PREDICTING HUMAN GRASP LOCATIONS ON CUP HANDLES BY USING DEEP NEURAL NETWORKS TO INFER HEAT SIGNATURES FROM DEPTH DATA

Yijun Jiang, Sean Banerjee, Natasha Kholgade Banerjee

Clarkson University, Potsdam, NY
{jiangy, sbanerje, nbanerje}@clarkson.edu

ABSTRACT

In automated assisted living where a robot assists a human to interact with physical objects, an important challenge is for a robot to understand where humans are likely to grasp objects, so that the robot can present the object to a user in the most tenable configuration. In this paper, we present an approach that uses encoder-decoder convolutional neural networks (CNNs) to predict human grasp location on cup handles. The primary challenge addressed by our work is that object occlusion induced by the human hand prevents direct imaging of grasp location. Our approach uses the insight that once the object is released, the hand leaves a heat signature on the object surface due to the temperature differences between the human body and the ambient environment. Our CNNs learn a mapping between images obtained from traditional depth sensors as input and heat signatures of grasp locations imaged using a thermal camera as output. Given the depth image of a novel cup, our approach uses the trained network to predict the grasp probability distribution over the cup. Using a leave-one-cup-out approach, we obtain a mean absolute pixel-wise prediction error of 5.67 on 17 cups imaged from 7 orientations.

Index Terms— human grasp, thermal maps, depth, encoder-decoder, neural network, grasp prediction

1. INTRODUCTION

Historically, robots have been viewed as independently operating tele-supervised or automated entities. However, with the spread of robotic technologies into healthcare for the purpose of automated assisted living and robotic quality of life improvement, it is becoming increasingly important to provide robotic systems that co-operate with human beings to enable successful accomplishment of human goals. One of the challenges involved in facilitating seamless human-robot interaction is to provide a fluid physical interface between a human and a robot when manipulating everyday objects. A large body of work in robotics research focuses on predicting the optimal approach for a robot hand to grasp an object by using human demonstrators [1–3], by generating stereo data [4–6], or by predicting grasp affordances [7–9]. However, these approaches perform grasp prediction for a robot operating independently. They do not address the task of predicting where a human would hold the object such that the robot can hand the object to a human in the most tenable configuration for human grasp. For instance, while a simulation may suggest that the ideal grasp of a hot cup for a two-fingered robot is by the handle, in a real-world environment it would be dangerous if a robot were to hand a human a hot cup by the handle, as the human would suffer injury by holding the cup around the main body. To provide optimal human grasp, the robot must recognize that the most likely location for a human to grasp the cup is by the handle.

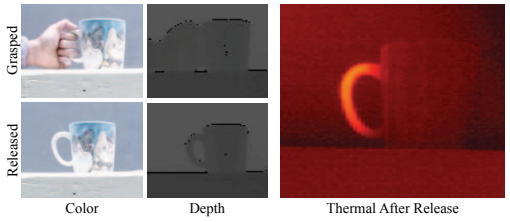

Fig. 1. Object occlusion by hand prevents detection of grasp location on object in color and depth images. On release, while the color and depth images show no information, the thermal image shows a heat signature at the grasp location. Our work uses the heat signature from thermal data to provide ground truth for training encoder-decoder networks to perform grasp location prediction from input depth data.

In this work, we address the problem of automatically predicting human grasp location on objects such as cups using depth data from a single viewpoint, such as may be captured by a depth sensor installed on a robotic arm. The main challenge in predicting human grasp lies in obtaining data on the location of human grasps on objects. When the interaction of a human hand with an object such as a cup is captured by a traditional sensor such as an RGB or depth camera, the object is significantly occluded by the hand due to the articulations of the finger, thereby preventing direct imaging of the grasp locations on the object as shown in the color and depth images of Figure 1. Our insight is that due to the temperature difference between the human body and ambient objects, hu-

978-1-5386-9214-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMEW.2019.00012

man contact with objects leaves behind a thermal signature which can be used to identify grasp locations. For instance, as shown in the thermal image of Figure 1, once a person lifts a cup exposed to the ambient environment by the handle and lets it go, the cup handle shows a region of high intensity at the location of the grasp. Our approach uses a thermal camera to image the heat variation over the surface of a cup induced by differences in the ambient environment and the human body temperature. In our experiments, the cups are placed at room temperature of 70° F, which is lower than human body temperature. This enables the thermal intensities to be interpreted as a probability distribution with higher values representing more likely grasp locations. Contact-based thermal data has been used to provide natural interactions on planar [10–12] and non-planar [13] surfaces. Our work is the first to use contact-based thermal data for grasp prediction.

