Hand-Object Interaction Recognition Techniques Embedded in Wearables

An Introduction and Recommendation of the Paper:
H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions

Lin Yuyu
Zhejiang University
Industrial Design Department
21821015
linyuyu@zju.edu.cn

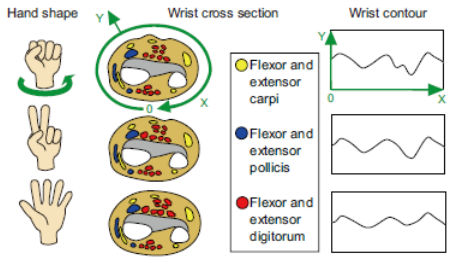
Abstract

The development of gesture interaction technology has gradually changed the way we interact with surrounding objects. Compared with many other methods of hand pose estimation, intuitively, vision-based method could directly recognize hand-object interaction in one step and support more kinds of gestures and objects. In this course final paper, I read and recommend the paper "H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions". I summarize the Hand-Object modal the paper present and extract the creative framework details, including the implementation steps, evaluation results and limitations. Finally, I envision the application scenarios where the methods could have practical uses, where I also present a simple object recognition prototype I made during the course and my inspiration after reading this paper.

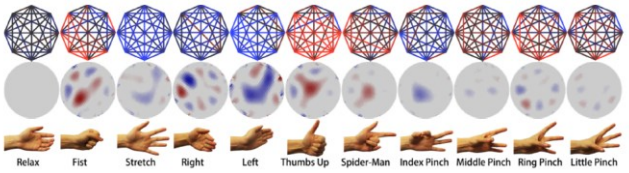
1. Introduction

Wearable gesture recognition bracelet has been brought out and developed for years, which indicates the promising future of NUI (Natural User Interface). However, in commercial and HCI research domain, EMW-based, contour-based and biological methods (Fig. 1), which respectively use electromagnetic wave transmitting and receiving, wrist contour recovering and bio-impedance or bio-capacitive sensing [2, 3, 4], are the most choices. They perform steadily in recognizing simple gestures to control the only constant object or express a limited number of meanings. Indeed, that covers the technical requirements of most hand-pose-based application scenarios. In contrast, vision-based gesture estimation methods attracted less attention when embedded in wearables because of the computational requirements and the bulky volume. However, hand-object interaction and action training require accuracy and real-time processing speed, [5] use electrical muscle stimulation and Unity to control and help users implement object behavior. I have also made a simple prototype of object recognition through MoblieNet as a necklace to test its feasibility in a wearable size [6] during this course.

In CVPR 2019, Microsoft and ETH Z¨urich proposes a novel unified framework in "H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions". They predicted simultaneously 3D hand and object poses, object classes and action categories through a data-driven architecture, and infer relations between hand and objects, directly in 3D and with egocentric cameras. In this course final paper, I read and propose a review of this paper including the Hand-Object model and experimental results. Finally, I present some of the application scenarios where the methods could have practical performances.



(a)



(b)

Figure 1: The gesture data of two popular methods beyond the CV field, the pictures are extracted from [2, 3]. (a) Contour-based methods with IR proximity sensors; and (b) Data visualization of Electrical Impedance Tomography (EIT) method.

2. Hand-Object Modal

As the author claimed in the paper, their model has following attributes:

1) Using single-shot regular RGB camera;

2) Reasoning about semantic meaning about the actions of the subject;

3) Understanding the environment by recovering the 3D object pose.

In my opinion, the most significant contribution is the 2nd item, which leads to the practical uses of wearable camera. To implement it, they propose a novel unified framework, which takes a sequence of images as an input and outputs per-frame 3D hand-object pose predictions, object and action classes along with the estimates of interactions for the entire sequence. The architecture shows in Fig. 2.

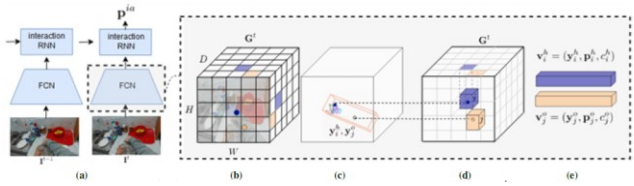


Figure 2: The framework of the H+O modal. (Extracted from the original paper)

2.1. Per-frame 3D hand and object poses

The authors start with a single image of a sequence. They parametrize both hand and object poses jointly with 3D control points, (21 skeleton joints for the hand pose and 3D locations of object keypoints with object bounding box.) That simplifies the regression task.

Interestingly, they further measure the distance (in metric space) between prediction and ground truth through the confidence of a prediction along with a given depth prediction.

2.2. Per-frame hand-object

This part predicts the object and action classes. As they defined, a *noun* and a *verb* respectively represent object and hand pose, so the action is a *(verb, noun)*. Given separate verb and noun predictions in 2.1, their model could recognize interactions. Ultimately, as the authors claimed, the network learns to jointly predict 3D hand and object poses along with object, action and interaction classes – with a single forward pass.

