section we want to demonstrate that the change of grasp type is a good feature for fine grain level manipulation action temporal segmentation.

Dataset and Experimental protocol

Cooking is an activity, requiring a variety of grasp types, that intelligent agents most likely need to learn. We conducted our experiments on a publicly available cooking video dataset collected from the world wide web and fully labeled, called the Youtube cooking dataset (YouCook) [5]. The data was prepared from 88 open-source Youtube cooking videos with unconstrained third-person view. These features make it a good empirical testing bed for our third hypothesis.

We conducted the experiment using the following protocols: (1) 8 video clips, which contain at least two fine grain activities, were reserved for testing; (2) all other video frames were used for training; (3) we randomly reserved 10% of the training data as validation set for training the CNNs. For training the grasp type recognition model, we extended the dataset by annotating image patches containing hands in the training frames. The image patches were resized to 64 × 64 pixels. The training set contains image patches that was labeled with the seven grasp types. We used the same GPU based CNN implementation [8] to train the neural network, following the same structures described above.

Action Fine Level Segmentation

For each testing clip, we first picked the top two hand proposals using [14] in the first frame, and then we applied a meanshift algorithm based tracking method [3] on both hands to continuously extract an image patch around each hand (Fig. 14). The image patches were further resized to 64 × 64 and pipelined to the trained CNN model. We then

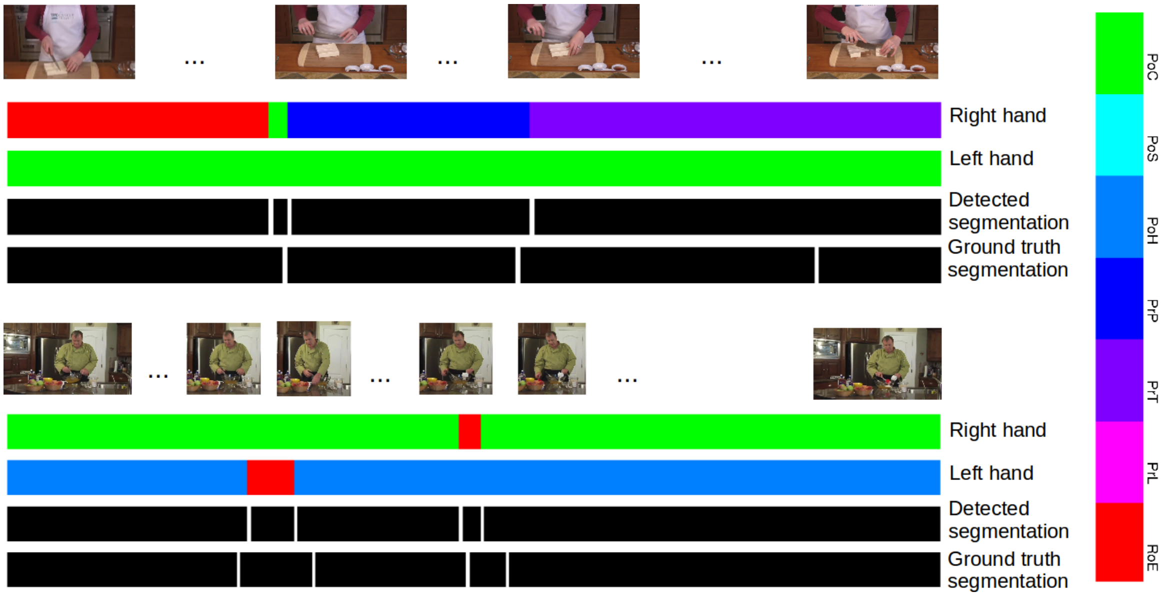

Figure 13. Left and right hand grasp type recognition along timeline and video segmentation results compared with ground truth segments.

labeled each hand with the grasp type of highest belief score in each frame. After applying a one dimensional mode filtering for temporal smoothing, we computed the grasp type evolution for each hand and segmented whenever one hand changes grasp type, as described in Sec. 3.6.

Fig. 13 shows two examples of intermediate grasp type recognition for the two hands and the detected segmentation. A key frame is considered correct, when a ground truth key frame lies within 10 frames around it. In the first example, the subject's right hand at the beginning holds the tofu using an Extension grasp, and then she cuts the tofu with a Pinch grasp holding the blade. Then using a precision Tripod grasp she separates one piece of tofu from the rest, and at the end using a Lumbrical grasp she further cuts the smaller piece of tofu. Using the grasp type evolution, our system can successfully detect two key frames out of the three ground truth ones. In the second video, the gentleman using a Cylindrical grasp whisks the bowl at the beginning. Then his left hand extends to reach a small cup, and then using a Hook grasp he holds the cup. After that, his right hand extends to reach a spatula and at the end his right hand scoops food out of the small cup using a Cylindrical grasp. Using the grasp type evolution, our system can successfully detect three key frames out of the four ground truth ones.

In the 8 test clips, there are 18 ground truth segmentation key frames, and 14 of them are successfully detected, which yields a recall of 78%. Among the 20 detected segmentation key frames, 16 are correct, which yields a precision of 80%.