Our work approaches grasp prediction as an image synthesis problem. We use a deep convolutional neural network (CNN) with encoder-decoder architecture to use the depth image of a cup as input and predict the thermal intensities depicting human grasp locations on the handle as output. We use a set of 17 cups of a variety of shapes and sizes, and image them from 7 viewpoints using a Microsoft Kinect v2 depth sensor and a Sierra-Olympic Viento-G thermal camera after a user has grasped and released each cup by the handle. Given each depth and thermal image, we use a leave-one-out approach to train encoder-decoder CNNs from sets of 16 cups and test on the left out cup. To perform grasp prediction as the synthesis of thermal intensities from depth images, our work draws inspiration from approaches that perform image-to-image translation [14–16]. To obtain high resolution decoded output, our approach uses the architecture proposed in [16] with transfer networks between the encoder and decoder layers. However, unlike [16] which uses the Rectified Linear Unit (ReLU) as the activation, our approach uses the Parametric Rectified Linear Unit (PReLU) [17] which provides higher average accuracy with respect to ground truth since the negative arm of PReLU avoids the dying neuron problem of ReLU.

Our contributions are summarized as follows: 1) To the best of our knowledge, we present the first approach that uses thermal cameras to address the hand occlusion problem from traditional cameras, and 2) we provide an encoder-decoder neural network with PReLU activation to accurately predict the human grasp locations on thermal images.

2. RELATED WORK

There exist a number of approaches in robotics to perform grasp understanding. Several approaches predict optimal grasp locations by searching for cylindrical shells that represent handle-like regions [18], or by learning from a sequence of grasp-and-drop actions [7]. Other learning-based techniques for grasp prediction include the use of support vector machines from pre-defined features describing the object and grasp [19], partially observable Markov decision processes [20], deep CNNs on segmented graspable objects [21], trial-and-error self-supervised CNN for grasp prediction without human labeling [8], and deep learning for successful grasp learned from the spatial relationship of a gripper and an object [9]. Unlike our work, these approaches estimate grasp for robot manipulators operating in independent environments, as opposed to considering environments where a robot may collaborate with a human.

In the area of collaborative human-robot grasp, approaches focus largely on manipulation as opposed to understanding of grasp for optimal human-robot interaction. For independent object handling, where a robot and a human interact with an object without simultaneous human-robot contact, a number of approaches focus on ensuring that human and robot trajectories do not collide [22–25]. In the area of handover of objects from robot to human and vice versa, a number of approaches have examined the influence of parameters such as object pose and orientation [26], timing [27, 28], spatial coordination [28], intent of human and/or robot [27, 28], preservation of distance, visibility and comfort constraints [29], and influence of secondary tasks performed by receivers on the primary handover task [30]. Some approaches use manual input to obtain priors on optimal robot pose for robot-to-human handover, e.g., using a set of manually defined rules derived from observation of human grasp [31], or by having users manually pose the 3D model of a robot arm holding an object via a GUI [26]. However, none of these approaches perform automated grasp understanding for collaborative interaction.

There exist many approaches that train robot grasp using human intervention. Detry et al. [32] hand objects to a robot to teach it to grasp. However, unlike our work, their aim is not to learn where a human would optimally grasp the object. Rather, the human may hold the object sub-optimally while the robot learns to hold the object. Kang et al. [1] provide a rule-based approach to deconstruct human finger configurations in a grasp around an object. Their method is tailored to simple objects such as cylinders and spheres, but does not readily generalize to objects of complex shapes. Takahashi et al. [2] propose a method to teach the robot performing tasks under human directions in virtual reality. Aleotti et al. [33] use motion tracking with a digital glove to track human grasp while subjects interact with objects in virtual reality environments. The use of digital gloves, the absence of tactile feedback, and incomplete physical and photo-realism hinder natural human motion in virtual reality environments. Our work uses non-invasive sensing and real-world objects to enable understanding of free-form human-object interactions.

