HAKE: Human Activity Knowledge Engine
Brief Report

Yong-Lu Li, Liang Xu, Xijie Huang, Xinpeng Liu, Ze Ma, Mingyang Chen,
Shiyi Wang, Hao-Shu Fang, Cewu Lu*

Shanghai Jiao Tong University
http://hake-mvig.cn
yonglu_li@sjtu.edu.com liangxu@sjtu.edu.cn huangxijie1108@gmail.com xinpengliu0907@gmail.com
maze1234556@sjtu.edu.cn cmy_123@sjtu.edu.cn Shiy.Wang@outlook.com fhaoshu@gmail.com
lucewu@sjtu.edu.cn

Abstract

*Human activity understanding is crucial for building automatic intelligent system. With the help of deep learning, activity understanding has made huge progress recently. But some challenges such as imbalanced data distribution, action ambiguity, complex visual patterns still remain. To address these and promote the activity understanding, we build a large-scale Human Activity Knowledge Engine (HAKE) based on the human body part states. Upon existing activity datasets, we annotate the part states of all the active persons in all images, thus establish the relationship between instance activity and body part states. Furthermore, we propose a HAKE based part state recognition model with a knowledge extractor named Activity2Vec and a corresponding part state based reasoning network. With HAKE, our method can alleviate the learning difficulty brought by the long-tail data distribution, and bring in interpretability. Now our HAKE has more than **7 M+** part state annotations and is still under construction. We first validate our approach on a part of HAKE in this preliminary paper, where we show **7.2** mAP performance improvement on Human-Object Interaction recognition, and **12.38** mAP improvement on the one-shot subsets.*

1. Introduction

Human activity understanding is an active topic in computer vision and has a large number of potential applications and business prospects. Facilitated by the growth of image and video data and the renaissance of Deep Neural Networks (DNNs), lots of works have been proposed to forward this direction. Activity recognition has strong relations with other research contents of computer vision, such as object detection[20], pose estimation [15], video analysis [11], visual relationship [18]. Recent works on activity and action recognition almost rely on the end-to-end supervised paradigm to address this high-level cognition task, *i.e.* perception from raw pixels directly to the activity classes in one stage. This paradigm shows poor performance on large-scale activity benchmarks, such as HICO [13], HICO-DET [12], AVA [10].

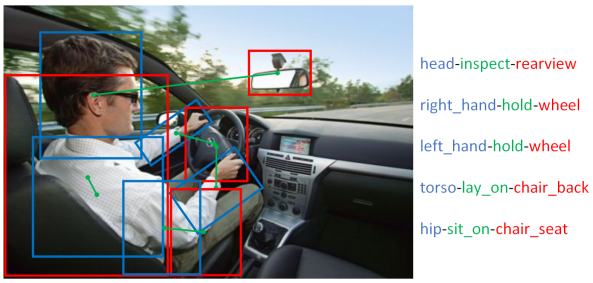


Figure 1. Body part state samples. A driving scene contains activity $\langle person, drive, car \rangle$. It can be decomposed into various body part states like $\langle head, inspect, rearview \rangle$, $\langle right\_hand, hold, wheel \rangle$, $\langle left\_hand, hold, wheel \rangle$.

The limited performance of present one-stage paradigm on these large-scale and exceedingly difficult datasets are possibly due to the own difficulties of activity understanding. For instance, activity recognition has many challenges such as long-tail data distribution, variability and complexity of action visual patterns, crowd background in daily scenes, various camera viewpoints and motions, occlusion

Draft, work in progress. *Cewu Lu is the corresponding author.