5. Conclusion and Future Work

Our experiments produced three results: (i) we achieved in average 59% accuracy using the CNN based method for grasp type recognition from unconstrained image patches; (ii) we achieved in average 65% prediction accuracy in inferring human intention using the grasp type only; (iii) using the grasp type temporal evolution, we achieved 78% recall and 80% precision in fine grain manipulation action segmentation tasks. Overall, the empirical results support our hypotheses (a-c) respectively.

Recognizing grasp type to infer human action intention or to do fine level segmentation of human manipulation actions are novel problems in computer vision. We have proposed a CNN based learning framework to address these problems with decent success. We hope our contributions can help advance the field of static scene understanding and human action fine level analysis, and we hope that they can be useful to other researchers in other applications. Additionally, we augmented a currently available hand data set and a cooking data set with grasp type labels, and provided human action intention labels for a subset of them. We will make this augmented data sets available for future research.

Our experiments indicate that there is still significant space for improving the recognition of grasp type and inference of human intention. We believe that advances in understanding high-level cognitive structure underlying human intention can help improve the performance. With the development of deep learning systems and more data, we can also expect a robust grasp type recognition system beyond the seven categories used in this paper. Moreover, we believe that progress in natural language processing, such as mining the relationship between grasp type and actions, can advance high-level reasoning about human action intention to improve computer vision methods.

6. Acknowledgements

This research was funded in part by the support of the European Union under the Cognitive Systems program (project POETICON++), the National Science Foundation under INSPIRE grant SMA 1248056, and by DARPA through U.S. Army grant W911NF-14-1-0384 under the Project: Shared Perception, Cognition and Reasoning for Autonomy. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

[1] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, 2013.

[2] D. C. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *CVPR 2012*, 2012.

[3] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 142–149. IEEE, 2000.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[5] P. Das, C. Xu, R. F. Doell, and J. J. Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[6] T. Feix, J. Romero, C. H. Ek, H. Schmiedmayer, and D. Kragic. A Metric for Comparing the Anthropomorphic Motion Capability of Artificial Hands. *Robotics, IEEE Transactions on*, 29(1):82–93, Feb. 2013.

[7] M. Jeannerod. The timing of natural prehension movements. *Journal of motor behavior*, 16(3):235–254, 1984.

[8] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. `http://caffe.berkeleyvision.org/`, 2013.

[9] J. Joo, W. Li, F. F. Steen, and S.-C. Zhu. Visual persuasion: Inferring communicative intents of images. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 216–223. IEEE, 2014.

[10] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS 2012*, 2013.

[11] Y. LeCun and Y. Bengio. The handbook of brain theory and neural networks. chapter Convolutional networks for images, speech, and time series, pages 255–258. MIT Press, Cambridge, MA, USA, 1998.

[12] I. Lenz, H. Lee, and A. Saxena. Deep learning for detecting robotic grasps. *International Journal of Robotics Research*, page to appear, 2014.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[14] A. Mittal, A. Zisserman, and P. H. Torr. Hand detection using multiple proposals. In *BMVC*, pages 1–11. Citeseer, 2011.

[15] Z. Mo and U. Neumann. Real-time hand pose recognition using low-resolution depth images. In *CVPR (2)*, pages 1499–1505, 2006.

[16] J. R. Napier. The prehensile movements of the human hand. *Journal of bone and joint surgery*, 38(4):902–913, 1956.

[17] I. Oikonomidis, N. Kyriazis, and A. Argyros. Efficient model-based 3D tracking of hand articulations using Kinect. In *Proceedings of the 2011 British Machine Vision Conference*, pages 1–11, Dundee, UK, 2011. BMVA.

[18] H. Pirsiavash, C. Vondrick, and A. Torralba. Inferring the why in images. *arXiv preprint arXiv:1406.5472*, 2014.

[19] A. Saxena, J. Driemeyer, and A. Y. Ng. Robotic grasping of novel objects using vision. *The International Journal of Robotics Research*, 27(2):157–173, 2008.

[20] K. B. Shimoga. Robot grasp synthesis algorithms: A survey. *The International Journal of Robotics Research*, 15(3):230–266, 1996.

[21] D. Song, N. Kyriazis, I. Oikonomidis, C. Papazov, A. Argyros, D. Burschka, and D. Kragic. Predicting human intention in visual observations of hand/object interactions. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 1608–1615. IEEE, 2013.

[22] R. Tubiana, J.-M. Thomine, and E. Mackin. *Examination of the Hand and the Wrist*. CRC Press, Boca Raton, FL, 1998.

[23] F. Warneken and M. Tomasello. Altruistic helping in human infants and young chimpanzees. *science*, 311(5765):1301–1303, 2006.

[24] D. Xie, S. Todorovic, and S.-C. Zhu. Inferring "dark matter" and "dark energy" from videos. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2224–2231. IEEE, 2013.

[25] Y. Yang, C. Fermüller, and Y. Aloimonos. Detection of manipulation action consequences (MAC). In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2563–2570, Portland, OR, 2013. IEEE.

[26] Y. Yang, A. Guha, C. Fermuller, and Y. Aloimonos. A cognitive system for understanding human manipulation actions. *Advances in Cognitive Sysytems*, 3:67–86, 2014.

[27] Y. Yang, Y. Li, C. Fermuller, and Y. Aloimonos. Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web. In *The Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, 2015.

[28] X. Zabulis, H. Baltzakis, and A. Argyros. Vision-based hand gesture recognition for human-computer interaction. *The Universal Access Handbook. LEA*, 2009.