2.3. Temporal interaction reasoning

This is the most important part of the modal. To reason along the temporal dimension, the authors added a RNN module to their architecture, and a Long Short-Term Memory (LSTM) was used. As a result, they explicitly model interactions of hands and objects, directly in 3D

That improves the performance and accuracy because the type of interaction isn't only superposition of words. The dependencies between hands and objects count. Although co-training 3D hand and object pose estimation networks may mostly account for interactions, they further propose to model hand-object interactions at the structured output level with an interaction RNN. They model dependencies with a composite learned function and give the resulting mapping as input to RNN.

2.4. Training

As a unified framework, the complete model is trained in two stages. As the paper writes, they first train on single frames to jointly predict 3D hand and object poses, object classes and action categories, then keep the weights of the initial model fixed and train our recurrent network to propagate information in the temporal domain and model interactions.

3. Evaluation

3.1. Dataset

1) Training data: the SynthHands 3D hand pose estimation dataset;

2) Validation: FPHA-HO, a subset of First-Person Hand Action (FPHA) dataset, contains annotations for objects' 6-dimensional poses along with corresponding mesh models for 4 objects involving 10 different action categories;

3) A part of the EgoDexter hand pose estimation dataset annotating by the authors.

However, first-person viewpoint is from eyesight, while the viewpoint of camera on wristband is from hand-side. Therefore, there are no suitable dataset from a bracelet device.

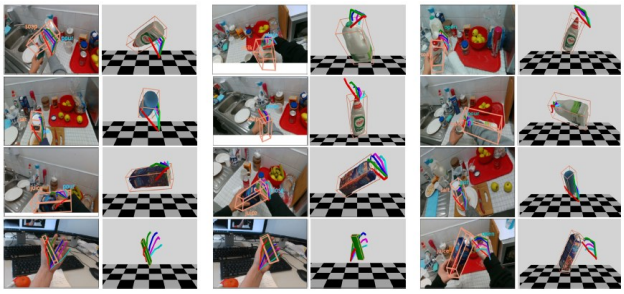


Figure 3: Visualization of the result. (Extracted from the original paper)

3.2. Results

The action accuracy of HAND + OBJECT POSE is 94.73%, while a state-of-the-art method from G. Garcia-Hernando is 91.97%. Besides, the action accuracy of HAND + OBJECT POSE + INTERACT is 96.99%.

Therefore, performance improvements are remarkable. The visual results show in Fig. 3.

The speed is feasible, their single pass network that produces per-frame predictions simultaneously for 3D hand poses, 6D object poses, object classes and action categories runs at real-time speed of 25 fps on an NVIDIA Tesla M40. The interaction RNN module further processes a sequence with virtually no time overhead, at an average of 0:003 seconds.

4. Limitation

This paper propose a novel framework to recognize hand-object interaction. Although the authors haven't mentioned its application scenarios, they would definitely never embed it in a wearable. However, personal hand-gesture control and communication would possibly better suit our daily needs in a personal device as small and light as possible in the near future. Therefore, the dataset of a first-person viewpoint is unfit for the case of a camera on wristband. Besides, hand-side recognition would encounter more severe occlusion issue. Although there are no suitable dataset from a bracelet device, the training architecture would be the better reference and inspiration to a brand new solution.

5. Applications

5.1. Handicraft preserving

In traditional handicraft, hand-tool or hand-medium interaction is the core content of recording and teaching. The creation process can be accurately recorded by recording the movement track of the hand tool, such as embroidery, sculpture, batik and so on. By the way, it would be able to support long-distance and cross-generation learning so that more people who have interest in them could have access to the masters' skilled action. However, it is difficult to ensure the precision of the handmade process through body action recognition and tracking movement teaching, because the creation of handicrafts requires attention not only to the posture, but also to the intensity. In addition, we hope to guide the learner's behaviors through reasonably interesting interactions, such as the electrical muscle stimulation in [4]. By promoting both learners and robot arm to imitate the creation process and improve their skills, the method the recommended framework can serve to be conducive to the inheritance and preservation of handicrafts.

5.2. Sensory augmentation & substitution

Sensory augmentation or substitution systems fulfill our expectations to enhance or compensate existing capabilities of a functional sensory system by providing complementary information. For people with sensory impairments, sensory augmentation devices, such as glasses and hearing-aided, serve to augment the weak sensory signals, while substitution devices compensate with another sense. A few studies work on enhancing real-world sense, although gesture and object pose information presents many meanings of the surroundings.

The applications of CNN models on embedded vision system become feasible since Depthwise Separable Convolution and MobileNet [6] were proposed, which greatly reduced the computational complexity of CNN models at the cost of accuracy, the application of CNN models on embedded system became possible. Now mobile object recognition devices have been widely developed, some of which create cross-modal immersive experiences.

For example, Hervé Goëau et al. [8] presented Pl@ntNet mobile app for plants recognition. However, research efforts have seldom connected computer vision with other senses.