3. DATA COLLECTION

Our experimental setup consists of a Microsoft Kinect v2 sensor of depth resolution 512×424 and a Viento-G thermal cam-

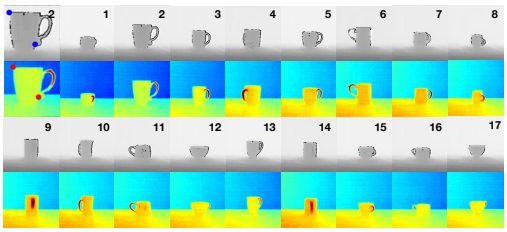

Fig. 2. Depth images from a Microsoft Kinect v2 sensor (first and third rows) and thermal images from a Viento-G thermal camera (second and fourth rows) captured from various viewpoints for cups of a range of shapes and sizes held and released by a user. For each cup, the corresponding thermal image shows the high intensities on the handle that arise due to heat transfer from the user's hand. The first depth and thermal image show the points marked manually on the largest cup (cup 2) for cropping all cup images.

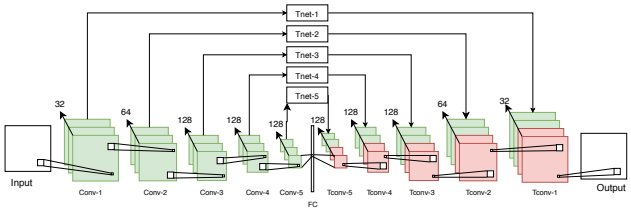

Fig. 3. Network Architecture. From left to right, each stack of rectangles represents: Input image, Conv-1, Conv-2, Conv-3, Conv-4, Conv-5, FC, Tconv-5, Tconv-4, Tconv-3, Tconv-2, Tconv-1, Output image. The green and red rectangles represent feature maps generated by Conv and Tconv blocks respectively.

era of resolution 640×480 placed at a distance of 5 cm from each other and 0.5 m from a tabletop. We use the sensors to capture the depth and thermal data for 17 cups of a variety of shapes and sizes. In our experiments, a single user lifts each cup by the handle outside the view of both sensors, holds the cup for 30 seconds to ensure stabilization of heat transfer, and places the cup in 7 different viewpoints on the tabletop. We immediately capture a thermal image using the Viento-G after the user has set down the cup in a particular viewpoint to prevent heat die-out, and we then capture a depth image using the Kinect. With 17 cups each in 7 orientations, we obtain 119 pairs of depth and thermal images. Figure 2 shows the depth and thermal images from a variety of viewpoints.

To obtain one-to-one pixel mapping from depth to thermal images for training the encoder-decoder CNN, we need to ensure that the depth and thermal images are aligned. While the accurate approach to perform the alignment is using stereo camera calibration, in this work, we obtain high accuracy of prediction with a simpler technique. We select two points on the largest cup at the locations shown in the first depth and thermal image of Figure 2. Given the two points, we estimate a 2D scale and translation to match the thermal image to the depth image. We use the scale to resize each thermal image to nearly the same resolution as the depth image, and the translation to align the cup in the two images to be at nearly the same location. We extract image crops of size 162×162 from the aligned depth and thermal images. Training depth and thermal images form the input and output respectively for our CNNs. To minimize overfitting, we perform data augmentation using random translations and scalings.