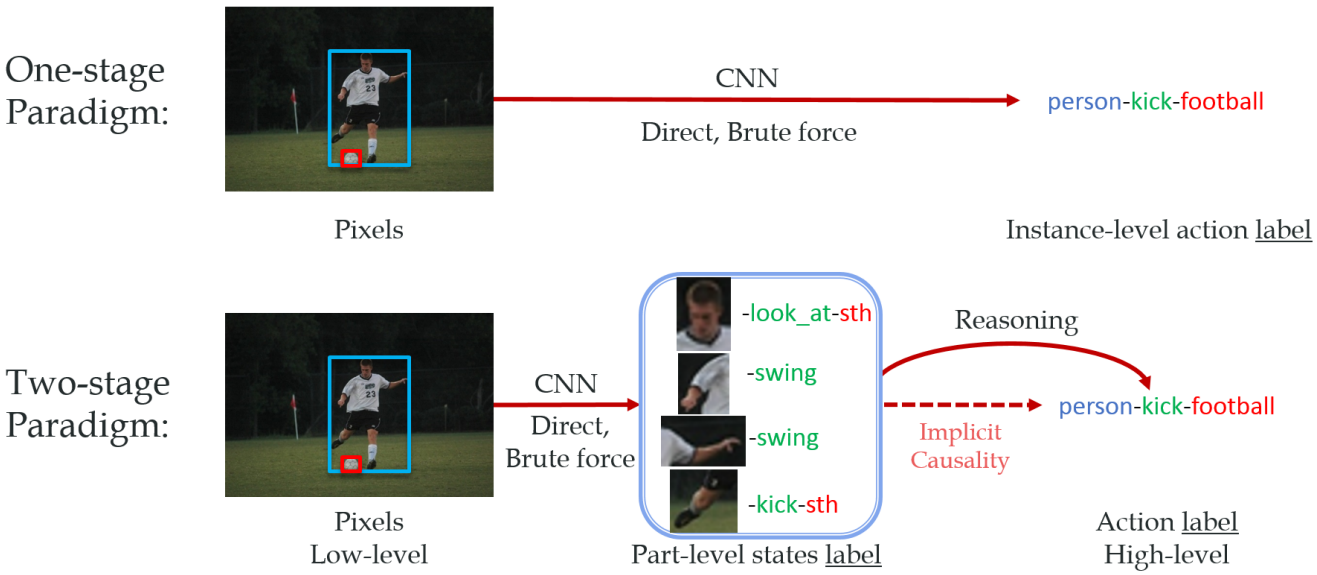


Figure 2. Previous one-stage paradigm and our hierarchical two-stage paradigm.

and self-occlusion, crowd-sourced annotations and data. In the absence of data, one-stage paradigm which needs to bridge the huge gap is powerless. To this end, we propose a new Human Activity Knowledge Engine (HAKE) based on body part states [9]. Based on HAKE, a new corresponding hierarchical two-stage paradigm for activity recognition is also presented.

Different from the one-stage paradigm, we divide the activity understanding into two phases: 1. Part states [9] recognition from the visual patterns, and the activity representation by combining visual and linguistic knowledge; 2. Reasoning the activities from part states, as seen in Fig. 2. Part states mean the finer level atomic body part actions [9] which compose the action of human instance. Fig. 1 shows an example of instance activity and its corresponding part states. Based on the reductionism [8], our assumption is that: the human instance action consists of the atomic actions or states of all the body parts. Thus we can divide the instance action recognition into two sub-tasks: body part state recognition and the recombination of part states.

The most obvious advantages of our hierarchical paradigm are three-fords. First, part states are the basic components of instance actions, their relationship can be in analogy with the amino acid and protein, letter and word. Different instance actions like "person hold an apple" and "person eats an apple" share the same part states "hand holds something" and "head looks at something". Thus the imbalanced data problem will be greatly alleviated, for that the samples per category largely increases on the same data scale. For supervised learning, it will effectively reduce the learning difficulty. Furthermore, part state recognition is much easier than the instance action recognition because of fewer categories and simpler visual patterns. In our experiment, a simple model consists of shallow CNNs and fully connected layers can achieve acceptable performance on part state recognition, which is generally relative 50% higher than the instance action recognition. Second, with part state recognition as the midpoint, the gap between the image space and the semantic space would be greatly narrowed. Third, we can obtain a more powerful representation of action patterns based on part states. In our experiment, the combinative visual-linguistic part state embeddings present obvious semantic meaning and better interpretation. When the model predicts what he/she is doing, we can easily know the reasons: what his/her body parts are doing.

The main contributions of this work are: 1. We construct a large-scale Human Activity Knowledge Engine named HAKE that bridges the relationship between instance activity and body part states. We will keep on enlarging and enriching it, and call on the community to help us make it more powerful to promote activity understanding. 2. A new hierarchical paradigm is proposed based on HAKE, which outperforms state-of-the-art methods on several activity recognition benchmarks. In particular, the performance of rare action categories on several benchmarks are significantly boosted.

