This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Effectiveness of Grasp Attributes and Motion-Constraints for Fine-Grained Recognition of Object Manipulation Actions

Kartik Gupta
School of Computing and Electrical Engineering
IIT Mandi, India
kk1153@gmail.com

Darius Burschka
Dept. of Computer Science
TU Munich, Germany
burschka@in.tum.de

Arnav Bhavsar
School of Computing and Electrical Engineering
IIT Mandi, India
arnav@iitmandi.ac.in

Abstract

In this work, we consider the problem of recognition of object manipulation actions. This is a challenging task for real everyday actions, as the same object can be grasped and moved in different ways depending on its functions and geometric constraints of the task. We propose to leverage grasp and motion-constraints information, using a suitable representation, to recognize and understand action intention with different objects. We also provide an extensive experimental evaluation on the recent Yale Human Grasping dataset consisting of large set of 455 manipulation actions. The evaluation involves a) Different contemporary multi-class classifiers, and binary classifiers with one-vs-one multi-class voting scheme, and b) Differential comparisons results based on subsets of attributes involving information of grasp and motion-constraints. Our results clearly demonstrate the usefulness of grasp characteristics and motion-constraints, to understand actions intended with an object.

1. Introduction

An important domain of contemporary innovations in technology is the development of new techniques to assist humans in various daily household tasks and work environment tasks. An important aim of these techniques is to target the problem of modeling and monitoring the behavior of the individuals, and also help in transferring the object manipulation capabilities to the robots for performing both the household and workforce tasks. The crucial step in developing such applications is to recognize everyday human actions in various environments. Such action recognition based technologies can also benefit various domains such as entertainment, smart homes, elderly care, health rehabilitation, analyzing productivity of human work-tasks etc.

Clearly, it is very intuitive to associate action recognition with object interactions. Although action recognition for the actions specific to the objects is a problem which has been studied in [10], [13], the recognition and understanding of everyday human actions is still difficult due to some hard challenges. Everyday manipulation tasks include considerable amount of variations in a particular task being performed. Different people may have their own personal style of performing a particular action. Some of the manipulation tasks contain very subtle variations in the observed motions for the task being intended. These factors make the problem of recognizing manipulation tasks a challenging job. Thus, an action recognition approach which works on a large set consisting of hundreds of action classes is a very useful but challenging problem.

Figure 1. Instances from Yale human grasping dataset depicting (a) Precision grasp for opening the bottle and (b) Power grasp for drinking.

In this paper, we propose an approach to utilize the grasp and object motion-constraints based information for fine-grained recognition of everyday manipulation actions. A particular object can be manipulated differently according to the actions or task intended by the subject. We believe that the ability of fine-grained classification of observed motion to task specific parameters can allow one to transfer observed actions to manipulators with suitable motion

capabilities. For instance, a particular object (e.g. bottle) can be opened using a precision grasp and can also be used for drinking with power grasp as illustrated in Figure 1. This type of fine-grained classification allows to identify very useful information about the task intended by the user based on the grasp information. This is especially useful for a large set of actions (which is natural for everyday manipulation actions), as it would often involve the same object to be handled differently for different tasks. As we focus on the task of action recognition, we assume that the information about grasp attributes and motion-constraints is available to us, as in the case of Yale human grasping dataset [3] which we have used in this paper.

2. Related work

The problem of human action recognition and understanding has attracted a lot of interest in computer vision in recent years. The majority of action recognition approaches are targeting small set of action classes and therefore these approaches are not useful when it comes to their application on real everyday actions. Research on action recognition has been mainly focused on full-body motions that can be characterized by movement and change of posture like walking, waving, etc.

Analyzing human dynamics from video sequences is typically considered in many of these approaches [2], [17], [18]. With the advent of cheap depth cameras like Kinect, the problem of action recognition has been dealt using motion trajectories. These approaches (e.g. see [5]) are typically considered to be more robust in considering pose information which inturn aids recognition. However, Kinect body pose recognition readily fails when the user interacts with objects. This is a bigger impediment to its use in manipulation action recognition than its lack of precision. Motion dynamics based action recognition still lacks the ability to represent the subtle manipulations performed using objects. Also, the performed actions might have similar human dynamics but might differ in the goal of the task.

