Master Thesis

Precise Affordance Annotation for Egocentric Action Video Datasets

(自己視点動作映像に対する厳密なアフォーダンスのアノテーション)

Zecheng Yu

48-206412

Department of Information and Communication Engineering
Graduate School of Information Science and Technology
the University of Tokyo

Thesis Supervisor: Yoichi Sato

January, 2022

Abstract

Object affordance has attracted a growing interest in computer vision. It is an important concept which builds a bridge between human ability and object property, and provides fine-grained information for other tasks like activity forecasting, scene understanding, etc. Although affordance has been investigated in many previous works, existing affordance datasets and methods failed to distinguish affordance from other concepts like action and function. In this paper, We propose an efficient affordance annotation scheme for egocentric action video datasets to address this issue. In our annotation scheme, we introduce a brand new affordance label form with the consideration of both object's property and agent's ability. we also develop a semi-automatic annotation scheme which could annotate affordance for large-scale video datasets with less effort, by utilizing the action labels of the datasets to locate video clips with the same affordance. Finally we apply this scheme to two large-scale egocentric video datasets: EPIC-KITCHENS and HOMAGE, and tested with various benchmark tasks: Tool-use/non-tool use action classification, mechanical action recognition, affordance recognition and grasp type recognition. Experimental results shows the rationality of our proposed affordance annotation scheme.

Acknowledgments

First, I would like to express my special acknowledgement to my respectable advisor Prof. Yoichi Sato for his continuous support of my Master's study and research. This thesis could not be done without his patient instructions and constructive suggestions. During these two years, he gave us enough space for finding our own research interest, and he will always guide us in time when we are stagnant. His enthusiasm on research also inspired me a lot.

Besides my supervisor, I want to thank my tutor Yifei Huang who helped me a lot both on my research and my life in Japan. Also, I would like to thank our research associate Ryosuke Furuta and project researcher Yusuke Goutsu as well. Their useful suggestions and advises helped me a lot on my research.

Thank everyone in Sato lab and Sugano lab for their questions and comments in our joint seminar, which helped me notice the problems exist in my research. And also for their company in my campus life.

At last, I would like to thank my parents: Yingqun Yu and Fengyun Zhang, for always giving me economic and spiritual support. I could not concentrate all my attention on my study and research without their support.

Contents

Figure List

Table List

Chapter 1

Introduction

1.1 Background

"I am washing the plate with a sponge". When we describe a human-object interaction in language, we naturally use verbs to denote the action and nouns for the interacting objects. Still, we don't need to describe how we perform this action on the object because we know how to interact with the object. The knowledge of how to interact with certain objects is learned from our daily experience, enabling us to know how to interact with a new object even if we see it for the first time. This is because we have learned that each type of handled object automatically activates afforded responses [9]. First defined by a psychologist named James Gibson in 1977 [10], the possibilities for action that objects or environments offer, i.e., "affordance," represents what the environment offers humans. After that, Norman [11] perfected the definition of affordance as a relationship between the object's properties and the agent's capabilities that determine the possible actions the object is afforded for. Studies from behavioral science [12, 13] and neuro science [14, 15] both showed that the afforded responses are activated automatically and unconsciously in our daily lives. While object affordances are implicitly hidden in our daily lives, understanding affordances is extremely important for a truly intelligent system. Since affordance can represent relationships between humans and objects, understanding affordance is an important step towards more robust action recognition, action forecasting, and other video understanding tasks, which

form the foundation of intelligent systems.

Unlike the actions or human-object interactions that are explicitly visible, affordance has been less studied in the field of computer vision because of its unconscious usage in our daily lives. Only recently have works been conducted on the scope of affordance understanding. For example, some researchers mainly focus on developing a model which could recognize object affordance from given pictures or videos [16, 17, 2, 7, 6]. Some works [18, 19, 20] also explored the use of affordance as context to improve other tasks such as action forecasting because of its ability to incorporate important knowledge. However, all of the previous works do not give an accurate definition of affordance. But instead, they simply regard actions or object functionalities as "affordance". The reason causing this issue is that existing datasets that contain affordance labels [4, 2, 3, 5] mainly directly use verbs as affordance labels, which is confusing since the same verb like "pick up" can represent different object affordances. For example, there is a huge difference in hand pose between picking up a box from the ground and picking up a shrimp using chopsticks. A well-structured, unambiguous definition of affordance is missing in all existing datasets.

Moreover, most existing datasets with affordance labels only contain videos taken from a third-person view, which does not adhere to our intuition. Since we learn to interact with the environment based on our experience, affordance is more sensible to humans from the first-person perspective. Fortunately, with the development of wearable devices, we can take first-person-view videos much more conveniently than before using head-mounted cameras. The videos taken by wearable head-mounted cameras observe exactly what the camera wearer sees, thus estimating object affordance from egocentric videos is obviously more reasonable because they are more in line with how humans recognize them. Therefore, we argue that besides constructing appropriate affordance labels, it is necessary to use datasets with first-person (or egocentric) videos so that the definition of affordance would be clearer and more intuitive.

In this thesis, we construct large-scale egocentric video affordance datasets with

a precise definition of affordance with the above motivations. We first make clear definitions of affordance by reviewing related psychology works. After getting the definitions, we develop an annotation platform and crowd-source the labels on two public representative datasets, namely EPIC-Kitchens [21] and HOMAGE [22]. We believe that our newly proposed definition of affordance and the annotated dataset can facilitate a deeper understanding of object affordance and further improve the subsequent tasks such as action recognition, anticipation, and robot imitation learning.

1.2 Challenges and Contributions

This thesis aims to construct large-scale egocentric video affordance datasets with a precise definition of affordance. The challenges are listed as follows: Firstly, the existing definition of affordance is confusing. Prior works chose possible actions on objects from verbs as affordance labels, and this makes different concepts such as human action, object function confused with affordance. We need to propose a criterion that could clearly distinguish affordance from other concepts such as human action and object functionality. Inspired by a recent psychological research [1], we propose a criterion which first classifies actions into two categories: *tool use action* and *non-tool use action* based on whether the actor is directly interacting with the object or through a tool, and then discriminate affordances and object functions by where they happened. Since we use affordance unconsciously in our daily life, it is impossible to use existing words such as "put" as the affordance label. Thus, in this thesis, we propose changing the form of affordance labels from one word to a combination of basic actions and grasp types. This can clearly represent the relationship between an object's property and an agent's ability and can make annotation much easier.

Secondly, collecting a large-scale egocentric video dataset that contains human-object interactions from scratch is not practical to be completed shortly. Instead, we leverage existing egocentric video datasets and give novel affordance annotations based on our definition. With the rise of interests in video understanding fields,

there are many egocentric action video datasets such as EPIC-KITCHENS [21] and HOMAGE [22]. They contain many human-object interaction video clips that perfectly fit our needs. However, manually annotating affordance labels for large-scale datasets which contain more than ten thousand video clips is also a laborious task. Therefore, we propose a semi-automatic annotation scheme by utilizing human habits on performing the same interaction with the same object, which could greatly reduce the workforce required for annotation. We use verb-noun-participant pairs to locate videos containing the same affordance, then manually annotate the affordance label for one of them. Finally, we assign this annotation to other videos with the same verb-noun-participant pair. This scheme could efficiently annotate affordance labels for video datasets recorded by limited participants and have action, object, and participant annotations.

Thirdly, it isn't easy to validate if the proposed affordance labels are reasonable or not. To show the rationality of our method, we apply our proposed annotation scheme on the EPIC-KITCHENS dataset and HOMAGE dataset. We have annotated 31,924 video clips for the EPIC-KITCHENS with an accuracy of 96.76%, and 14,689 video clips for the HOMAGE with an accuracy of 92.81%. Then we benchmark four tasks on the EPIC-KITCHENS dataset: Tool-use/non-tool-use action classification, mechanical action recognition, affordance recognition, and grasp recognition. Experimental results demonstrate that datasets constructed by our annotation scheme successfully categorize different affordances.

The main contributions of this thesis are summarized as follows:

1. We propose an efficient affordance annotation scheme for egocentric action video datasets with an precise definition of affordance.

2. We apply this annotation scheme on two large scale egocentric action video datasets: EPIC-KITCHENS, HOMAGE, and benchmark various tasks on them.

1.3 Organization of this Thesis

The rest of this thesis is organized as follows. In Chapter 2, we survey works related to ours, including the concept of affordance, affordance datasets, HOI datasets, and affordance understanding tasks. Chapter 3 proposes an efficient affordance annotation scheme with a precise affordance definition and applies it to two large-scale action video datasets. Chapter 4 tests our dataset on various benchmark tasks, including tool-use/non-tool use action classification, mechanical action recognition, affordance recognition, and grasp type recognition. Finally, Chapter 5 concludes our work and shows future works.

Chapter 2

Related Work

We propose an efficient, precise affordance annotation scheme and apply them to two large-scale action video datasets in this thesis. This chapter introduces previous research related to our work on affordance, affordance datasets, HOI video datasets, and affordance understanding methods.

2.1 Affordance

Since Gibson [10] first introduced the term "affordance" in 1977, many researchers in neuroscience explained their understanding of affordance. Gibson thinks affordances are action possibilities the environment offers the animal, which regards affordance as a property of the environment. This concept is widely applied in previous affordance works. They use action labels to denote possible actions on objects, making affordance confused with other concepts like action and function. Norman [11] perfected the definition of affordance as a relationship between the object's properties and the agent's capabilities that determine the possible actions the object is afforded for. This is more reasonable since the agent's ability also plays an important part in interacting with an object. For example, it is easy for an adult to move a chair but impossible for a baby. But Norman's definition of affordance doesn't mention tools. As an extension of human hands, we could utilize different tools to interact with objects in various new ways.

After reviewing previous works [23, 24, 25, 26, 27] of affordance in psychology,

Osiurak et al. [1] develops the three action-system (3AS) model to distinguish affordance from other concepts. As shown in Figure 2.1, they introduced 3 different concepts to represent different relationships that exist in human-object interactions: (a) affordance is the relationship between hand and object, which is hand centered; (b) mechanical action is the relationship between objects, which is tool-centered; (c) function knowledge is the contextual relationship that could help us to find the specific tool for some tasks. With the 3AS system, we could easily tell the differences between affordance, action, function, and other concepts. Besides, they also divide human-object interactions into two classes: (a) tool use actions: interacting with an object through an intermediate tool; (b) non-tool use actions: interacting with an object directly. According to the definition of the 3AS system, it is easy to spot that affordances present in both tool-use actions and non-tool use actions, but mechanical actions only exist in the tool-use actions.