5.2.1 A preliminary prototype

During the course, I try to connect vision and olfaction and make a preliminary prototype. I propose a compact design integrated camera and multi-scent display to be a pendant for hyposmia patients, which augments real-world olfactory perception synchronizing visual information. In detail, I set a brief goal to recognize seven common scent-evoked plants (the labels in Fig. 5) and trained a MobileNet model with 8 outputs (with a blank output indicating "no flowers in the image"). The architecture of the model is in Fig. 4. A pre-trained model with ImageNet dataset accelerates the convergence. I modify the last fully-connected layer of MobileNet and fine-tune it on our own dataset with positive samples of 7 scented objects and some negative samples. The model gets an accuracy of 79% on test-dataset and is deployed on Raspberry Pi zero with Tengine [7]. Fig.5 illustrates the Confusion for the performance. In an actual measurement, the system identified most of the testing objects no more than 2 meters straight away. Therefore, I improve the system with automatic clipping, and achieve the recognition in an extended range of 3-4 meters, with a reduced processing speed of 2s interval between 2 shots.

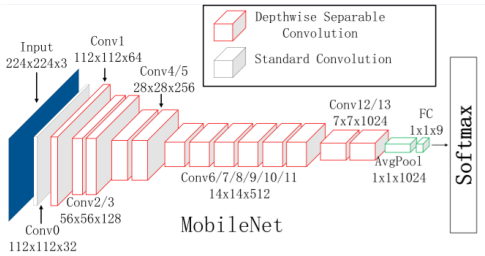


Figure 4. The architecture of MobileNet. Input: a resized image with the size of 224x224 and 3 channels(RGB); Conv: the output

of convolution layer; AvgPool: the output of average pooling layer; FC: the output of fully-connected layer; Softmax: a classification layer that output probability of each class.

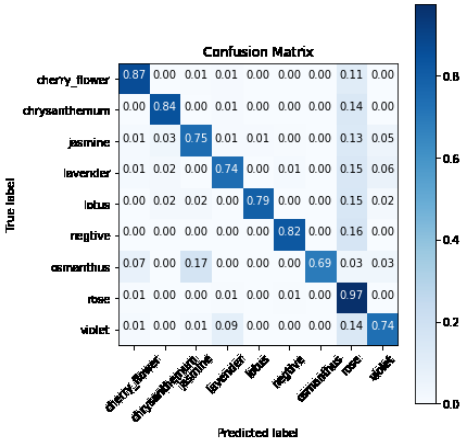


Figure 5. Confusion matrix for the performance of our model. The horizontal axis is the label predicted by MobileNet while vertical axis the true label of the test dataset. All data are normalized by the number of true labels of each class.

5.2.2 Inspiration

Hand-object interaction presents more information of the living environment, which is an inspiring method for sensory, or olfactory, to be exactly, augmentation and substitution. Therefore, embedded H+O model in a wearable would be a promising application for context and human's motivation awareness.

6. Conclusion and future work

In this CV course final paper, I focus on a promising applications of hand-object interaction in a wearable. I read and recommend the related model from the CVPR 2019 *H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions*. I raise a limitation of the method in the aspect of the use of a wrist device and envision the future applications. In the future work, I would fix the model and apply the fixed framework to practical use.

References

[1] Bugra Tekin, Federica Bogo, Marc Pollefeys. H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions. CVPR 2019.

[2] Rui Fukui, MasahikoWatanabe, Masamichi Shimosaka, Tomomasa SatoHand, Tomoaki Gyota. Shape Classification With a Wrist Contour Sensor: Development of a Prototype Device. UbiComp'11. ACM.

[3] Yang Zhang, Chris Harrison. Tomo: Wearable, Low-Cost, Electrical Impedance Tomography for Hand Gesture Recognition. UIST '15. ACM.

[4] Pedro Lopes, Patrik Jonell, and Patrick Baudisch. Affordance++: allowing objects to communicate dynamic use. CHI 2015. ACM.

[5] Yuyu Lin, Kai Zheng, Lijuan Liu, Yang Chen, Jiahao Guo, Shuo Li, Cheng Yao, Fangtian Ying. OlfacEnhancer: A Vision-Based Scented Necklace for Cross-Modal Perception and Olfaction Augmentation.

[6] Howard AG., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., & Adam H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861.

[7] https://github.com/OAID/Tengine

[8] Hervé Goëau, Pierre Bonnet, Alexis Joly, Antoine Affouard, Vera Bakic, Julien Barbe, Samuel Dufour, Souheil Selmi, Itheri Yahiaoui, Christel Vignau, Daniel Barthélémy, and Nozha Boujemaa. Pl@ntNet Mobile 2014: Android port and new features. 2014. In Proceedings of International Conference on Multimedia Retrieval (ICMR '14). ACM New York, NY, USA, 527. DOI: https://doi.org/10.1145/2578726.2582618