4. ENCODER-DECODER NEURAL NETWORK

Our neural network architecture, shown in Figure 3 is similar to the one in [16], with the exception that we use a shallow network with transposed convolution (Tconv) instead of the maxpool-unmaxpool framework of [16], fewer kernels per layer, and the Parametric Rectified Linear Unit (PReLU) [17] as the activation function instead of ReLU. The architecture consists of four parts: 5 convolution blocks (Conv), 1 fully connected layer (FC), 5 transfer networks (Tnets) [16], and 5 transposed convolution blocks (Tconv). The notations Conv-n, Tnet-n, and Tconv-n represent the n^{th} convolution block, the n^{th} Tnet and the n^{th} transposed convolution block. The Conv-n block is similar to that of VGG [34]. Each Conv-n block consists of a collection of two 3×3 layers that perform convolution, batch normalization (BN) [35], and PReLU activation, followed by a 2×2 max pooling layer.

Given a depth image of size $n \times n$, Conv-1 generates 32 feature maps with size of $n/2 \times n/2$. In Conv-2, we double the number of feature maps to compensate for the reduced complexity by max pooling from the previous block, i.e., we generate 64 feature maps with size of $n/4 \times n/4$. We repeat the process for Conv-3. However, we do not increase the number of feature maps from Conv-4 to Conv-5 to minimize overfitting. The output feature map size of Conv-5 is 32 times smaller than that of the input image. After Conv-5, we generate an FC layer using an additional 5×5 convolution-BN-PReLU layer. We adopt the Tnet cross-connection to facilitate the gradient flow from input to deeper layers and to convert the information from input to output domain. The Tnet is similar to [16] except that the activation functions are changed to PReLU. Tconv-1 to Tconv-4 each consist of a 2×2 transposed convolution with a 2×2 stride, BN and PReLU. Tconv-5 uses 5×5 transposed convolution with 1×1 stride for recovering the feature map size from Conv-5. Tnet-5 generates feature maps from Conv-5. We concatenate feature maps generated by Tnet-5 with those from Tconv-5, and feed them to Tconv-4. The remaining Tconv blocks repeat the process, yielding Tconv-n outputs with the same size as Conv-n. Finally, we use an extra Tconv block to generate output images with the same size as the input images.

Instead of using pre-trained networks, we train our neural networks from scratch. We define the loss

function as the mean absolute error (MAE), given as $(\sum_{j=1}^{m}\sum_{i=1}^{n}\left|p_i^j - l_i^j\right|)/mn$, where p_i^j and l_i^j denote the i^{th} pixel intensity of the j^{th} prediction and label samples respectively. We use stochastic gradient descent (SGD) where we set the learning rate to start from 1.0 and to be divided by 10 when the error plateaus, with a momentum of 0.9 and a weight decay of 0.0001. We train our neural networks for 100 epochs and select the model with lowest validation loss for testing, using 20% of training samples selected at random for validation. We perform testing using leave-one-cup-out cross validation, i.e., given n cups, we train with $n-1$ cups and test with the left out cup. Our neural networks are trained on four computers each containing an Intel Core i7-4790K 4.0GHz processor, 32 GB of RAM, and one NVIDIA GTX 1080 Ti GPU. Training time for each leave-one-out round is around 7 hours and testing time for each image is 3.7 milliseconds.

5. RESULTS

To verify the effectiveness of our neural networks using PReLU activation function, termed 'N-PReLU', we compare testing performance with a reference neural network which uses ReLU, termed 'N-ReLU'. The reference network is similar to that of [16]. In Table 1, we show the mean absolute error for N-PReLU and N-ReLU, averaged over all 7 orientations for each cup. Overall, we achieve 5.67 and 6.34 pixel-wise mean absolute error (MAE) for N-PReLU and N-ReLU respectively averaged over all cups. We define the pixel-wise mean range error (MRE) as $MAE_j/(\max_i(l_i^j) - \min_i(l_i^j))$, where l_j^i denotes the i^{th} pixel for the j^{th} ground truth image. The MRE provides an interpretation of network performance as a percentage of the range of ground truth intensity values. We achieve an MRE of 0.033 and 0.037 for N-PReLU and N-ReLU respectively. Figure 5 shows grasp likelihood prediction using N-PReLU for cups 5, 7, and 9 in our dataset in all seven orientations. In Figure 6, we show the results from various testing samples and viewpoints using N-PReLU.