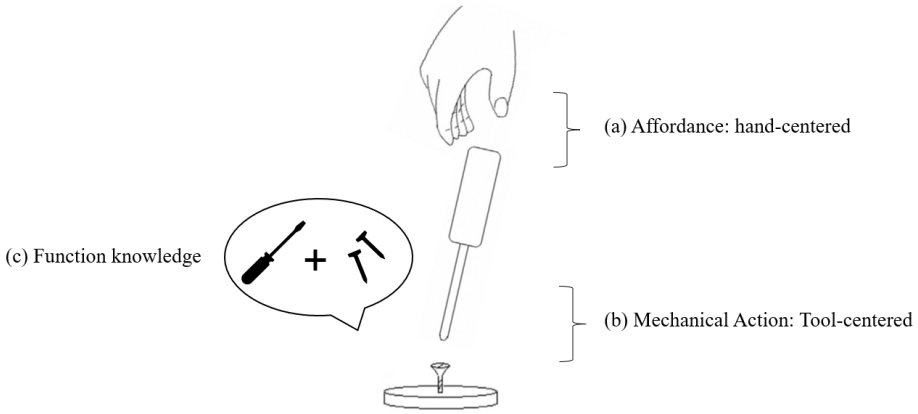

Fig. 2.1: Overview of three action-system [1]. (a) affordance is a hand-centered relationship between agent and object. (b) mechanical action is a tool-centered relationship between object and object. (c) function knowledge is the possible mechanical actions of object pairs we remembered.

2.2 Affordance datasets

Most earlier affordance datasets focus on images. They annotate possible actions and the exact region where these actions could occur for objects. Here are some of the representative works 2.2:

IIT-AFF: The IIT-AFF dataset is introduced by [2]. It provides 14,462 object bounding boxes and 24,677 affordance part pixel-wise annotations for 8835 RGB images. In this dataset, object parts that afford different actions are annotated with different colors. For example, in Figure 2.2 (a), the handle is annotated with "grasp", and the head is annotated with "hit". Actions and object functions are regraded as affordance in this dataset.

RGB-D Part Affordance Dataset: This dataset [3] contains RGB-D images and ground-truth affordance labels for 105 kitchens, workshop, and garden tools, and 3 cluttered scenes. Like IIT-AFF, different concepts such as "cut" and "scoop" are all used as affordance, confusing the definition of affordance.

(a) (b)

Fig. 2.2: Image based affordance datasets: (a) IIT-AFF [2]: object parts that afford different actions are annotated with different colors. Actions and object functions are regraded as affordance in this dataset. (b)RGB-D Part Affordance Dataset [3] contains RGB-D images and ground-truth affordance labels for 105 kitchens, workshop, and garden tools, and 3 cluttered scenes. Different concepts such as "cut" and "scoop" are all used as affordance.

Images have the disadvantages of only carrying limited information like the appearance of the objects, which is not enough for studying their affordance. Videos, especially human-object interaction videos, contain information on how humans interact with objects, which is a great reference for the model to learn about affordance. Recently, research interest in this field started to move to videos, including human-object interactions. Many works construct affordance datasets based on HOI videos, as shown in Figure 2.3:

CAD-120: The Cornell Activity Dataset is introduced by [4]. It contains 120 third-person view RGB-D videos annotated with 10 action labels and 12 affordance labels based on the action labels. Some affordance labels of this dataset, like "cuttable", "scrubbable," confuse affordance with mechanical action between objects. And the data amount of 120 videos is not enough for further studies.

SOR3D: This dataset [5] contains 20k RGB-D videos of human-object interactions recorded by asking annotators doing specific action on objects, annotated with affordance labels and pixel-wise affordance segments. "cut", "hammer", "paint"– object functions is also utilized as affordance in this dataset. And the scale of the dataset is too small for us to model daily used affordances.

OPRA: The Online Product Review dataset for Affordance (OPRA) [6] contains 11,505 demonstration video clips and 2,512 object images from YouTube. As shown in Figure 2.3, every instance of this dataset consists of a demonstration video clip, an object image annotated with interaction hotspots, and an action label. The hotspots indicate where this action could occur on the object. Different from prior datasets, affordance labels used in OPRA are pretty accurate. There are no affordance labels confused with other concepts like object functions. But it is still difficult to tell the differences between action and affordance if we define affordance as possible actions following Gibson's concept.

EPIC-KITCHENS Affordance: Nagarajan et al. [7] proposed a weakly supervised method for affordance heatmap generation by utilizing supervision from the EPIC-KITCHENS dataset's action label. To evaluate their work, they manually annotated the interaction region from 1,817 images. Using a large-scale action

video dataset greatly expends the data amount, and utilizing original action annotations of the dataset also reduces the annotation efforts. But this also leads to the problem of regarding affordance as possible actions without the consideration of the agent's abilities.

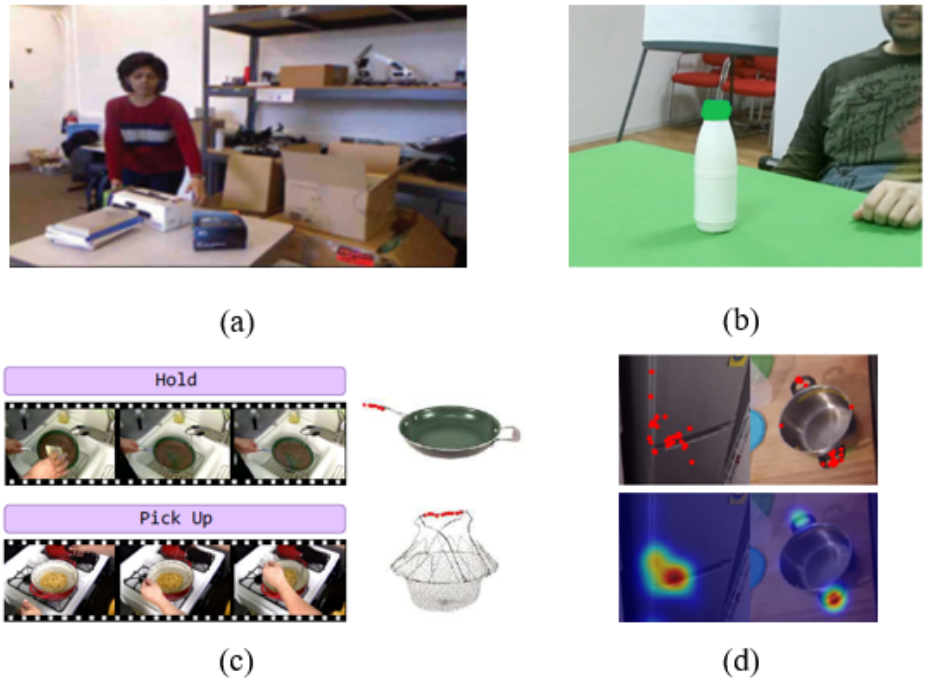

Fig. 2.3: Video based affordance datasets: (a) CAD-120 [4] contains 120 third-person view RGB-D videos annotated with 10 action labels and 12 affordance labels based on the action labels. (b)SOR3D [5] contains 20k RGB-D videos of human-object interactions annotated with affordance labels and pixel-wise affordance segments. (c)OPRA [6] contains 11,505 demonstration video clips and 2,512 object images from YouTube. Object images are annotated with interaction hotspots, and an action label. (d)EPIC-KITCHENS Affordance [7] manually annotated the interaction region from 1,817 images within the EPIC-KITCHENS dataset.

Table. 2.1: Comparison between different affordance datasets

	Format	Categories	Image/Video	Interaction Region	View	Affordance Labels
IIT-AFF	RGB Image	9	8,835	√	-	contain, cut, display, engine, grasp, hit, pound, support, w-grasp
RGB-D Part Affordance Dataset	RGB-D Image	7	105	√	-	grasp, cut, scoop, contain, pound, support, warp-grasp
CAD-120	RGB Video	6	130	×	third-person view	openable, cuttable, pourable, containable, supportable, holdable
SOR3D	RGB-D Video	13	9,683	√	third-person view	grasp, lift, push, rotate, open, hammer, cut, pour, squeeze, unlock, paint, write, type
OPRA	RGB Video	7	11,505	√	third-person view	hold, touch, rotate, push, pull, pick up, put down
EPIC-KITCHENS Affordance	RGB Image	20	1,871	√	egocentric	subset of EPIC-KITCHENS action set

As shown in the table above, these datasets failed to provide a clear definition of affordance, which is important to distinguish affordance from other concepts. For example, "cut" is a mechanical action between two objects. However, in these datasets, it is used as an affordance label. "Drink" is an action utilizing multiple affordances of the cup to achieve the goal of "drinking," but in CAD-120, it is used as the affordance label. These misusing cases in existing datasets confused affordance with action. Therefore we need an affordance dataset with a clear definition of affordance. Besides, datasets in egocentric view are missing, and this is not intuitive since we use affordance unconsciously in first-person view in our daily life. We could better investigate affordance with data in the egocentric view.

2.3 Action Video Datasets

Compared with existing affordance datasets, there are more large-scale action video datasets such as Kinetics [28], Something Something [29], EPIC-KITCHENS [21], HOMAGE [22], and so on. Many researchers use these datasets to investigate video understanding tasks like action recognition, action segmentation, etc. But fewer people utilize these large-scale video datasets for tasks related to affordance since there is no affordance annotation for them, and it is laborious to manually annotate a such number of video clips. We could save a lot of data collecting time if we could make use of these datasets.

Some of the action video datasets such as Kinetics and Something Something collect videos from online video sites like YouTube, which contains videos shot by

various uploaders. It is challenging to make use of those videos. Human-object interactions even don't exist in many of those videos. But there are also datasets built by limited participants with much more indoor human-object interactions. EPIC-KITCHENS and HOMAGE are two representations.

EPIC-KITCHENS: The EPIC-KITCHENS dataset [21] is the largest dataset in egocentric vision, and it contains 90k action segments annotated with action labels, 97 action classes, and 300 object classes. 32 participants in 45 kitchen environments recorded it. There are plenty of human-object interactions in kitchens scenes, benefiting affordance learning. Also, videos in egocentric view help us understand how affordances participate in our daily activities. A limited number of participants and scenes is also essential for reducing the efforts needed for annotation. Since different participants may interact with the same object in different ways, and object with the same name could be different in multiple scenes.

HOMAGE: The Home Action Genome dataset [22] (HOMEAGE) is a recently published large-scale multi-view video dataset of daily activities at home. It contains 24.6k action segments annotated with 453 classes of atomic action labels. 36 participants in limited indoor scenes also recorded it. They provide third-person view videos as well as egocentric videos. Same as the EPIC-KITCHENS dataset, it is also suitable for affordance annotation because of the large number of indoor human-object interactions, egocentric view, and limited participants and scenes.

Egocentric view, plenty of human-object interactions, limited participants, and scenes are all fit for affordance investigations. The only barrier is the laborious affordance annotation. To address this problem, we propose an efficient, precise affordance annotation scheme for this kind of dataset. We will illustrate it in chapter 3.

2.4 Affordance Understanding

Affordance understanding in computer vision has been researched from different perspectives. They can be mainly divided into four categories: *Affordance*

Recognition, *Affordance Semantic Segmentation*, *Affordance Hotspots Prediction*, and *Affordance as Context*. Given a set of images/videos, the task of affordance recognition aims to estimate object affordances from them. Azuma et al. [16] estimate object affordances using CNN and visual attention. Functions are used as affordance labels in this work. Ji et al. [30] first estimate object functions from its pictures, then feed the feature together with the category feature and the query feature into a linear layer to get the answer to the question. Pieropan et al. [31] proposed to recognize object affordance by modeling the Spatio-temporal relationship between objects using graphs.