2. Construction of HAKE

In this section, we will illustrate the construction of HAKE. Considering the complexity, we first construct part states annotations on still images, and then expand to the

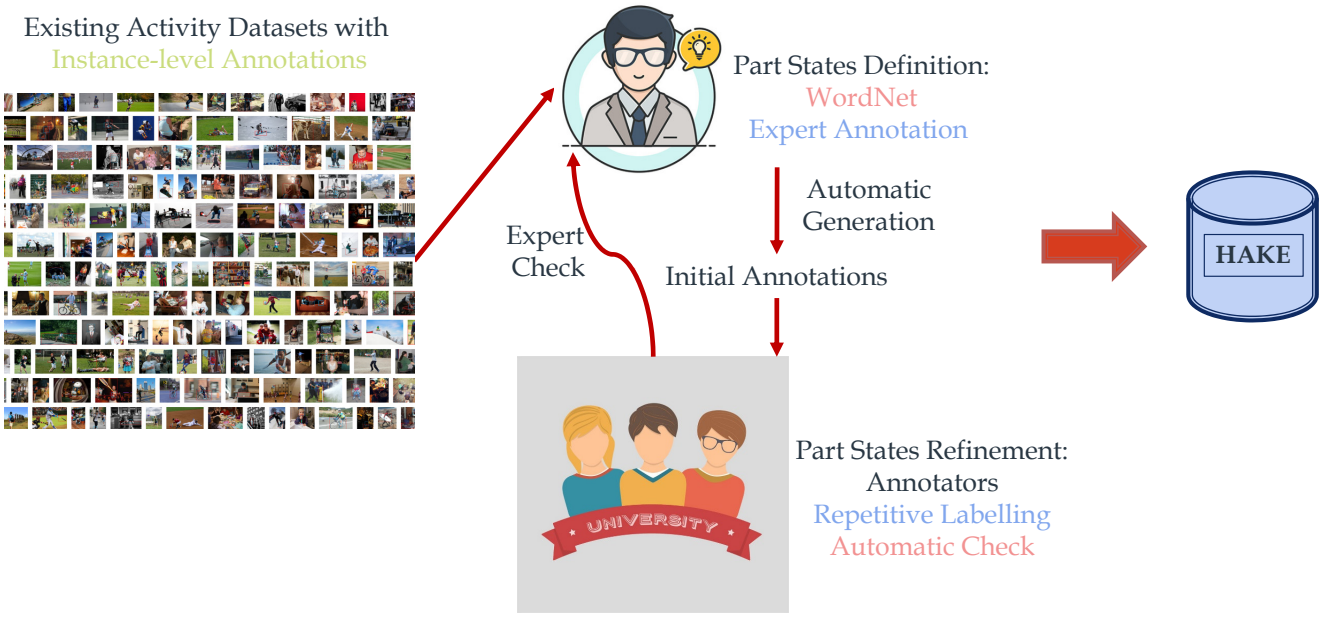


Figure 3. The construction of our HAKE.

consecutive frames of videos. The key characteristic of our HAKE are: the definition of part states are based on atomic and composite actions, crowd-sourced images from several widely-used activity datasets, realistic visual contexts, diversity and variability of activities.

Part States Definition. HAKE is based on the existing well-designed datasets, for example, HICO-DET [12], V-COCO [16], OpenImage [17], HCVRD [22], HICO [13], MPII [1], AVA [10], which are structured around a rich semantic ontology. The activity categories contained in HAKE are chosen according to the most common human daily actions/activities, social interactions with daily objects and person. We first select 154 instance activity categories from above datasets in the case of hierarchical activity structure [11]. All the part states will be annotated upon instance level activities. Then we decompose the human body into ten body parts following [14], namely head, arms, hands, hip, legs, feet. Third, we select about 200 part states based on the verbs from WordNet [2] as the candidates to build a part state pool, e.g. "hold", "push", "pick" for hands, "listen to", "eat", "talk to" for head and so on.

To ensure the quality of part state selection, we invite several experts to use their own understandings to depict the selected 154 instance actions in the body part level. For example, when we show an image with activity "person drive a car" to them, they may describe it as "hip sit on something", "hands hold something", "head look at something". Based on their choices, we use the Normalized Point-wise Mutual Information (NPMI) [3] to calculate the co-occurrence between the instance action categories and part state candidates. Finally, we choose 92 candidates with the highest NPMI values as the final part states.