In Figure 7, we compare the testing results of N-ReLU and N-PReLU. Although the numerical results of the two networks seem similar, N-ReLU occasionally predicts overheated temperature. As stated in [36], ReLU suffers from dying neurons, i.e., if the neurons are not initially activated, they are always in the off state as zero gradient flows through them. In our case, the predicted heat maps would be undesirable if neurons for the cup handle are never activated as shown in Figure 4. Specifically, the N-ReLU model predicts high intensities for cups 6 to 8 and cups 12 to 15 since its training process is negatively affected by dying neurons as shown in Figure 7. While the mean absolute error in Table 1 for some cups is higher with PReLU, e.g., for cups 4 and 7, the increase in error is due to the contribution from background pixels that are uninformative of grasp location. As shown in Figure 7, cups 4 and 7 show an intensity distribution more representative of high grasp likelihood at the handle for PReLU, whereas for ReLU, due to dying neurons the entire cup receives similar intensity information.

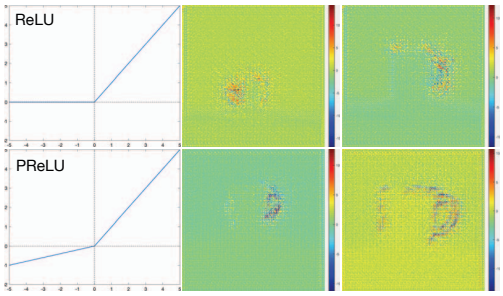

Fig. 4. Examples of intermediate feature maps after random weight initialization using ReLU (top row) and PReLU (bottom row). For ReLU feature maps, negative inputs exist around the handle (the left and right side of the first and second feature map respectively), which leads to dying neurons in those spots. The PReLU function avoids dying neurons enabling the gradients to propagate to the output.

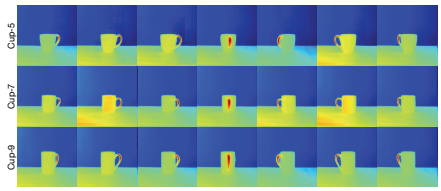

Fig. 5. PReLU results with all viewpoints of cups 5,7, and 9.

6. DISCUSSION

In this work, we estimate human grasp locations on cup handles by predicting heat maps from depth data using deep encoder-decoder neural networks. The advantage of using thermal information is that it enables imaging of grasp locations as heat signatures in comparison to traditional imaging techniques which prove ineffective due to hand occlusions during grasp. We validate the effectiveness of our approach by testing on novel cups. Our results demonstrate that our approach achieves small pixel-wise error for novel cups.

One issue with our work is that the prediction is inaccurate when the region corresponding to high grasp likelihood is not visible in the depth image but slightly visible in the ground truth thermal image as in the last column of Figure 3. This issue may be resolved by including training examples where the grasp region is completely occluded in both the depth and thermal images. Another limitation is that prediction depends upon the object being cooler than the human body, and will not work for hotter objects. One method of resolving this issue is to model human body temperature distribution and filter out thermal intensities that lie outside the distribution.

Table 1. Mean absolute pixel-wise errors of 17 cups using PReLU (top row) and ReLU (bottom row).

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Mean
PReLU	12.74	11.78	5.72	6.25	4.08	5.49	4.19	3.76	3.28	3.88	4.30	5.53	4.57	6.23	3.95	6.77	3.84	5.67
ReLU	12.59	12.89	4.90	4.78	7.78	5.71	3.31	3.31	3.59	6.62	4.20	8.16	4.07	6.65	6.23	7.87	5.01	6.34