Instead of only predicting affordance labels, Affordance semantic segmentation aims to segment an image/video frame into a set of regions that are labeled with an affordance category. Lüddecke et al. [17] proposed a ResNet50-based model for estimating affordance map from an RGB image. They use both the action and function of the object as affordance labels. [2] detects object affordances and their region from RGB images using CNN and CRF. Affordance labels are verbs including actions and functions. Moreover, [32] designs an encoder-decoder model for learning affordance segmentation mask from a given video clip, but it confuses affordance with function and action.

Affordance hotspots prediction [6, 7, 33] tries to predict affordance hotspot maps which indicate possible interaction areas for an object. Demo2Vec proposed by Fang et al. [6], as shown in Figure 2.4, utilizes action labels and heatmaps as supervision of the model to learn object affordance from demonstration videos. This work inspired some "learning from demonstration" works that partly solve video affordance annotation shortage. They use basic action labels: hold, touch, rotate, push, pull, pick up, put down. Although these labels are not confused with functions, they didn't consider the presence of the agent's ability, simply regrading affordance as possible actions on objects. Nagarajan et al. [7], use action labels from the EPIC-KITCHENS as supervision of their weakly supervised affordance hotspots methods. They first train an active anticipation network that could anticipate the active state of an inactive image and an action classifier using demonstration videos. At inference phase, the input object image is first fed into

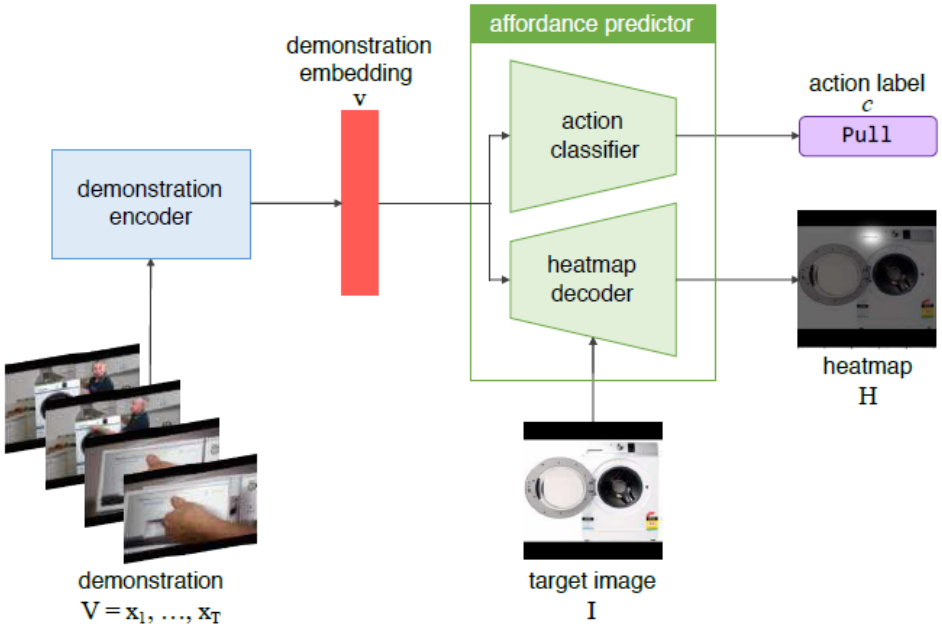

Fig. 2.4: Model Overview of Demo2Vec: The demonstration video is first encoded to a demonstration vector, then this vector is used for predicting the action label. The heatmap is generated by feeding the vector together with the target image into a heatmap decoder.

the active anticipation network and then fed into the action classifier to predict the action label. The affordance heatmap is generated by deriving the gradient-weighted attention maps over the original image. Since most video datasets provide action labels, they don't need any laborious annotation for training.

Instead of affordance detection and recognition, many prior works also use affordance as context because of its ability to incorporate important knowledge. Koppula et al. [18] use affordance to anticipate human activities. They first predict an object affordance heatmap from RGB-D video input, then use this heatmap as a clue to anticipate human movement trajectories in the future. Liu et al. [19] also utilize affordance for forecasting future actions. This work first estimates motor attention heatmap and interaction heatmap from the given video and then

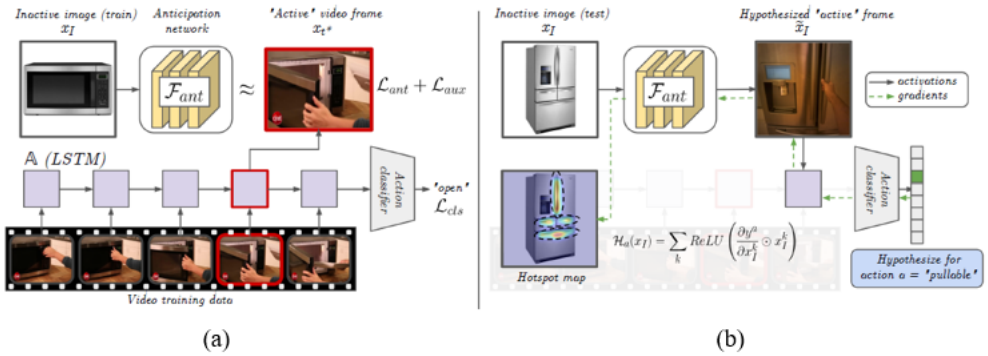

Fig. 2.5: Model overview of [7]: (a) Training phase: They first train an active anticipation network that could anticipate the active state of an inactive image and an action classifier using demonstration videos. (b) Inference phase: The input object image is first fed into the active anticipation network and then fed into the action classifier to predict the action label. The affordance heatmap is generated by deriving the gradient-weighted attention maps over the original image.

uses this heatmap to guide the action anticipation model. These two works show affordance's capability of indicating possible interactions on objects. Besides, Nagarajan et al. [20] proposed a method to anticipate future activities with the help of scene affordances, which stands for the possible actions in a scene.

Among these affordance understanding methods, most of them confused affordance with other concepts like action and function. This is caused by the lacking of datasets with precise affordance annotation. Although some of them try to avoid this problem by using weakly supervised methods, actions still can not replace the role of affordance as the relationship between the object's properties and the agent's ability. In the next chapter, we will introduce an efficient, precise affordance annotation scheme that can perfectly address the problems above.

Chapter 3

Proposed Method

The goal of this thesis is to develop an affordance annotation scheme which could distinguish affordance with other concepts, as well as annotate more video clips with less manual effort. Following the definition of the 3AS [1], we first propose an affordance label form which could both represent object's properties and agent's ability. Then we introduce a semi-automatic annotation scheme which needs little manual effort for annotation, and could be applied to egocentric video datasets which are recorded by limited participants and have action, object, participant annotations.

3.1 Affordance Annotation for Egocentric Video Datasets

When we talk about hand-object interactions, we focus on the interaction between our hand and the target object, for example, "Grab a ball", "put down knife", "open drawer". Besides, there are also certain type of interactions that need an intermediate object to perform: "cut a cucumber", "drink water", "wash dishes". We use *Non-Tool-Use Action*, *Tool-Use Action* to distinguish these two kinds of actions. As shown in Figure 3.1, tool-use actions utilize tools to interact with other objects, and tools can be seen as the extension of hands. Non-tool-use actions directly interact with objects. According to the definition of 3AS, affordances are hand-centered, animal-relative, and goal irrelevant properties of object. Hand-centered means affordance only presents between hand-object interfaces, and

this confirms the location of affordance. Animal-relative means object affordance is also related to the agents interacting with the object, and this makes affordance not only determined by the object's properties but also the agent's ability. Goal irrelevant differentiate affordance with the goal, to make sure that we could achieve different goals using affordance. For example, we could use the "grasp" and "rotate" affordance of a screwdriver to achieve the goal of "screwing a nail" because "grasp" and "rotate" are all goal irrelevant actions that could be the affordance of any object affords them.

On the other side, mechanical actions are tool-centered, mechanical action possibilities between objects. Tool-centered means mechanical action only presents between object-object interfaces, which confirms the location of mechanical action. Although many of us haven't seen this concept before, we can think of it as the function of tool objects. The verbs we usually use to describe actions between tools and objects in our daily lives are mechanical actions. For example, "cut" is the mechanical action between the knife and other objects, "screw" is the mechanical action between the screwdriver and other objects. In our proposed annotation scheme, actions between objects are used as mechanical actions. From these definitions, we can easily tell that mechanical actions only exist at the tool-object interface of tool-use actions, and affordances exist at the hand-object interface of both tool-use actions and non-tool-use actions. Therefore, we first need to classify original action labels into tool-use actions and non-tool-use actions and then annotate mechanical actions and affordances for them.

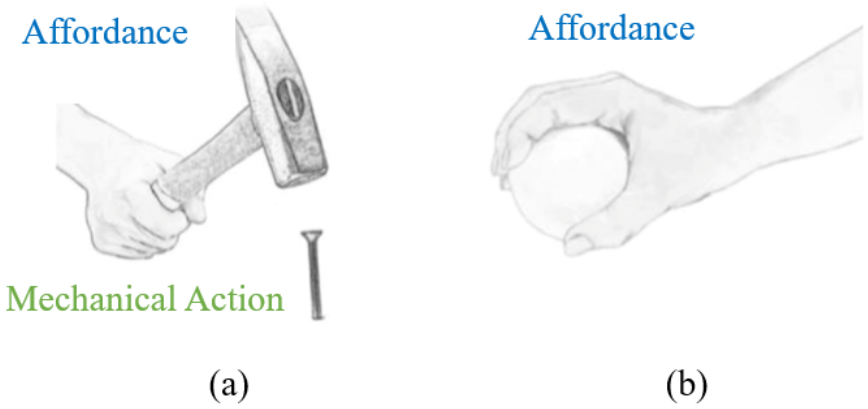

Fig. 3.1: Typical examples of (a)Tool-Use Action. (b)Non-Tool-Use Action. Tool-use actions utilize tools to interact with other objects, and tools can be seen as the extension of hands. Non-tool-use actions directly interact with objects. Therefore affordances exist in both tool-use actions and non-tool-use actions. Mechanical actions only exist in tool-use actions.

3.1.1 Tool-Use/Non-Tool-Use Action Annotation

Tool-use/non-tool-use action annotation for action video datasets can be done by dividing original action labels from the dataset into three categories: tool-use action, non-tool-use action, and both, according to the meaning of the action label. For example, all the instances of "take", "put", and "open" are non-tool use actions, while "fill", "cut", and "scrape" are tool use actions. Some action labels could include both tool-use action and non-tool-use action simultaneously, for example, "wash", "eat", "wear". We can simply ignore these labels during annotation because of their ambiguity. After grouping original action labels into tool-use/non-tool-use actions, we assign tool-use/non-tool-use action labels to all video clips based on their original action label's category.

3.1.2 Mechanical Action Annotation

Since mechanical actions exists in tool-object interfaces according to the definition, we only need to annotate mechanical actions for tool-use actions. Mechanical actions are relationships between tool and object. If you look into the action labels of tool-use action video clips, you may find that almost every original action label denotes the interaction between the tool and the object. For example: “stir food” – “stir” stands for the mechanical action between the slice and the food, “cut cucumber” – “cut” represents the mechanical action between the knife and the cucumber. On the basis of this rule, we could automatically annotate mechanical actions for all tool-use action video clips by using the original action annotation.