More closer to the problem considered here, is that of hand gesture recognition, which has also been addressed using depth data generated from Kinect in Kurakin *et al*. [14] and Wang *et al*. [19]. But these techniques mainly target sign language gestures and not the human hand-object interactions.

At this point, we note here that the above mentioned works involve processing low-level information (e.g. feature extraction from videos/images), whereas our goal in this work is to convey the importance of grasp and motion-constraints information at the higher semantic level (e.g. types of grasps and motion-constraints). Such high level attributes for manipulating actions, are indeed available as a part of the Yale human grasping dataset that we are considering in this work.

The problem of understanding manipulation actions is of great interest in robotics as well, where the focus is on simplifying methods to implement action execution on robots. Much work has been devoted to robot task planning based on imitation learning [1], which is essentially the problem of object manipulation through robots by imitating the real world trajectory observed on people performing the action to the robot body.

To the best of our knowledge, apart from [20] there has been no work using grasp information for action recognition. Yang *et al*. [20] semantically group action intentions using grasp based information into three coarse and somewhat abstract classes: Force-oriented, Skill-oriented, and Casual actions. Feix *et al*. [7] considers the problem of grasp classification on Yale human grasping dataset, again based on the coarsely defined task attributes such as force (interaction and weight), motion-constraints on objects and functional class (use and hold), whereas we propose a solution to task or manipulation action classification based on the grasp information, motion-constraints, and object class. To our knowledge, there has been no work on the problem of manipulation action recognition based on the motion-constraints on the objects being involved.

Unlike the works of [20] and [7], we consider fine-grained and physically interpretable action categories, also including object information. For instance, we consider the manipulation action of towel wiping and cloth wiping as two different tasks whereas Feix *et al*. [7] consider it as a single task. We believe that manipulation actions need to be classified at such a finer level to be able to serve the purpose of recognition of everyday manipulation actions and transferring complex task capabilities to the robotic manipulations. We differentiate between object manipulation actions, focusing on the functional property of an object. Thus, we demonstrate that information related to grasp, objects, and their motion-constraints are useful in achieving high recognition accuracy for a large set of action classes in an everyday manipulation action dataset.

Thus, the important aspects of our work include: a) A compact representation of the grasp and motion-constraints using some popular and some contemporary schemes. b) Demonstrating the usefulness of information from grasp attributes as well as motion-constraints for fine-grained action recognition. c) A differential experimental analysis involving subsets of grasp and motion-constraints features, to provide more insights on the usefulness of grasp information alone, motion-constraints information alone, and grasp and motion-constraints based information together for intended classification problem. d) An extensive experimental evaluation using different contemporary multi-class and binary classifiers (with a multi-class voting strategy), which also serves as a useful comparative study of popular classifiers for the manipulation action recognition problem. This