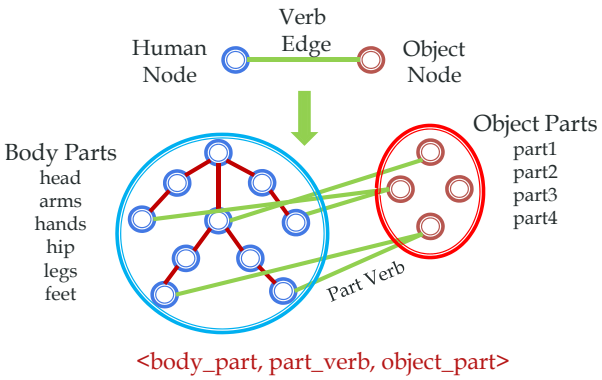


Figure 4. The graphic model of instance activity and part states.

Based on 154 instance activities and 92 part states, we can construct a hierarchical graph structure shown in Fig.4. Actions and part states are represented as the edges between the subject and object nodes (for the non-object actions, the edge is a loop).

Part State Annotation. We annotate all part states belonging to all the actions of all the active persons in all the collected images exhaustively. To be specific, existing datasets already have the human and object bounding box annotations and the relationship links between them. We then use pose estimation [15] to obtain the pose keypoints of all the annotated persons, and their part bounding boxes following [14]. We adopt a semiautomatic method to build HAKE.

First, we invite nine experts to annotate 10 thousands of images with all the 154 instance actions as the basis, and generate the initial part states for all the rest of images based on their annotation distribution. Thus the other annotators will use our tool to amend and refine the initial annotations according to their understanding of these actions. To ensure the quality, one instance with multiple actions would be annotated multiple times for each activity. Furthermore, each image will be checked at least twice by the automatic procedure and experts. We cluster these labels and discard the obvious outliers to obtain the robust label agreements.

It is worth noting that, action recognition is a multi-label classification problem, an active person may have more than one actions. For each instance action, we annotate its corresponding part states respectively and then combine all sets of part states in the final round. In other words, a body part can also have multiple states at the same time, e.g., activity "person cuts an apple" would have part states "right-hand holds a knife" and "right-hand uses something to cut something" at the same time.

At present, we have finished the annotations of 104 K+ images, which include 677 k+ human instances, 278 K+ interacted objects, 733 K+ instance actions, 7 M+ human body part states. In addition, our labeling is still in progress, and we have build a project page (http://hake-mvig.cn) and an online annotation tool. We hope the volunteers from all over the world to help us enlarge and enrich HAKE to advance the activity understanding. With these densely annotated part states, we believe in that we can deepen and promote the activity understanding significantly.

3. Hierarchical Paradigm

On the basis of our human activity knowledge engine, we can address the activity recognition in a hierarchical way: 1. Part State Recognition with knowledge extraction via Activity2Vec; 2. Reasoning from Part States to Instance Activity. This hierarchy would bring in more interpretability and a new possibility for the following-up researches.

3.1. Part State Recognition and Activity2Vec

In the first phase, we utilize the canonical pipeline to address the part state recognition. With the object detection of images, we can obtain the bounding boxes of all the detected person and objects. Second, we extract the ROI pooling features of all body parts as the input of the Part States Classification Network (PSC), as shown in Fig. 6. Within HAKE, we have annotated all the part states of all the human instances, thus we can construct part state classification loss for each part. As mentioned before, part state recognition is much easier in the supervised learning paradigm, which is also proven by our experiments.

To enhance the representation ability and promote the subsequent activity reasoning, we additionally utilize the

Figure 5. Samples of activity reasoning via part states. By combining the classified part states, we can reason out the activity from co-occurrence and prior knowledge. Inner relations between part states can also be helpful. For instance, holding an apple by hand and eating an apple with mouth often appear together. These two part states can derive the activity $\langle human, eat, apple \rangle$.

uncased BERT-Base pre-trained model [4] as the language feature extractor. Bert [4] is a language understanding model trained on large-scale word corpus, it can generate contextual embeddings for many types of downstream NLP tasks. Bert has considered the surrounding (left and right) of a word, and use a deep bidirectional Transformer to extract the general embeddings of words. Thus its pre-trained representations carry the contextual information from the enormous text data, *e.g.* Wikipedia. These features usually contain implicit semantic knowledge about the activity and part states, which is clearly different from visual information. In specific, we divide a part state into tokens, *e.g.* "head"(body part), "eat"(verb), "apple"(object). Each part state will be converted to a 2304 vector (three 768 vectors for the part, verb, object respectively). Second, we multiply the fixed language features with predicted part state probabilities from the part state classifiers, which can also be seen as an attention mechanism. A more possible part state will get larger attention.