3.1.3 Affordance Annotation

Previous affordance datasets and affordance understanding methods failed to accurately define affordance and discriminate it with other concepts. The first reason causing this issue is that different concepts are all mixed as affordance labels, such as object functions. For example, ”cut” is the most misused affordance label in the existing dataset. Some use it as the possible action for cut-able objects, and others use it as the function of knife-like objects. This misuse of affordance breaks the consistency of affordance labels and confuses affordance-related tasks with other tasks such as action recognition task and function recognition task. Secondly, they regard affordance as possible actions on the object instead of relationships between the object’s properties and the agent’s ability following the definition of affordance. This makes affordance more like another kind of action and deprives its ability to incorporate contextual information between agents and objects. To address this issue, we need a proper affordance label form to represent both agent’s ability and object properties.

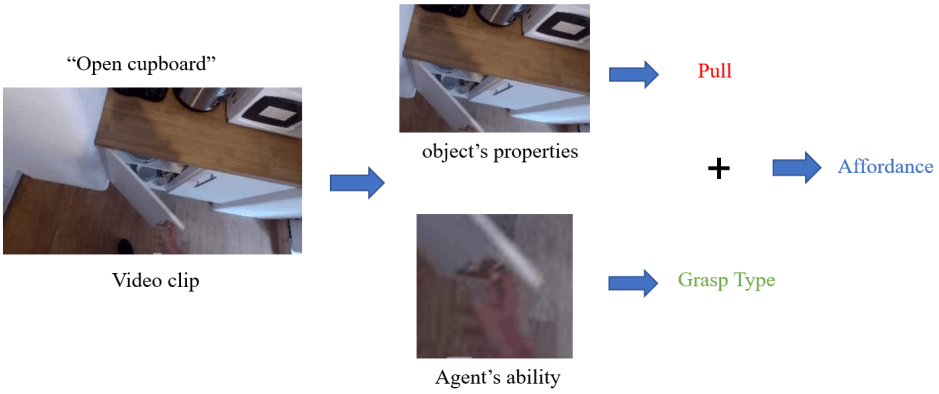

Fig. 3.2: Given an unlabeled video clip, we first confirm the original action label and the object. Then we come up with a goal irrelevant action according to the object property used in this action. Here we use "pull" to represent the "pullable" property of the cupboard. We use grasp types to represent the agent's ability. Finally, we combine the goal irrelevant action label with the grasp type label to get the affordance label of this video clip.

As shown in Figure 3.2, given an unlabeled video clip, we first confirm the original action label and the object. Then we come up with a goal irrelevant action according to the object property used in this action. Here we use "pull" to represent the "pullable" property of the cupboard. As for the agent's ability, we use grasp types. Possible grasp types determine what we can do when interacting with an object, so we use them to represent the agent's ability. Finally, we combine the goal irrelevant action label with the grasp type label to get the affordance label of this video clip. This affordance label form perfectly models the relationship between hand and object. At the same time, it also reduces the difficulty of affordance annotation. We use affordance unconsciously in our daily life, so it is hard to annotate affordance used in video clips. With our proposed affordance label, the annotator only needs to annotate actions and grasp types based on the video, which is much easier and introduces less bias.

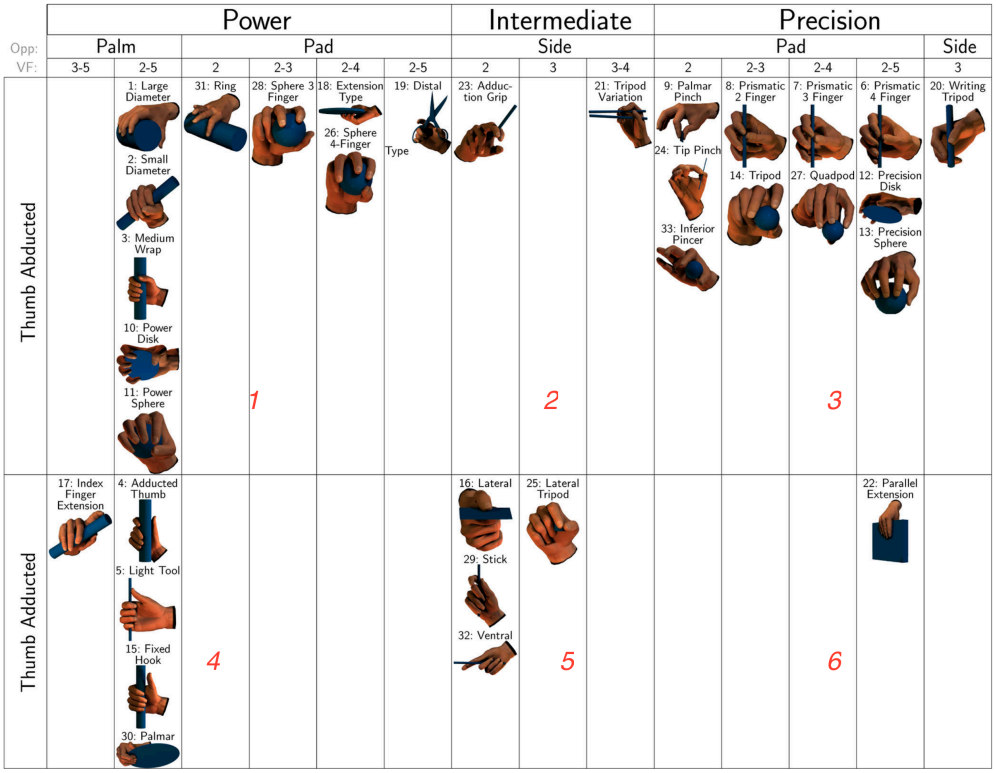

Fig. 3.3: The 33-class grasp types taxonomy introduced by [8] covers most daily used grasp types. We further narrow the grasp types into 6 categories according to the power of the grasp type and the thumb's posture to simplify the annotation. Grasp type 1 are powerful, thumb abducted grasps usually used to hold weight objects. Grasp type 2 are intermediate, thumb abducted grasps usually used to clamp objects. Grasp type 3 are precise, thumb abducted grasps usually used to do precise operations such as writing. Grasp type 4 are powerful, thumb abducted grasps usually used to grasp door handles. Grasp type 5 are intermediate, thumb abducted grasps usually used to grasp stick-shaped tools such as a knife. Grasp type 6 are precise, thumb abducted grasps usually used to grasp flat objects.

We use the 33-class grasp type taxonomy from [8], they divided grasp types into 33 categories. It could cover mostly grasp types we use in our daily life. We further narrow the grasp types into 6 categories according to the power of the grasp type

and the thumb's posture and use grasp type 0 for interactions that don't have a complete grasp, such as "touch". Results are shown in 3.3. This further simplifies the procedure of annotation. During annotation, grasp type is chosen based on both the semantic meaning of the interaction in the video clip and the appearance of hands, instead of only focusing on the appearance. Since we use grasp types to represent the agent's ability, we need to analyze why the agent is using this grasp type, such as fitting the object's shape or exerting more force on the object, rather than investigating the precise shape of the grasp. This could also help us eliminate ambiguities. As shown in Figure 3.4, if we only focus on the hand's appearance when annotating the "open fridge" video clip, it is hard to tell whether it is grasp type 4 or grasp type 5. But if we consider the semantic meaning of "open fridge", we should know that we need more power to open the fridge, and this can not be done by using grasp type 5. Consequently, annotation's accuracy is higher than before.

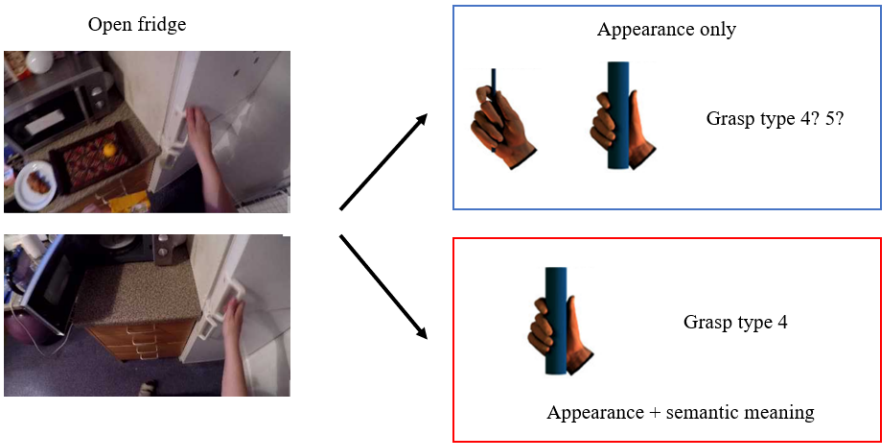

Fig. 3.4: On the left are two video clips in the EPIC-KITCHENS dataset, their original action label are all open fridge. If we chose grasp type only by the appearance of hands, it is difficult to tell whether it is grasp type 4 or grasp type 5. But if we consider the semantic meaning of "open fridge", we should know that we need more power to open the fridge, and this can not be done by using grasp type 5.

Only having an accurate affordance label form is not enough. Because manually annotating thousands of video clips is impractical, a more efficient automatic annotation scheme is imperative. Inspired by our nature that people are used to interacting with an object in a specific way. For example, different people may have different habits of grasping a pencil, but they will not change their manner of grasping the pencil in the future. Based on this phenomenon, we could assume that the same participants will perform the same interaction with the same object in a fixed way. This is also why we chose the EPIC-KITCHENS dataset and HOMAGE dataset. They are all shot by limited participants in limited scenes, simplifying our annotation process.

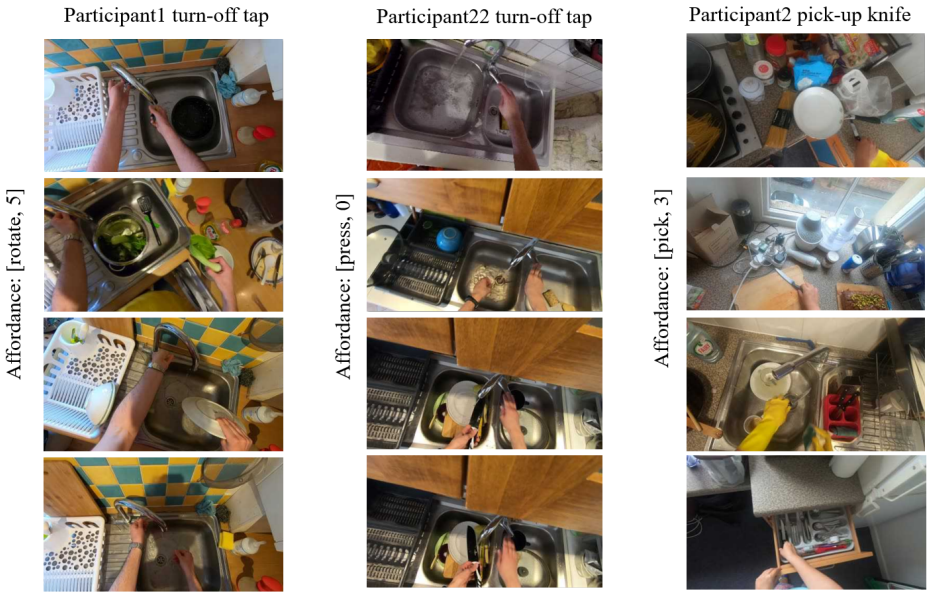

Fig. 3.5: Each column lists video clips within each verb-noun-participant pair. The first column are videos labeled with turn-off tap and performed by participant 1. He/She used affordance [rotate, 5] in these video clips. The second column are videos labeled with turn-off tap and performed by participant 22. He/She used affordance [press, 0] in these video clips. The third column are videos labeled with pick-up knife and performed by participant 2. He/She used affordance [pick, 3] in these video clips. It is easy to notice from each column that one participant always performs an action on the same object using the same affordance. By comparing columns 1 and 2, we could tell that the different participants may perform an action with different affordances because of different objects and personal habits.