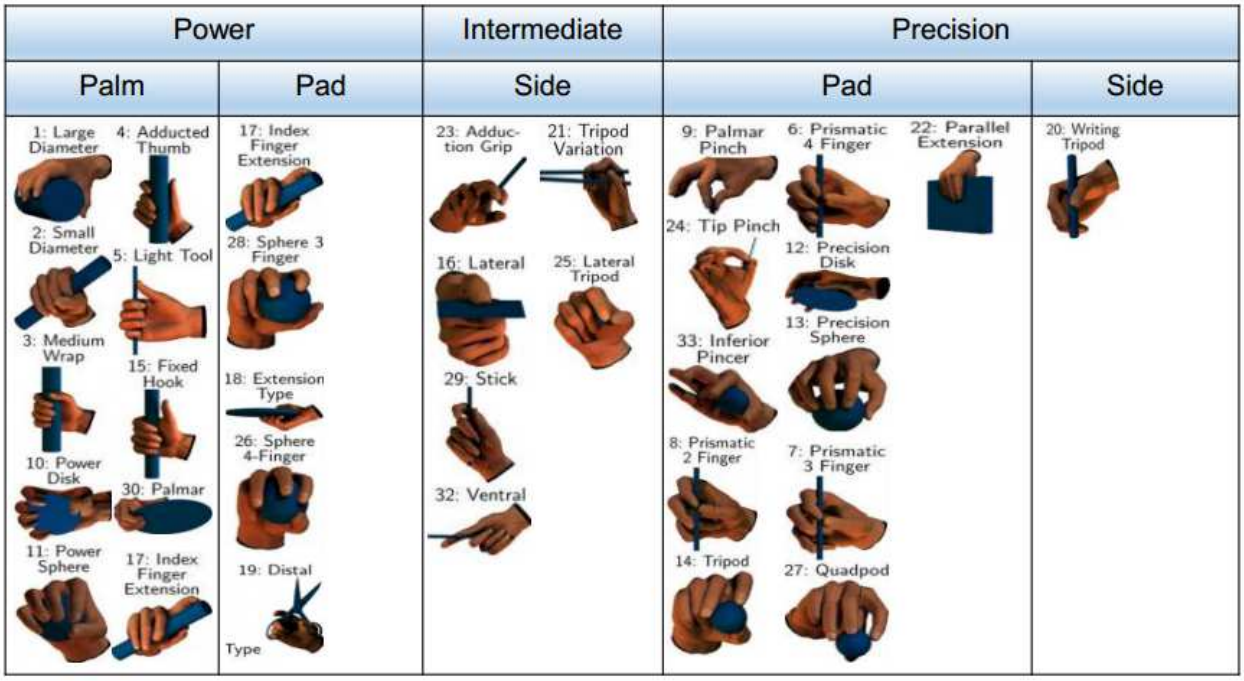

Figure 2. Coarse grasp categorization based on grasp taxonomy [9] where power, precision, and intermediate are grasp types and palm, side, and pad are opposition types.

analysis also helps to demonstrate that different classification frameworks, largely arrive at a consensus with respect to our hypothesis about using grasp and motion-constraints for fine-grained action classification. We demonstrate our results on a large Yale Human Grasping dataset [3] which involves various tasks on different objects.

The rest of the paper is organized into the following sections. In Section 3, we discuss in detail, the attributes used in our study, outlining the representation of the grasp and motion-constraints based information. Section 4 discusses the classification strategies that we have used. We present the experimental results and comparisons, with related discussions for various cases in section 5. We conclude in section 6.

3. Proposed approach: Attributes

Our approach for recognition of manipulation actions considers object information, grasps, and motion-constraints of objects.

3.1. Object

The object name (or corresponding symbols) serves as a simple string data on the information on the name of the object. As we want to perform classification of actions based on the grasp and motion-constraints information of the known object, we use the object name in the feature representation of an instance.

3.2. Grasp attributes

There are large number of grasp taxonomies available based on earlier research on grasp types. Although grasp types have also been classified at much finer level (e.g. [9]), we use it at a coarser level with grasp type as Power, Precision and Intermediate grasps, as also discussed in [9]. This type of coarse level grasp categorization is quite popular and relatively simple. We note that our assumption about the availability of grasp attributes, is much more valid for the coarse level attributes than the finer level ones, as the latter are arguably, more difficult to estimate. That is another reason for us to not use the grasp information at such fine-grained level for our classification of manipulation actions problem.

At the coarse level, each grasp can be classified by its need for precision or power to be properly executed. The differentiation is very important, and the idea has influenced many previous studies. In the power grasp, there is a rigid contact between the object and the hand that infers all the motion for the object is based on the human arm. For the precision grasp, the hand is able to perform intrinsic movements on the object without having to move the arm. In the third category i.e. Intermediate grasp, characteristics of power and precision grasps are present in roughly the equal proportion. We demonstrate that such a coarse division among grasp attributes is also useful for the purpose of manipulation action recognition. Figure 2 illustrates coarse

level grasp categorization for 33 grasp types specified in Feix *et al*. [9].

Apart from these categorizations, we further use three basic directions relative to the hand coordinate frame as illustrated in Figure 2 for 33 grasp types [9], in which the hand can apply forces on the object to hold it securely. Pad Opposition occurs between hand surfaces along a direction generally parallel to the palm. Palm Opposition occurs between hand surfaces along a direction generally perpendicular to the palm. Side Opposition occurs between hand surfaces along a direction generally transverse to the palm.

Opposition type mainly contains the information about the direction of grasp of the object whereas Grasp type contains the information about the force on the object. Both opposition type and grasp type consists of complementary information.