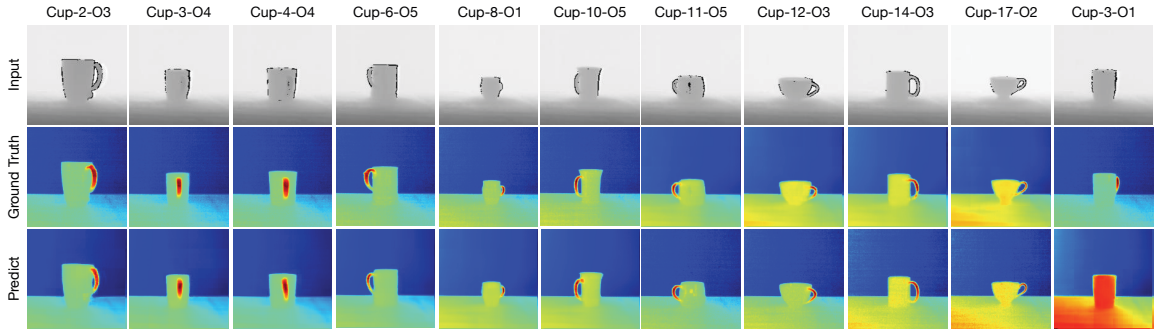

Fig. 6. N-PReLU testing results from various samples from different viewpoints. 'Ground Truth' represents the ground truth heat map. 'Predict' represents the testing results generated by networks using PReLU activation function. The last column shows the limitation of the neural network when the cup handle is occluded.

However, object and environment temperatures drive human-object interaction. For instance, a user may lift a cold cup in summer by holding the cup body, and a hot mug by the handle. On the other hand, in winter, a user may choose to hold the hot mug by the cup body to increase hand warmth. In future work, we will perform large-scale studies on the influence of object, the human body, and environment temperatures on human-object interaction.

Our future work also includes performing single-person and multi-person grasp predictions for left, right, and both hands on a variety of objects with a range of shapes, sizes, and weights. While everyday objects show instance-specific diversity, they can be clustered into categories based on human use. E.g., bottles, cans, jars, and small plant pots may be held similarly by a hand curl; pans, pots, spoons, and toothbrushes may be held by a a four finger grip on a long handle; and cartons and boxes may be held by a rectilinear grip. In future work, we will use a large object dataset to learn use-based object categorization, and perform grasp detection on novel objects through the categorization.

7. ACKNOWLEDGEMENTS

This work was partially supported by NSF grant #1730183.

8. REFERENCES

[1] S. B. Kang and K. Ikeuchi, "Toward automatic robot instruction from perception-mapping human grasps to manipulator grasps," *IEEE Trans. Robotics & Automation*, 1997.

[2] T. Takahashi and H. Ogata, "Robotic assembly operation based on task-level teaching in virtual reality," in *ICRA*, 1992.

[3] C. de Granville, J. Southerland, and A. H. Fagg, "Learning grasp affordances through human demonstration," in *ICDL*, 2006.

[4] P. Azad, T. Asfour, and R. Dillmann, "Stereo-based 6d object localization for grasping with humanoid robot systems," in *IROS*, 2007.

[5] P. K. Allen, A. Timcenko, B. Yoshimi, and P. Michelman, "Automated tracking and grasping of a moving object with a robotic hand-eye system," *IEEE Trans. Robotics & Automation*, 1993.

[6] R. B. Rusu, A. Holzbach, R. Diankov, G. Bradski, and M. Beetz, "Perception for mobile manipulation and grasping using active stereo," in *HUMANOIDS*, 2009.

[7] R. Detry, D. Kraft, O. Kroemer, L. Bodenhagen, J. Peters, N. Krüger, and J. Piater, "Learning grasp affordance densities," *Paladyn, Journal of Behavioral Robotics*, 2011.

[8] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *ICRA*, 2016.

[9] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *IJRR*, 2018.

[10] K. Oka, Y. Sato, and H. Koike, "Real-time tracking of multiple fingertips and gesture recognition for augmented desk interface systems," in *AFGR*, 2002.

[11] E. N. Saba, E. C. Larson, and S. N. Patel, "Dante vision: In-air and touch gesture sensing for natural surface in-