With the original action, object, participant labels of the action video dataset. We could easily locate video clips using the same affordance by "verb-noun-participant" pairs."verb-noun-participant" pair consists of three important elements: "verb" is the action label of the video clip, "noun" is the interaction object's label, and "participant" is the participant's id. "Verb-noun" pairs could find out unique interaction video clips from the dataset. In most cases, we use a specific affordance for the same verb-noun pair as we have the same body structure. But with the

consideration that different participants could have different habits on interacting with an object, we add the participant to increase the precision of our annotation scheme further.

Given an action video dataset, we first find out unique verb-noun-participant pairs. We could determine a specific affordance for each pair with these three elements. Then assign this affordance to all video clips with the same verb-noun-participant pair. As shown in Figure 3.5, most affordances used in different video clips within the same verb-noun-participant pair are the same.

This method greatly reduces the burden of labeling, and we could annotate 10 thousand video clips by manually annotating a few hundred of them. But we also need to notice the scene and object change inside each pair because this happens sometimes.

We could efficiently annotate an egocentric action video dataset with precise affordance labels with this affordance annotation scheme. We will apply this scheme to two large-scale egocentric video datasets in the next part.

3.2 EPIC-KITCHENS Dataset

In this section, we will introduce the process of annotating EPIC-KITCHENS dataset with our affordance annotation scheme, and show the results of annotation. The EPIC-KITCHENS dataset [21] is one of the largest dataset of egocentric videos, which contains 90k action segments annotated with action labels, and in total 97 action classes, and 300 object classes. It was recorded by 32 participants in 45 kitchen environments. We leverage the existing verb and noun labels to construct our new annotation.

3.2.1 Tool-Use/Non-Tool-Use Action Annotation

We annotate tool-use/non-tool-use actions for EPIC-KITCHENS dataset based on the meaning of their original action labels, the results is shown in 3.1:

Table. 3.1: Tool-use/non-tool-use action annotation for EPIC-KITCHENS

Category	Video Clips	Action labels
Non-tool-use	51.5k	take, put, open, close, insert, turn-on, turn-off, move, remove, throw, shake, adjust, squeeze, empty, press, turn, check, apply, fold, break, pull, pat, lift, hold, wrap, look, unroll, sort, hang, sprinkle, rip, search, crush, stretch, knead, divide, set, feel, drop, slide, gather, turn-down, transition, increase, wait, lower, smell, let-go, finish, serve, uncover, unwrap, choose, lock, flatten, switch, carry, unlock, bend, unfreeze
Tool-use	8.5k	cut, pour, mix, dry, scoop, peel, flip, scrape, fill, scrub, filter, spray, cook, add, rub, soak, brush, sharpen, drink, water, attach, coat, measure, unscrew, form, use, grate, screw, stab, season, prepare, bake, mark

3.2.2 Mechanical Action Annotation

There are 33 tool-use action labels inside the EPIC-KITCHENS dataset. As we mentioned above, original action labels of tool-use action videos represent the interaction between tool and object. So we use these original action labels as mechanical action annotations. We have got 8.5k mechanical action annotations without any manual annotations.

3.2.3 Affordance Annotation

We apply our proposed semi-automatic affordance annotation scheme to the EPIC-KITCHENS dataset. First, we need to manually annotate affordance labels

for each verb-noun-participant pair:

1. Find out top 300 verb-noun pairs from the original annotation which could cover half of all video clips.
2. Randomly sample 5 video clips for every participants inside every verb-noun pairs.
3. Determine from the gallery if there are different scenes or objects inside these clips .
4. If no, annotate one video clip.
5. If yes, annotate one video clip for each scene or object.
6. If active object is invisible due to occlusion or blur, annotate it with 'none', then annotate next video clip instead.

Then we can assign annotated affordances to their neighborhoods. Since EPIC-KITCHENS has both new videos and old videos, and we can distinguish them by their narration id. As shown in Figure 3.6, our automatic annotation scheme is as follows: (a) One scene: if there is no scene change among these video clips, we apply this annotation to all video clips inside this verb-noun-participant pair. (b) Two scenes without boundary: use annotation of the later scene as a boundary, video clips earlier than it is annotated with annotation 1, those later than it is annotated with annotation 2. Video clips are divided into two groups based on their relative position to the boundary. Videos of each group are annotated with the annotation inside their group. (c) Two scenes with boundary: Video clips are divided into two groups based on their relative position to the boundary. Videos of each group are annotated with the annotation inside their group.

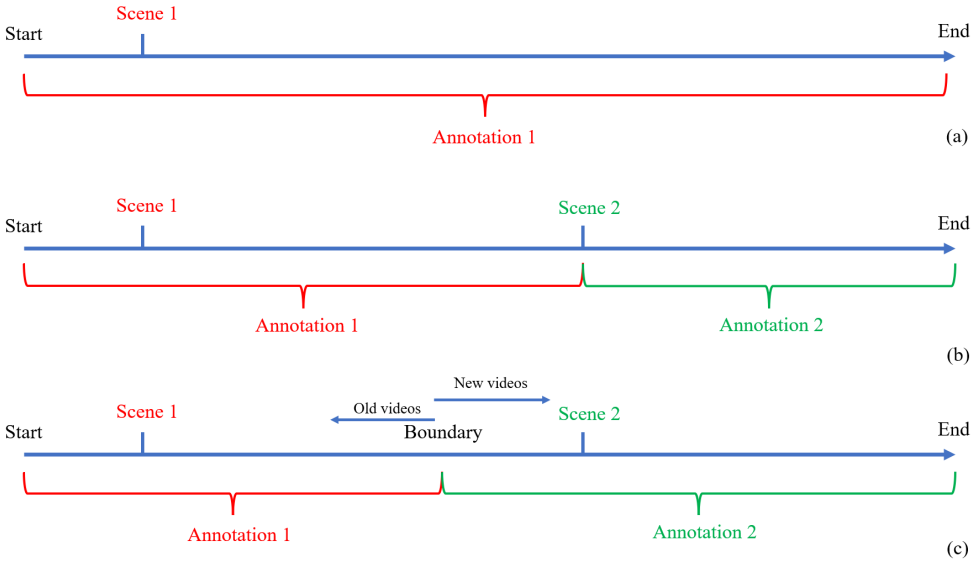

Fig. 3.6: Three cases in automatic affordance annotation of the EPIC-KITCHENS: (a) One scene: if there is no scene change among these video clips, we apply the manual annotation to all video clips inside this verb-noun-participant pair. (b) Two scenes without boundary: If there are two different manual affordance annotations but no old-new video boundary. We use later scene as a boundary, video clips earlier than it are annotated with annotation 1, those later than it are annotated with annotation 2. (c) Two scenes with boundary: If there are two different manual affordance annotations, locating on both sides of the old-new video boundary. Video clips are divided into two groups based on their happened time, videos of each group are annotated with the annotation inside their group.

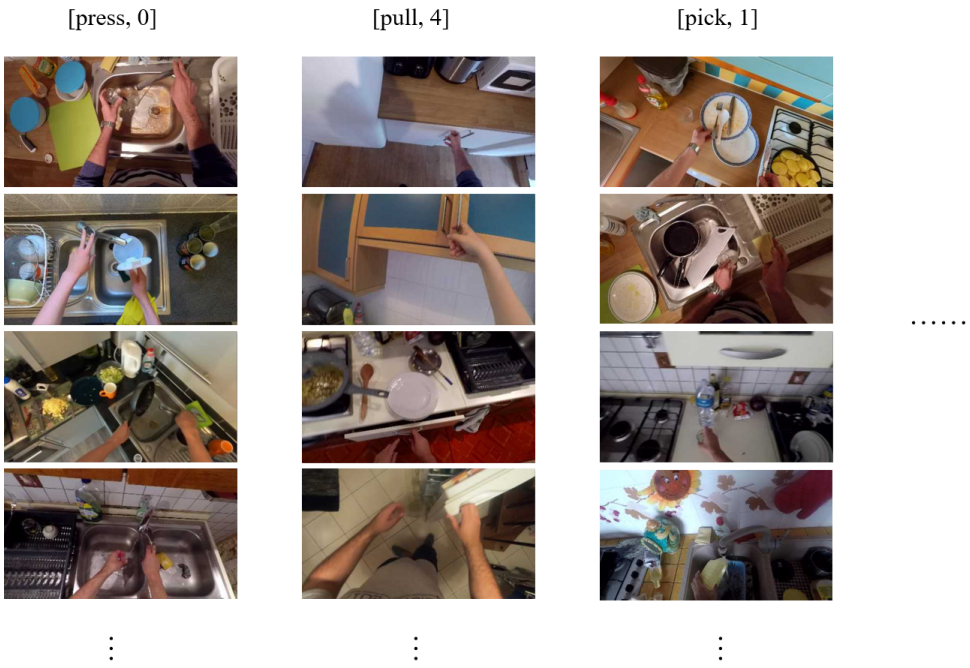

Fig. 3.7: In this figure, each column shows one example affordance annotation instance for the EPIC-KITCHENS dataset. Each column lists video clips annotated with the affordance label. First column: video clips annotated with the affordance [press, 0]. Original action labels of them are all turn off-tap. Second column: video clips annotated with the affordance [pull, 4]. Original action labels of them are: open-cupboard, open-cupboard, open-drawer, and open-fridge. Third column: video clips annotated with the affordance [pick, 1]. Original action labels of them are: pick up-plate, pick up-glass, pick up-glass, and pick-up bowl.

After annotation, we sampled 1,000 video clips from those not chosen in the first sampling phase. Then, we manually checked the accuracy of this automatic annotation scheme. As a result, we have annotated 31,924 video clips by manually annotating 300 verb-noun pairs, getting an accuracy of 96.76%. Some examples of the affordance annotation are shown in Figure 3.7.

3.3 HOMAGE

In this section, we will introduce the process of annotating the HOMAGE dataset with our affordance annotation scheme and show the annotation results. The Home Action Genome dataset [22] (HOMEAGE) is a recently published large-scale multi-view video dataset of daily activities at home. It contains 24.6k action segments annotated with 453 classes of atomic action labels. It was recorded by 36 participants in a limited number of indoor scenes. Similarly, with the EPIC-KITCHENS dataset, we also leverage the existing labels for constructing our new affordance annotation.