In addition, we also employ grasped dimension as another feature for representation, which signifies the specific dimensions (sides) of the object along which the object is grasped. For instance, a knife needs to be grasped along the blade to be able to be used for cutting purpose. We use the grasped dimension stated in [8] as the part of the object that lies between the fingers when grasped. The values are from the prior set (a, b, c) to indicate which axes best determine the hand opening. Here a is along the longest object dimension and c is along the shortest dimension. An example is illustrated in Figure 3. The grasped dimension contains crucial information about the handling of the object. It gives a spatial relationship between the human hand and object.

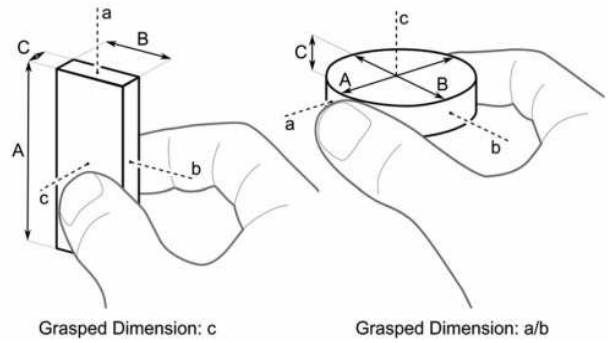

Figure 3. (Reproduced from [8]) Grasped dimensions for cuboid and round objects. For the round object grasp opening could be along both a and b dimensions.

3.3. Motion-Constraints on object being manipulated

Depending on the task (and also the object properties), an object is only allowed to translate and rotate in certain directions in order to successfully complete the task. In order to categorize motion-constraints for manipulation action, each of the three axes is assigned a symbol for the motion-constraints as abbreviated in Table 1. Thus, the resultant attribute can be represented as a string with three characters (symbols). Moreover, not all the combinations for three axes (i.e. 4^3 combinations) are practically valid, and only a set of 20 possible relative motions between two rigid bodies specified in [16], [15], are used. The nomenclature defines the relationship between the object and the environment (a fixed reference frame). Table 1 illustrates the symbols used for the motion-constraints along each axes of the object being manipulated by human hand to show whether motion for the object along an axis is unconstrained or allows translation/rotation or fixed.

Symbol	Translation/Rotation	Interpretation
u	unconstrained/unconstrained	unconstrained
t	unconstrained/fixed	translation
r	fixed/unconstrained	rotation
x	fixed/fixed	fixed

Table 1. Each of the three axes can either be free to move (u), only allow translation (t), only allow rotation (r) or do not allow any movement around that axis (x). Motion-constraints along x, y and z axes of the object is categorized using these generalizations.

4. Proposed approach: Classification

As discussed above, we represent an instance of a manipulation action using grasp label (power, precision and intermediate), opposition type (palm, side and pad), grasped dimensions of the object, object name, and motion-constraints on the object. To represent an instance i, we concatenate

Variable	Abbreviation
Grasp Type	ν
Opposition Type	ρ
Object	κ
Grasp Dimension	χ
Motion-Constraints	α

Table 2. Abbreviations used for different grasp and motion-constraints attributes.

these string data abbreviated in Table 2 to form a feature vector x_i.

$$x_i = [\nu, \rho, \kappa, \chi, \alpha]^T \quad (1)$$

During our experimentation for differential analysis, i.e. to see the effect of subsets of the above attributes, we define the instance x_i by removing one or more attributes from equation (1).

4.1. Classification models

As to our knowledge, there is no other work related to grasp and motion-constraints attributes for fine-grained

Figure 4. Sample frames of Yale human grasping dataset depicting variation in grasps for the objects - screwdriver, hammer and pen respectively.

classification of manipulation actions. We provide classification results using various contemporary classification frameworks. These include multi-class decision forests and multi-class neural networks, and binary classifiers. We also employ multi-class classifiers constructed from binary classifiers (using one-vs-one multi-class voting scheme). Such methods include locally deep support vector machines, support vector machines, binary boosted decision tree, and binary neural networks.