Our goal is to bridge the gap between part states and instance activity. The combination of the visual and linguistic knowledge thus can be a powerful clue for establishing this mapping. We align the visual and linguistic features by using triplet loss [5], and concatenate them as the out-

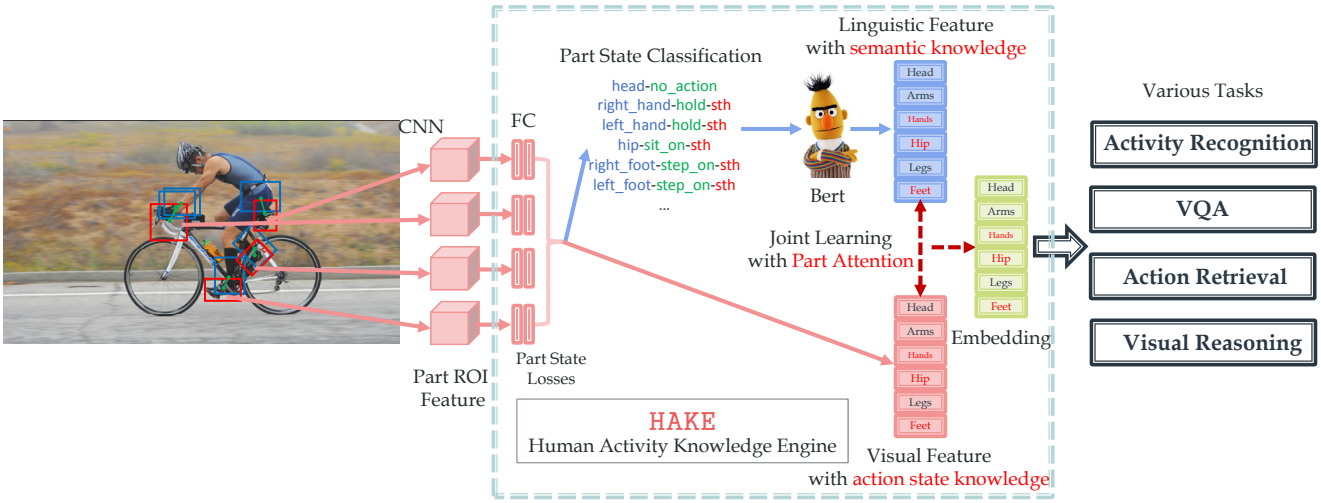


Figure 6. The overview of Part state recognition and Activity2Vec.

put, this process is called as Activity2Vec. (Seen in Fig. 6) The output embedding is 3584 sized and as the input of the downstream tasks, *e.g.* activity recognition, Visual Question Answering, action retrieval. Especially, before utilizing this embedding, we will first use the activity recognition task to pre-train it to capture the activity knowledge. From the experiment results, we can find that the embedding generated by our Activity2Vec can significantly improve the performance of multiple activity related benchmarks.

3.2. Reasoning from Part States to Instance Activity

With the embeddings from the Activity2Vec, we can better infer the relationship between part states and activities. If all the activities can be seen as the nodes at a higher level within a hierarchical graph, the part states will be the nodes in the lower level. Inner relationships between part state nodes can be seen as their co-occurrence, so do the activity nodes. The edges linked the instance activity and part states are the key elements in activity understanding, as seen in Fig. 7. We propose a Part States Reasoning Network (PSR) to estimate these cross-level edges between activity nodes and part state nodes.

More details of the proposed PSC and PSR models will be illustrated in our official version paper.

