論 文 の 内 容 の 要 旨

論文題目　Hand-Object Interaction Mining from First-Person Videos
（一人称視点映像からの手-物体インタラクション マイニング）

氏　名　八木　拓真

Hands are our primary way to interact with the world. Understanding the interaction between human hands and the environment offers valuable insights into fields such as robotics, human-computer interaction, human-robot interaction, and virtual reality. Recently, hand-object interaction understanding from visual inputs has been gaining interest due to the widespread of mobile cameras. Numerous hand-object interaction recognition methods have been developed to recognize the user's short-term actions and spatial configuration of hands and interacting objects. These studies have been conducted in a controlled environment where user action is simple, the target object is evident, and the scene is static. However, the world we live in is far more complicated than we expect. People move around places and perform various actions to meet their needs. Multiple objects are simultaneously involved in an activity and their spatial configuration and appearance change over time by actions performed by the user. While this makes it difficult to even figure out the right object which is in interaction, such real-world aspects have not been taken seriously.

In this thesis, I present methods for recognizing when and for which object the hand-object interaction occurs that generalize to unknown, dynamic, and cluttered scenes in the real world. Specifically, I study the problems of (1) recognizing the contact state between a hand and an object, (2) identifying unique objects appearing in a real-world environment, and (3) their application on assisting users in finding lost objects. Towards developing models that work in real-world environments, they are designed through the unified concept of hand-object

interaction mining, which comprises the following properties: (i) learning from unlabeled data, (ii) category-agnostic formulation, and (iii) minimum user intervention. Off-the-shelf object detection, tracking, and segmentation techniques are used as a common component for automatically extracting useful knowledge from large-scale unlabeled data. Extensive data collection is conducted for evaluating and discovering unique difficulties that appear in a real-world setting.

In the first work, a method to predict contact states between hands and objects is introduced. Specifically, a video-based method that predicts a sequence of binary contact states (contact or no-contact) from a video and a pair of hand and object tracks is introduced. By predicting hand-object contacts, we can detect objects involved in interactions. However, annotating a large number of hand-object tracks and contact labels is costly. To overcome this difficulty, a semi-supervised framework with two new techniques is introduced: (i) automatic collection of training data with motion-based pseudo-labels and (ii) guided progressive label correction (gPLC) which corrects noisy pseudo-labels with a small amount of trusted data. Because there are no suitable datasets are available for evaluation in real-world environments, a new benchmark on a popular first-person video dataset is introduced. Experiments show that the learned model shows superior performance against existing baseline methods and generalizes well against novel objects and environments.

In the second work, the problem of category-agnostic object instance identification is studied. On understanding hand-object interactions across time, recognizing whether an object is the same one that appeared before will be one of the essential abilities. Because diverse objects appear in real-world environments, it is not realistic to pre-define the target category, and a class-agnostic solution will be demanded. However, no prior works exist on this challenging task, and fundamental difficulties in recognizing object instances in real-world environments were unknown. To this end, a large-scale, challenging benchmark consisting of more than 1,500 unique instances is built on top of unscripted, large-scale first-person videos. Strong metric learning-based baseline models, an in-depth evaluation of the dataset, and a performance comparison against previous datasets are introduced. The analysis shows that the trained model using the created dataset shows better robustness against

significant clutters in real-world environments.

In the third work, a practical use-case of hand-object interaction in assisting users in finding lost objects is introduced. People spend an enormous amount of time and effort looking for lost objects. To help remind people of the location of lost objects, various computational systems that provide information on them have been developed. However, prior systems for assisting people in finding objects require users to register the target objects in advance. This requirement imposes a cumbersome burden on the users, and the system cannot help remind them of unexpectedly lost objects. In this study, I propose GO-Finder ("Generic Object Finder"), a registration-free wearable camera-based system for assisting people in finding an arbitrary number of objects based on two key features: automatic discovery of hand-held objects and image-based candidate selection. Given a video taken from a wearable camera, GO-Finder automatically detects and groups hand-held objects to form a visual timeline of the objects. Users can retrieve the last appearance of the object by browsing the timeline through a smartphone app. To investigate how users benefit from using GO-Finder, two user studies are conducted. In the first study, the usefulness of GO-Finder is evaluated by a realistic object retrieval task. In the second study, the system's usability on a longer and realistic scenario is verified, accompanied by an additional feature of context-based candidate filtering. The usefulness of GO-Finder in realistic scenarios where more than one hundred objects appear is verified through experimental results and user feedback.