3.3.1 Tool-Use/Non-Tool-Use Action Annotation

We annotate tool-use/non-tool-use actions for HOMAGE dataset based on its action label list, and the results are shown in 3.2:

Table. 3.2: Tool-use/non-tool-use action annotation for HOMAGE

Category	Video Clips	Action labels
Non-tool-use	21.3k	hold, take, put, open, close, throw, fold, tear, turn, do, type, lie, read, squeeze, twisting, zip, press, flush
Tool-use	2.1k	wipe, put/pour, eat, brush, cut, iron, apply, write, sew, pour, spray, blow, dry, mop, drink, shave

Since the proportion of tool-use action video clips is too small in this dataset, we didn't annotate mechanical action for the HOMAGE dataset.

3.3.2 Affordance Annotation

Unlike the EPIC-KITCHENS dataset, HOMAGE asked the participants to perform specific activities inside different rooms. The annotation of room numbers helps us reduce the variation between different scenes, facilitating the annotation speed and accuracy. Following our proposed annotation scheme, we first find out the top 100 verb-noun pairs (action classes in this dataset). Then we randomly sample 5 video clips for every participant inside every room for each action class. Next, we manually annotate the affordance label for each action-participant-room pair, and finally we automatically assign these affordances labels to their neighborhoods.

We annotate 14,689 video clips automatically by manually annotating 100 action classes. We also sampled 500 instances from automatically annotated video clips and manually checked their affordance annotation. As a result, we got an accuracy of 92.81%. Some examples of the affordance annotation are shown in Figure 3.8.

3.4 Annotation Tools

As for annotation tools, we use CVAT [34]. CVAT is an online, interactive video and image annotation tool, which is also easy to deploy. Once we deployed it on the server, it could be accessed through the browser everywhere and do the annotation job. The whole process of using CVAT in our affordance annotation scheme is as follows. Take EPIC-KITCHENS as an example, we first sample video clips from the original annotation of the dataset according to the verb-noun-participant pair. For example, "P01_turn-off–tap" denotes video clips whose action label is "turn-off", object label is "tap", and recorded by participant 1. After getting the list of these sampled clips, we use the command-line tools provided by CVAT to generate annotation tasks automatically. By the way, the original dataset was on the server, so we don't need to upload video clips. Then we could check annotation tasks related to any verb-noun-participant pairs by searching keywords. As shown in Figure 3.9, the gallery picture of each task shows the scene change inside each

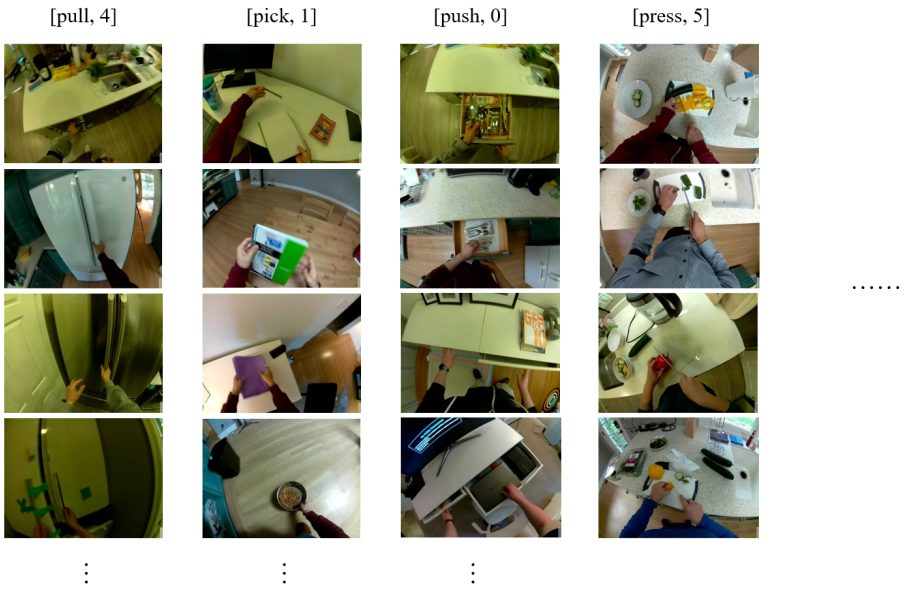

Fig. 3.8: Affordance annotation instances for the HOMAGE dataset. Each column lists video clips annotated with the affordance label. First column: video clips annotated with the affordance [pull, 4]. Original action labels of them are: open-drawer, open-fridge, open-fridge, and open cabinet. Second column: video clips annotated with the affordance [pick, 1]. Original action labels of them are: take-paper, take-book, take-paper, and take-bowl. Third column: video clips annotated with the affordance [push, 0]. Original action labels of them are all close drawer. Fourth column: videos clips annotated with the affordance [press, 5]. Original action labels of them are all cut-something.

pair. In this case, video number "P01_02_122" differs from other tasks. We need to annotate two affordance labels, one for it and one for other tasks in another scene.

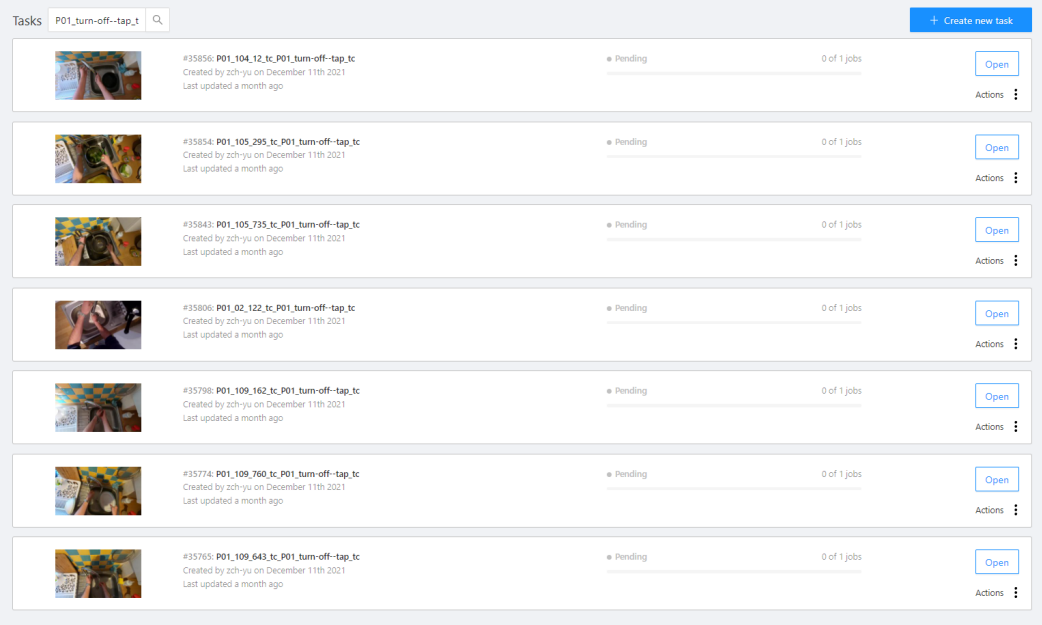

Fig. 3.9: The task list interface of the CVAT annotation tool. We first search related annotation tasks by searching the "verb-noun-participant". Then we check if there are different scenes inside the results. Finally, we manually annotate affordance labels for one video clip each scene.

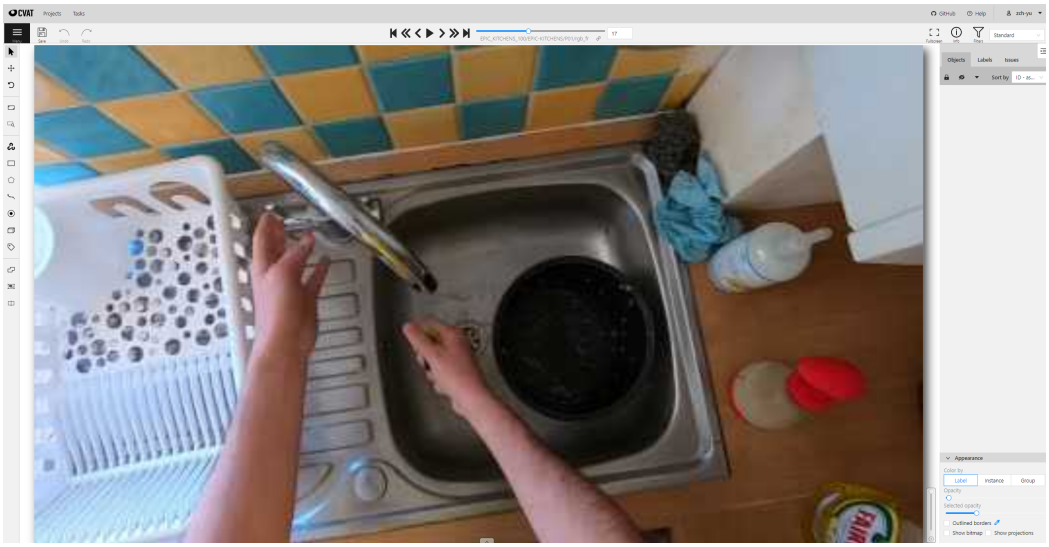

Fig. 3.10: The annotation interface of the CVAT annotation tool. We watch the video clip here first, and then annotate affordance label for the video.

During annotation, the annotator first watches the whole video clip through 3.10, and then annotates the goal irrelevant action label and grasp type label for it. We have thought about training a grasp type classification model to automatically annotate the grasp types to reduce annotation burden. The EPIC-KITCHENS dataset provides hand bounding boxes generated using Shan's method [35]. We first trained a sample grasp type classifier on the GNU-71 [36] dataset. This dataset used a 73-class grasp type taxonomy [37] which is an extension of the 33-class taxonomy we used in our method. We only use images within these 33 classes for training and narrowed grasp types into 6 as shown in Figure 3.3. Finally, we got an accuracy of 59.79% on the validation set. Due to the unacceptable classification accuracy and the semantic meaning of grasp types we discussed in Section 3.1.3, we choose to manually annotate grasp types.

Chapter 4

Experiments

In this chapter, we test our affordance annotation for the EPIC-KITCHENS dataset on various benchmark tasks to evaluate the rationality of our proposed efficient, precise affordance annotation scheme. We present quantitative and qualitative results for tool-use/non-tool-use action classification, mechanical action recognition, affordance recognition, and grasp type recognition. The experimental results show that a regular video understanding model could learn the affordance annotation labeled by our methods.

4.1 Benchmark Tasks

4.1.1 Tool-Use/Non-Tool-Use Action Classification Task

To evaluate the rationality of our proposed tool-use/non-tool-use action annotation method illustrated in 3.1.1, we trained an action recognition model using the EPIC-KITCHENS dataset with our tool-use/non-tool-use action annotation. The dataset is partitioned into the training set and validation set based on the train/val split of the EPIC-KITCHENS, as shown in Table 4.1. We also trained another model using annotations by chance to compare our annotation.

Table. 4.1: The number of video clips for train/validation

Label	Tool-use actions	Non-tool-use actions
#Train	8.5k	51.5k
#Validation	0.4k	2.5k

We use a slowfast [38] model as the action recognition model. The backbone is ResNet3d-50. Here the speed ratio is $\alpha = 4$, channel ratio is $\beta = 1/8$, and $\tau = 4$. Non-degeerate temporal filters are underlined.