4. Experiments

4.1. An analogy: simplified Action Recognition

In this section, we design a simplified experiment to give a better intuition about our approach. We build a dataset derived from MNIST which consists of handwritten digits from 0 to 9 and with size $28 \times 28 \times 1$. To test our assumption, we generate a set of new $128 \times 128 \times 1$ images which is a combination of 3 to 5 handwritten digits randomly selected from MNIST and randomly distributed in images. And the corresponding label of each new image is the sum of the largest and second largest value of the digits within this image. Thus the total number of categories is 19 (0 to 18). This problem is a simplified analogy of human activity recognition, as shown in Fig. 8. We argue that actions can be decomposed into a set of part states, which highly resembles the relationship between the digits and the sum function in the above case. The random amount of digits and their distribution aim to simulate the randomness of part states. Taking the influence of background into consideration, we also add Gaussian noise on the whole generated image samples.

To compare the one-stage (instance based) paradigm and two-stage (part based) paradigm, we adopt a simple network to conduct a test. The network is composed of shallow sequential convolution layers and fully connected layers (shown in Fig.9).

Two paradigms are trained with the same optimizer, learning rate and epochs. The results are shown in Fig.10 and Tab. 1, which show the significant superiority of part based paradigm (174% relative increases on accuracy) over instance based paradigm. To some extent, this result supports our assumption about the decomposability of human instance activity and the effectiveness of part state knowledge representation.

4.2. Human-Object Interaction Recognition

To verify the effectiveness of our HAKE and hierarchical paradigm, we perform experiments on Human-Object Interaction (HOI) recognition task [13] in still images. HOI [12, 23, 21] usually accounts for a large proportion of daily life activities. And it is more complex than the non-HOI activity, *e.g.* "dance", "swim". In the initial version,

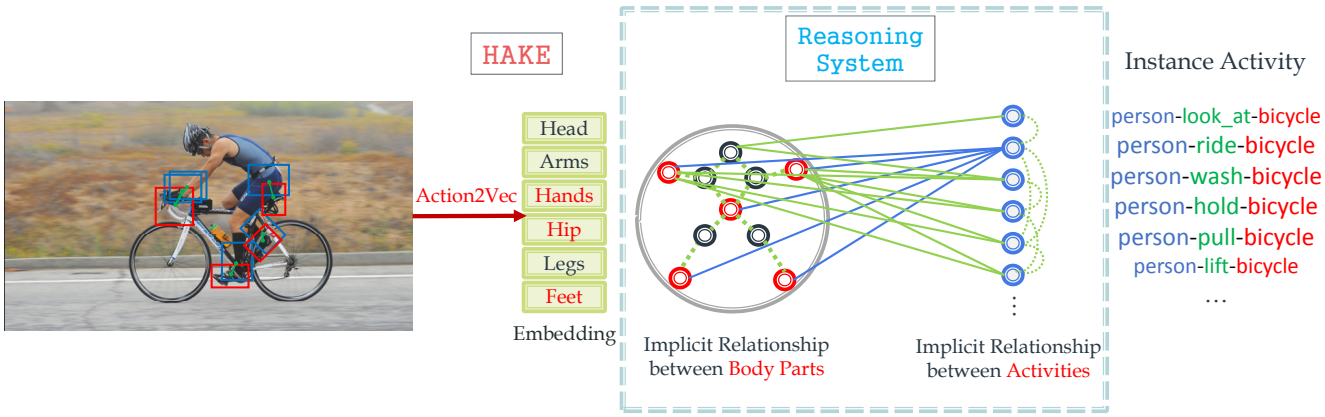


Figure 7. Reasoning from part states to instance activities.

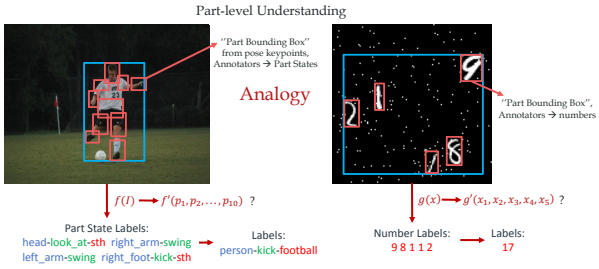


Figure 8. An analogy to activity recognition.