Multi-class decision forests [6] and binary boosted decision trees [11], are extensions of decision tree based classifiers. A decision forest is an ensemble model that very rapidly builds a series of decision trees, learning from labeled data. Decision trees subdivide the feature space into regions with largely the same label. These can be regions of consistent category or of constant value, depending on whether we are doing classification or regression. Boosted decision trees avoid overfitting by limiting how many times they can subdivide and how few data points are allowed in each region.

In both multi-class and binary neural networks which we use, input features are passed forward (never backward) through a sequence of layers before being turned into outputs. In each layer, inputs are weighted in various combinations, summed, and passed on to the next layer. This combination of simple calculations results in the ability to learn non-linear class boundaries and data trends.

Support vector machines (SVMs) [4] find the boundary that separates classes by as wide a margin as possible. When the two classes cannot be linearly separated, one can use kernel transformation to project the data into higher dimension, wherein classes may be arguably more separable. Two-class locally deep SVM is a non-linear variant of SVM proposed in Jose *et al*. [12].

4.1.1 One-vs-One multi-class scheme

As indicated above, one can perform a multi-class classification using binary classifiers. Typically, such schemes use one-vs-one classification, and construct one classifier per pair of classes. This approach requires the modeling of $N(N-1)/2$ classifiers, where N denotes the number of classes. During the testing stage, the test sample receiving the most votes from any class label is assigned that label. In the event of a tie (among two classes with equal number of votes), the label selection is based on the class with the highest aggregate classification confidence by summing over the pair-wise classification confidence levels computed by the underlying binary classifiers.

5. Experiments and results

As mentioned earlier, we evaluate our proposed hypothesis on large Yale human grasping dataset [3] consisting of everyday manipulation actions. We also emphasize that, to our knowledge, this is the only publicly available dataset which considers such a large set of everyday manipulation action in an unstructured environment. We evaluate the classification using different multi-class classifiers and binary classifiers with one-vs-one multi-class voting scheme to model the grasp and motion-constraints information. We also provide some differential analysis over the attributes, to study their effect on classification.

5.1. Yale human grasping dataset

This dataset consists of large annotated videos of housekeeper and machinist grasping in unstructured environments. The full dataset contains 27.7 hours of tagged video and represents a wide range of manipulative behaviors spanning much of the typical human hand usage. It involves total of 455 distinct manipulation actions (excluding holding actions and the action without proper grasp information) performed by two machinists and two housekeepers. This dataset is illustrated in Figure 4 using some example images of different grasps on some of the common objects like screwdriver, hammer, and pen. The videos are acquired by a head mounted camera on each subject. All subjects have normal physical ability, are right handed, and have been able to generate at least 8 hours of data. The labels for each of the task attributes, grasp attributes and object attributes are available with the dataset itself. This dataset is annotated by the raters experienced in the domain.

Classifier	Grasp Type(PIP)	Opposition Type	Grasped Dimension	Grasp Information(All)
Multi-class decision forest	0.6460	0.6532	0.6508	**0.6966**
Multi-class neural network	**0.6810**	0.6820	0.6474	0.6929
Locally deep SVM (Binary)	0.6688	0.6908	**0.6677**	0.6943
Support vector machine (Binary)	0.6789	**0.6912**	**0.6644**	0.6854
Boosted decision tree (Binary)	0.6152	0.6291	0.5687	0.5721
Neural network (Binary)	**0.6973**	**0.7041**	0.6508	**0.7085**

Table 3. Action recognition results (top two accuracies in bold) for two fold cross validation evaluation based on different grasp attributes using different multi-class and binary classifiers.

5.2. Experimental settings

We evaluate the proposed hypothesis on Yale human grasping dataset using two fold cross validation scheme where 50% of instances of each action associated with an object are used for training purpose and rest are used for testing purpose. As it is not necessary that all the actions performed by one machinist/housekeeper are performed by other machinist/housekeeper, we do not use a cross subject evaluation here. We remove the instances of task for which raters are not able to annotate any grasp information. Also, the task $holding$ is trivial as a manipulation action so we get rid of those instances too. We ultimately concatenate the object and task string data for each instance to get manipulation action labels. These labels serves as our manipulation actions as the goal for us is to solve the problem of finding which action is being done using a particular object. We are left with 455 different manipulation actions after the cleaning of dataset for our purpose with a total of 6188 manipulation action instances.