We use stochastic gradient decent [39] optimizer with an initial learning rate of 0.1, momentum 0.9, and weight decay of 1e-4. Data augmentation is applied during the training phase. The input videos have been randomly resized, cropped, resized to 224×224, and flipped. After that, they are normalized with their average pixel value and standard deviation.

This model is trained on both our tool-use/non-tool use action annotations and random annotations. We do label balancing when training on our annotation since the label imbalance between tool-use and non-tool use is serious. The result is shown in Table 4.2, we could tell that the model learns to recognize tool-use/non-tool-use actions well with the supervision of our annotation.

Table. 4.2: Tool-use/non-tool use action classification results

Dataset	Tool-use actions(Acc)	Non-tool-use actions(Acc)
By chance	0.4720	0.5282
Ours	0.8580	0.7867

Figure 4.1 shows the visualization results generated by GradCam [40]. The successful cases of tool-use actions and failure cases of non-tool-use actions indicate that the model has learned the key to recognizing tool-use actions – interactions between tool and object. Failure cases of tool-use actions and successful cases of

non-tool-use actions illustrated this result from the opposite side.

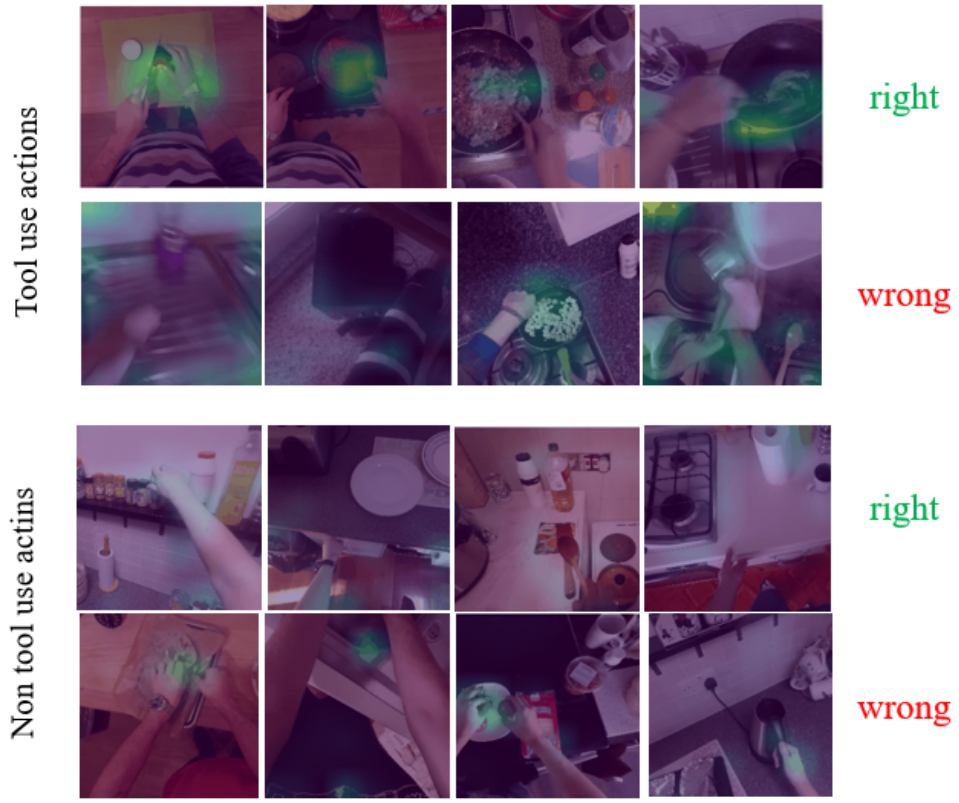

Fig. 4.1: Visualization results of tool-use/non-tool use classification model, the first row and third row are successful cases for recognizing tool-use-action and non-tool-use action, separately. The second row and fourth row are failure cases of for recognizing tool-use-action and non-tool-use action.

4.1.2 Mechanical Action Recognition

To test our annotation scheme for mechanical action, we test it with a mechanical action recognition task. 33 original action labels of tool-use action video clips are used as the mechanical action label of the EPIC-KITCHENS dataset. The data distribution of mechanical actions is shown in Figure 4.2 The dataset is par-

titioned into training set and validation set based on the train/val split of the EPIC-KITCHENS, as shown in Table 4.3. We use the slowfast action recognition model, which has the same structure as the one used in the tool-use/non-tool-use action classification task. We change the class numbers into 33, and use cross-entropy loss instead of binary cross-entropy loss.

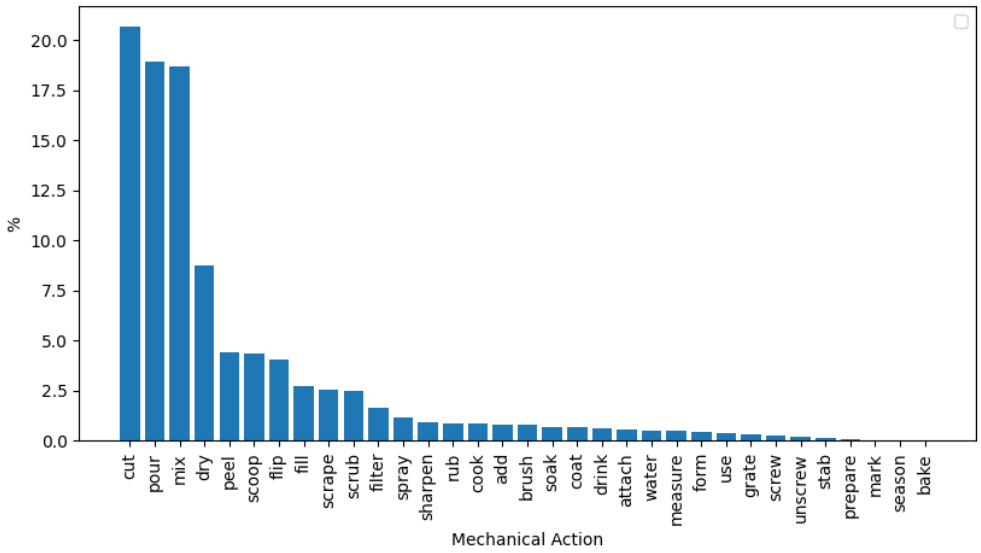

Fig. 4.2: Data distribution of the 33 mechanical actions of the EPIC-KITCHENS dataset

Table. 4.3: The number of video clips for train/validation of mechanical action recognition

	#Train	#validation
Mechanical action	8,425	1,260

Table 4.4 shows the results. Although tool-use action video clips are much less than non-tool-use action video clips, the results shows that regular action recognition model could learn to discriminate different mechanical actions from labels annotated with our method.

Table. 4.4: Mechanical action recognition results

	Top1 Acc	Top5 Acc
Mechanical action	0.5190	0.8643

4.1.3 Affordance Recognition

Affordance labels in our proposed affordance annotation scheme consist of grasp type and goal irrelevant actions. It is intuitive that this kind of affordance label fits the definition of affordance and discriminates affordance with other concepts. But can they easily be recognized by action recognition models? We trained a slowfast action recognition model to estimate affordance for given video clips. The distribution of affordance labels is shown in Figure 4.3.

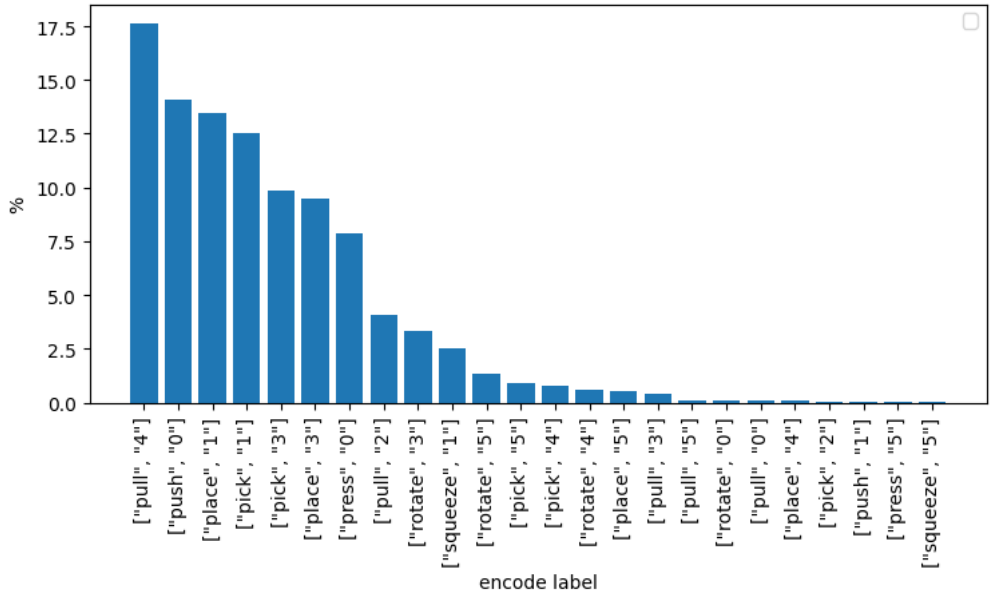

Fig. 4.3: Data distribution of the 24 affordance labels of the EPIC-KITCHENS dataset

Table 4.5 shows the quantitative results, and this indicates that regular action recognition methods are not suitable for our affordance recognition task. The rea-

son of this problem could be: Contextual information and semantic information played an important role during the annotation, the appearance has been less considered. This also reviews the essential difference between affordance and action.

Table. 4.5: Affordance recognition results

	Top1 Acc	Top5 Acc
Affordance	0.3107	0.6153

The visualization results of affordance recognition, mechanical action recognition, and tool-use/non-tool-use action classification are shown in Figure 4.4. From the first row, we could find out that the affordance recognition model focuses more on hands than objects. The second row shows that the mechanical action recognition model cares more about the interaction between objects. Let's make a comparison between the second row and the third row. It is clear that the mechanical action model focuses on tool object interactions, and the tool-use/non-tool-use action classification model focuses on the existence of tools. These results confirm the definition of affordance and mechanical action and evaluate the rationality of our proposed method.

4.1.4 Grasp Type Recognition

The data distribution of grasp types is shown in Figure 4.5. Instead of using the whole affordance label as annotation, we test our dataset on the grasp type recognition task. Given a video clip, the model will predict the grasp type. We still use the slowfast model, and the result is shown in Table 4.6.