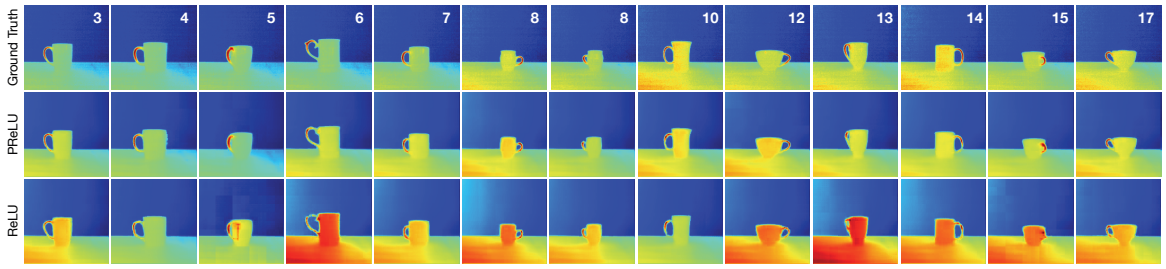

Fig. 7. Compared Results from various testing samples. The 'Ground Truth' represents the ground truth heat map. The 'PReLU' and 'ReLU' represents testing results generated by networks using PReLU and ReLU activation functions respectively.

teraction with combined depth and thermal cameras," in *ESPA*, 2012.

[12] T. Dunn, S. Banerjee, and N. K. Banerjee, "User-independent detection of swipe pressure using a thermal camera for natural surface interaction," in *MMSP*, 2018.

[13] D. Kurz, "Thermal touch: Thermography-enabled everywhere touch interfaces for mobile augmented reality applications," in *ISMAR*, 2014.

[14] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *ICCV*, 2015.

[15] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[16] L. He, G. Wang, and Z. Hu, "Learning depth from single images with deep neural network embedding focal length," *IEEE Trans. Image Processing*, 2018.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *ICCV*, 2015.

[18] A. Ten Pas and R. Platt, "Localizing handle-like grasp affordances in 3d point clouds," in *Experimental Robotics*, 2016.

[19] R. Pelossof, A. Miller, P. Allen, and T. Jebara, "An svm learning approach to robotic grasping," in *ICRA*, 2004.

[20] K. Hsiao, L. P. Kaelbling, and T. Lozano-Perez, "Grasping pomdps," in *ICRA*, 2007.

[21] Z. Wang, Z. Li, B. Wang, and H. Liu, "Robot grasp detection using multimodal deep convolutional neural networks," *Adv. Mech. Engg.*, 2016.

[22] J. Mainprice and D. Berenson, "Human-robot collaborative manipulation planning using early prediction of human motion," in *IROS*, 2013.

[23] P. A. Lasota, G. F. Rossano, and J. A. Shah, "Toward safe close-proximity human-robot interaction with standard industrial robots," in *CASE*, 2014.

[24] C. Pérez-D'Arpino and J. A. Shah, "Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classification," in *ICRA*, 2015.

[25] J. Mainprice, R. Hayne, and D. Berenson, "Predicting human reaching motion in collaborative tasks using inverse optimal control and iterative re-planning," in *ICRA*, 2015.

[26] M. Cakmak, S. S. Srinivasa, M. K. Lee, J. Forlizzi, and S. Kiesler, "Human preferences for robot-human hand-over configurations," in *IROS*, 2011.

[27] M. Cakmak, S. S. Srinivasa, M. K. Lee, S. Kiesler, and J. Forlizzi, "Using spatial and temporal contrast for fluent robot-human hand-overs," in *HRI*, 2011.

[28] S. Glasauer, M. Huber, P. Basili, A. Knoll, and T. Brandt, "Interacting in time and space: Investigating human-human and human-robot joint action," in *RO-MAN*, 2010.

[29] J. Mainprice, E. A. Sisbot, T. Siméon, and R. Alami, "Planning safe and legible hand-over motions for human-robot interaction," in *IARP*, 2010.

[30] C-M. Huang, M. Cakmak, and B. Mutlu, "Adaptive coordination strategies for human-robot handovers.," in *Robotics: Science and Systems*, 2015.

[31] A. H. Quispe, H. B. Amor, and M. Stilman, "Handover planning for every occasion," in *HUMANOIDS*, 2014.

[32] R. Detry, C. H. Ek, M. Madry, and D. Kragic, "Learning a dictionary of prototypical grasp-predicting parts from grasping experience," in *ICRA*, 2013.

[33] J. Aleotti and S. Caselli, "Grasp recognition in virtual reality for robot pregrasp planning by demonstration," in *ICRA*, 2006.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[36] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *ICML*, 2013.