5.3. Results and discussion

We first provide recognition results (Table 3) using only the object and grasp attributes (without motion-constraints). These results indicate that even parts of grasp information is quite useful enough to classify a large set of 455 complex manipulation actions. This information is useful to understand that even with methods to recognize grasp types at much coarser level, one can distinguish between the complex manipulation actions to some extent. Table 3 also shows differential recognition rates based on the individual grasp attributes (grasp type, grasped dimension, and opposition type). From these results, we can infer that opposition type contributes relatively more to the recognition results. However, most of the classifiers agree that the combined attributes do perform better than individual ones (as expected). In general, this clearly highlights that grasp attributes indeed provide quite useful information for manipulation action recognition, and the fact that we are using a large dataset, support such a hypothesis. Even with a large set of action classes, we are able to differentiate tasks based on the object and grasp information at a rate of 0.7085.

Classifier	Avg. accuracy
Multi-class decision forest	0.8235
Multi-class neural network	0.8262
Locally deep SVM (Binary)	**0.8445**
Support vector machine (Binary)	**0.8408**
Boosted decision tree (Binary)	0.7376
Neural network (Binary)	0.8327

Table 4. Action recognition results (top two accuracies in bold) for two fold cross validation evaluation based on motion-constraints attributes using different multi-class and binary classifiers.

We next provide, in Table 4 the recognition results with objects and motion-constraints alone (without grasp attributes), and in Table 5, results with all attributes combined together. These results indicate that motion-constraints appears to help the manipulation action classification, much more than grasp information. However, in Table 5, one can notice that most classifiers agree that grasp attributes further improves the overall classification up to some extent. Below, we take a closer look at the difference between grasp and motion-constraints, considering certain specific classes.

Classifier	Avg. accuracy
Multi-class decision forest	0.8310
Multi-class neural network	**0.8388**
Locally deep SVM (Binary)	0.8293
Support vector machine (Binary)	0.8150
Boosted decision tree (Binary)	0.7587
Neural networks (Binary)	**0.8446**

Table 5. Action recognition results (top two accuracies in bold) for two fold cross validation evaluation based on the grasp and motion-constraints attributes using different multi-class and binary classifiers.

The failure cases to the action recognition based on grasp are mainly of the objects which do not have any rigid structure. Such objects do not have a particular way of handling to complete an action, for e.g. towel, paper etc. The reason for lower recognition rates for manipulation actions using these objects based on grasp information is the non-rigid

structure of the objects. Out of 6188 total action instances 19% of total instances i.e. 1189 instances consists of manipulation actions using towel. These object manipulation actions still are able to achieve better recognition rates based on the motion-constraints attributes as most of the actions based on these objects allow limited degree of freedom for the motion of object, for e.g. $cloth/towel\ wiping$ on plane surface does not usually consists of rotation along two axes and translation along one axis.

We perform another experiment to support this hypothesis, by removing instances of object - towel, cloth, and paper (where $towel$ constitute 19% of instances of whole dataset). In Table 6, we provide the results for this experiment. One can clearly notice in the earlier recognition results (across Tables 3 and 4), the difference between the results with grasp and motion-constraints is of the order of 10% to 15%. However, after removing the "non-informative" classes from the grasp perspective, one can observe that the classification using grasp attributes has also improved dramatically. While, the motion-constraints still contribute more for the recognition, the difference between recognition using grasp and motion-constraints is now reduced to 2% - 3%. Such a differential analysis highlights that while motion-constraints are generally useful for recognition, grasp attributes are also important, except for a small fraction of classes.

Classifiers	Grasp	Motion-Const.	Both
Multi-class decision forest	0.7840	0.8022	0.8088
Multi-class neural networks	**0.7913**	0.8166	**0.8378**
Locally Deep SVM (Binary)	0.7876	**0.8236**	0.8286
SVM (Binary)	0.7819	0.8205	**0.8495**
Boosted Decision tree	0.6323	0.6948	0.7198
Neural Networks (Binary)	**0.8045**	**0.8218**	0.8318

Table 6. Action recognition results (top two accuracies in bold) for two fold cross validation evaluation using different classifiers after removing instances involving objects - towel, cloth, and paper.