Table. 4.6: Grasp type recognition results

	Top1 Acc	-
Grasp Type	0.5986	-

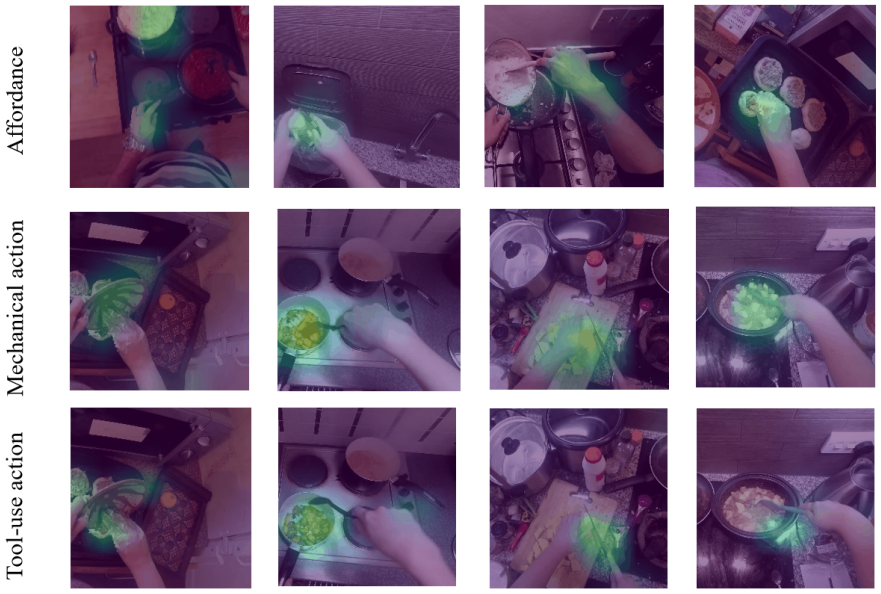

Fig. 4.4: Visualization results of affordance recognition, mechanical recognition, and tool-use/non-tool-use action recognition. The first row shows that the affordance recognition model focuses more on hands than objects. The second row shows that the mechanical action recognition model cares more about the interaction between objects. and the third row shows that tool-use/non-tool-use action classification model focuses on the existence of tools.

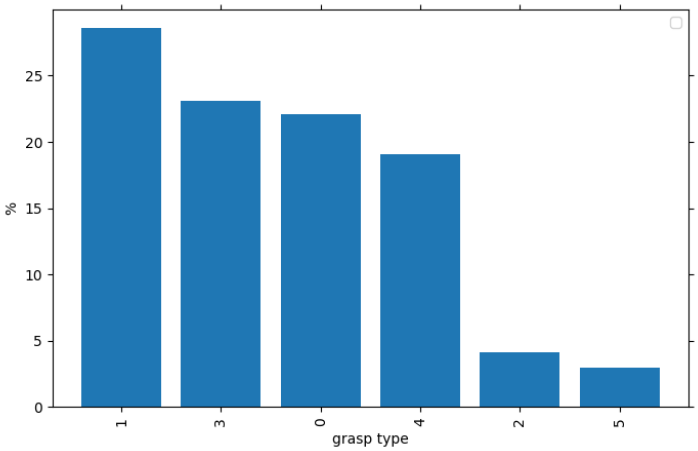

Fig. 4.5: Data distribution of grasp types

Chapter 5

Conclusion

Many works have investigated affordance in the computer vision field because of its capability of incorporating important contextual information. But previous research, especially affordance datasets, failed to give affordance an accurate definition, which is important to distinguish affordance from other concepts such as actions and object functions. We need an affordance annotation scheme for egocentric video datasets with a precise definition of affordance to address this issue. Besides, the affordance annotation scheme must be efficient because manually annotating affordance for large-scale action video datasets could be laborious.

This thesis introduced a precise affordance annotation scheme for egocentric videos. Firstly, we successfully distinguish affordance from other concepts by introducing the 3AS, including tool/non-tool use action, mechanical action, affordance. By combining grasp type and goal irrelevant actions, we proposed an affordance label form that can represent the relationship between the human's ability and the object's property. Secondly, we proposed a semi-automatic affordance annotation scheme for egocentric action video datasets recorded by limited participants inside limited scenes. The automatic annotation is based on the human's nature of consistently interacting with the same object in the same way. With this scheme, we could annotate affordance for a large number of video clips with less manual effort. Then we successfully applied our proposed annotation scheme to two large-scale datasets: the EPIC-KITCHENS dataset and the HOMAGE dataset. As a result, we got 46,613 videos annotated with affordance labels.

To evaluate the rationality of our proposed method, we test our affordance annotation of the EPIC-KITCHENS dataset on four benchmark tasks: tool-use/non-tool-use action classification, mechanical action recognition, affordance recognition, and grasp type recognition. Both quantitative and qualitative results show that our proposed method could clearly distinguish affordance from other concepts. Furthermore, the performance distance between the first two tasks and the affordance recognition task using the same action recognition model shows the giant gap between action and affordance. Leading to future research direction of affordance recognition models.

Future Work

As discussed in Section 4.1.3, the existing action recognition model is not suitable for affordance recognition tasks. This suggests that we need a novel model for affordance recognition. In our affordance annotation scheme, affordances are relationships between hand and object, and mechanical actions are relationships between tool and object. This inspired me that, to investigate affordance, we shouldn't pay much attention to appearance. We must focus on these contextual relationships. In the future, we could also investigate models good at modeling relationships.

It is also important to utilize affordance in other tasks such as action anticipation, gaze prediction, scene understanding, and so on. Using affordance as contextual information of the environment and assistant other tasks.

Furthermore, it would be interesting if we could transfer our affordance knowledge to robots. The main challenge is how to measure the difference in agents' abilities.

References

[1] F. Osiurak, Y. Rossetti, and A. Badets, "What is an affordance? 40 years later," *Neuroscience & Biobehavioral Reviews*, vol. 77, pp. 403–417, 2017.

[2] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Object-based affordances detection with convolutional neural networks and dense conditional random fields," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5908–5915.

[3] A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos, "Affordance detection of tool parts from geometric features," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 1374–1381.

[4] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *The International Journal of Robotics Research*, vol. 32, no. 8, pp. 951–970, 2013.

[5] S. Thermos, G. T. Papadopoulos, P. Daras, and G. Potamianos, "Deep affordance-grounded sensorimotor object recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6167–6175.

[6] K. Fang, T.-L. Wu, D. Yang, S. Savarese, and J. J. Lim, "Demo2vec: Reasoning object affordances from online videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2139–2147.

[7] T. Nagarajan, C. Feichtenhofer, and K. Grauman, "Grounded human-object

interaction hotspots from video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8688–8697.

[8] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The grasp taxonomy of human grasp types," *IEEE Transactions on human-machine systems*, vol. 46, no. 1, pp. 66–77, 2015.

[9] Y. Yamani, A. Ariga, and Y. Yamada, "Object affordances potentiate responses but do not guide attentional prioritization," *Frontiers in integrative neuroscience*, vol. 9, p. 74, 2016.

[10] J. J. Gibson, "The concept of affordances," *Perceiving, acting, and knowing*, vol. 1, 1977.

[11] D. A. Norman, *The psychology of everyday things.* Basic books, 1988.

[12] S. A. Linkenauger, J. K. Witt, J. K. Stefanucci, J. Z. Bakdash, and D. R. Proffitt, "The effects of handedness and reachability on perceived distance." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 35, no. 6, p. 1649, 2009.

[13] J. C. Phillips and R. Ward, "Sr correspondence effects of irrelevant visual affordance: Time course and specificity of response activation," *Visual cognition*, vol. 9, no. 4-5, pp. 540–558, 2002.

[14] M. J. Riddoch, G. W. Humphreys, S. Edwards, T. Baker, and K. Willson, "Seeing the action: Neuropsychological evidence for action-based effects on object selection," *Nature neuroscience*, vol. 6, no. 1, pp. 82–89, 2003.

[15] J. Grèzes, M. Tucker, J. Armony, R. Ellis, and R. E. Passingham, "Objects automatically potentiate action: an fmri study of implicit processing," *European Journal of Neuroscience*, vol. 17, no. 12, pp. 2735–2740, 2003.

[16] R. Azuma, T. Takiguchi, and Y. Ariki, "Estimation of Object Functions Using Visual Attention," p. 4, 2018.

[17] T. Luddecke and F. Worgotter, "Learning to segment affordances," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 769–776.

[18] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 14–29, 2015.

[19] M. Liu, S. Tang, Y. Li, and J. M. Rehg, "Forecasting human-object interaction: joint prediction of motor attention and actions in first person video," in *European Conference on Computer Vision*. Springer, 2020, pp. 704–721.

[20] T. Nagarajan, Y. Li, C. Feichtenhofer, and K. Grauman, "Ego-topo: Environment affordances from egocentric video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 163–172.

[21] D. Damen, H. Doughty, G. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "The epic-kitchens dataset: Collection, challenges and baselines," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 01, pp. 1–1, 2020.

[22] N. Rai, H. Chen, J. Ji, R. Desai, K. Kozuka, S. Ishizaka, E. Adeli, and J. C. Niebles, "Home action genome: Cooperative compositional action understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 184–11 193.

[23] G. Young, "Are different affordances subserved by different neural pathways?" *Brain and cognition*, vol. 62, no. 2, pp. 134–142, 2006.

[24] G. Humphreys, "Objects, affordances⋯ action!" *The Psychologist*, 2001.

[25] A. M. Borghi and L. Riggio, "Stable and variable affordances are both automatic and flexible," *Frontiers in human neuroscience*, vol. 9, p. 351, 2015.

[26] R. Ellis and M. Tucker, "Micro-affordance: The potentiation of components of action by seen objects," *British journal of psychology*, vol. 91, no. 4, pp. 451–471, 2000.

[27] R. Ellis, M. Tucker, E. Symes, and L. Vainio, "Does selecting one visual object from several require inhibition of the actions associated with nonselected objects?" *Journal of Experimental Psychology: Human Perception and Performance*, vol. 33, no. 3, p. 670, 2007.

[28] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2019.

[29] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5842–5850.

[30] L. Ji, B. Shi, X. Guo, and X. Chen, "Functionality Discovery and Prediction of Physical Objects," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, pp. 123–130, Apr. 2020. [Online]. Available: https://www.aaai.org/ojs/index.php/AAAI/article/view/5342

[31] A. Pieropan, C. H. Ek, and H. Kjellström, "Recognizing object affordances in terms of spatio-temporal object-object relationships," in *2014 IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2014, pp. 52–58.

[32] S. Thermos, P. Daras, and G. Potamianos, "A deep learning approach to object affordance segmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2358–2362.

[33] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, "Learning visual affordance grounding from demonstration videos," *arXiv preprint arXiv:2108.05675*, 2021.

[34] B. Sekachev, N. Manovich, M. Zhiltsov, A. Zhavoronkov, D. Kalinin, B. Hoff, TOsmanov, D. Kruchinin, A. Zankevich, DmitriySidnev, M. Markelov, Johannes222, M. Chenuet, a andre, telenachos, A. Melnikov, J. Kim, L. Ilouz,

N. Glazov, Priya4607, R. Tehrani, S. Jeong, V. Skubriev, S. Yonekura, vugia truong, zliang7, lizhming, and T. Truong, "opencv/cvat: v1.1.0," Aug. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4009388

[35] D. Shan, J. Geng, M. Shu, and D. F. Fouhey, "Understanding human hands in contact at internet scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9869–9878.

[36] G. Rogez, J. S. Supancic, and D. Ramanan, "Understanding everyday hands in action from rgb-d images," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3889–3897.

[37] J. Liu, F. Feng, Y. C. Nakamura, and N. S. Pollard, "A taxonomy of everyday grasps in action," in *IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2014, pp. 573–580.

[38] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[39] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International conference on machine learning*. PMLR, 2013, pp. 1139–1147.

[40] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.