Such an inference is vital considering that the motion-constraints information i.e. degrees of freedom of object for the manipulation action, is relatively difficult to understand from the manipulation actions as compared to grasp information at a coarser level, using existing methods. Thus, one can appreciate that such coarse grasp information (which is easier to compute) can still prove useful to the manipulation action recognition.

The above analysis also serves to provide a comparison among different contemporary classifiers, for the current task involving categorical features provided in Yale human grasping dataset. We note that in majority of the cases binary neural network yields high classification accuracies. In addition, SVMs and multi-class neural networks also perform well, and often provide close to highest accuracies. It is also observed that the decision forest classifiers yield relatively low classification rates.

6. Conclusion

We propose a novel approach to the recognition of everyday manipulation actions based on the grasp and motion-constraints information. We evaluate our hypothesis on large Yale human grasping dataset consisting of 455 action classes. Our results and detailed analysis clearly shows that grasp information contains important clue to the everyday manipulation actions. We consider the differentiation between the functionality of the object and show that this approach for recognition has a clear advantage over the traditional methods of action recognition based on the human dynamics. Another overall advantage to this approach is that this type of action analysis can work over huge set of action classes with very subtle variations in their motion dynamics. This approach works well with everyday manipulation actions thus have capability of transferring advance task capabilities to the robotics applications and modeling the human behavior in complex environment.

References

[1] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

[2] A. F. Bobick and A. D. Wilson. A state-based approach to the representation and recognition of gesture. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(12):1325–1337, 1997.

[3] I. M. Bullock, T. Feix, and A. M. Dollar. The yale human grasping dataset: Grasp, object, and task data in household and machine shop environments. *The International Journal of Robotics Research*, 34(3):251–255, 2015.

[4] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[5] L. Chen, H. Wei, and J. Ferryman. A survey of human motion analysis using depth imagery. *Pattern Recognition Letters*, 34(15):1995–2006, 2013.

[6] A. Criminisi, J. Shotton, and E. Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision*, 7(2–3):81–227, 2012.

[7] T. Feix, I. M. Bullock, and A. M. Dollar. Analysis of human grasping behavior: Correlating tasks, objects and grasps. *Haptics, IEEE Transactions on*, 7(4):430–441, 2014.

[8] T. Feix, I. M. Bullock, and A. M. Dollar. Analysis of human grasping behavior: Object characteristics and grasp type. *Haptics, IEEE Transactions on*, 7(3):311–323, 2014.

[9] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic. The grasp taxonomy of human grasp types.

[10] R. Filipovych and E. Ribeiro. Recognizing primitive interactions by exploring actor-object states. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–7. IEEE, 2008.

[11] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[12] C. Jose, P. Goyal, P. Aggrwal, and M. Varma. Local deep kernel learning for efficient non-linear svm prediction. In *Proceedings of the 30th international conference on machine learning (ICML-13)*, pages 486–494, 2013.

[13] H. Kjellström, J. Romero, and D. Kragić. Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding*, 115(1):81–90, 2011.

[14] A. Kurakin, Z. Zhang, and Z. Liu. A real time system for dynamic hand gesture recognition with a depth sensor. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 1975–1979. IEEE, 2012.

[15] G. H. Morris and L. S. Haynes. Robotic assembly by constraints. In *Robotics and Automation. Proceedings. 1987 IEEE International Conference on*, volume 4, pages 1507–1515. IEEE, 1987.

[16] J. D. Morrow and P. K. Khosla. Manipulation task primitives for composing robot skills. In *Robotics and Automation, 1997. Proceedings., 1997 IEEE International Conference on*, volume 4, pages 3354–3359. IEEE, 1997.

[17] C. Rao, A. Yilmaz, and M. Shah. View-invariant representation and recognition of actions. *International Journal of Computer Vision*, 50(2):203–226, 2002.

[18] J. Sullivan and S. Carlsson. Recognizing and tracking human action. In *Computer Vision–ECCV 2002*, pages 629–644. Springer, 2002.

[19] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3d action recognition with random occupancy patterns. In *Computer vision–ECCV 2012*, pages 872–885. Springer, 2012.

[20] Y. Yang, C. Fermuller, Y. Li, and Y. Aloimonos. Grasp type revisited: A modern perspective on a classical feature for vision. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 400–408. IEEE, 2015.