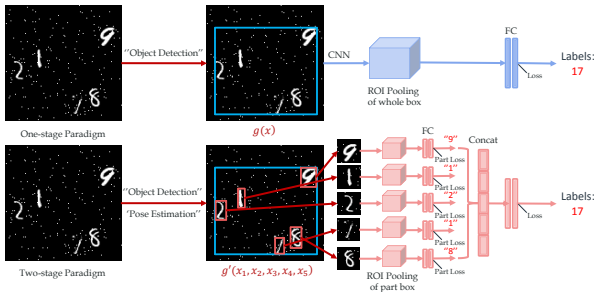


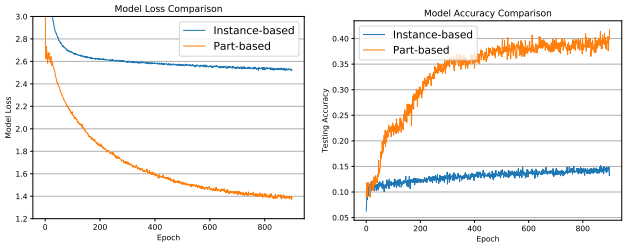
Figure 9. Instance based model and part based model.

we just report the results on HICO [13] to evaluate the improvements brought by HAKE, especially on one-shot and few-shot learning problems. More results on other activity understanding tasks will be reported in the official paper.

Method	Test Accuracy
Instance Based Paradigm	15.2
Part Based Paradigm	**41.7**

Table 1. Comparison of accuracy on our dataset

HICO [13] contains 38,116 images in train set and 9,658 images in test set. To be fair, we follow the experiment settings of [14] to compare the recognition performance.



(a) Loss (b) Test Accuracy

Figure 10. Comparison of loss and accuracy

Method	mAP
AlexNet+SVM [13]	19.4
R*CNN [7]	28.5
Girdhar—&Ramanan [6]	34.6
Mallya—&Lazebnik [19]	36.1
Pairwise-Part [14]	39.9
Pairwise-Part [14]+HAKE	**47.1**
Gain	**7.2**

Table 2. Comparison with previous methods on the HICO test set.

From Tab. 2 we can find that our method achieve 7.2 mAP gain over the state-of-the-art result on HICO [13]. Moreover, on the one-shot and few-shot sets (training images are less than 5 and 10) of HICO, our method can achieve more than 11 mAP improvement. Specific results can be seen in Tab. 3. These results show that our HAKE and HAKE-based hierarchical paradigm can significantly enhance the learning ability of model under few-shot circumstances.

Figure 11. Some predictions of our method. Triplets under images are predicted activities. Body part, part verb and object part are represented in blue, green and red, so are the activity results. Green tick means right prediction and red cross is the opposite.

Method	Few@1	Few@5	Few@10
Pairwise-Part [14]	13.02	19.79	22.28
Pairwise-Part [14]+HAKE	**25.40**	**32.48**	**33.71**
Gain	**12.38**	**12.69**	**11.43**

Table 3. Effectiveness on few-shot problems. Few@i represent the average mAP on few-shot activity sets. @i means the number of training images is less than i, if i is 1 then it means one-shot problem. On HICO [13], there is obvious positive correlation between performance and the quantity of training samples. Our approach can obviously improve the recognition effect on few-shot problem, for the reason of reusability and composability of part states.

Some qualitative results on HOI recognition are shown in Fig.11. The $\langle body\_part, part\_verb, object\_part \rangle$ with the highest scores are visualized in blue, green and red bounding boxes, and their corresponding labels are demonstrated under each image with colors consisted with boxes. The final predictions with the highest scores are represented too.

5. Conclusion

In this paper, we proposed a novel body part state knowledge base named Human Activity Knowledge Engine (HAKE) for human activity understanding, and a corresponding hierarchical paradigm. Our two-stage method consists of two components: Part State Recognition with Activity2Vec, and Part State based Reasoning. With HAKE, we can obtain a new embedding combined both visual and linguistic semantic knowledge, which brings the interpretability in the human activity recognition. Part states can significantly heighten the activity reasoning to alleviate the learning difficulties brought by the imbalanced data, and utilize the co-occurrence relation to bridge the semantic interspace between part states and activity. Our experiment results on HOI recognition show that HAKE can significantly improve the performance, especially under the few-shot circumstances.

6. Future Work

Considering that our HAKE is still under construction, we will keep on enriching and enlarging it to promote the research in human activity understanding. We will also use HAKE to promote other tasks related to human activity understanding, *e.g.* video-based activity understanding, action retrieval, Visual Question Answering and so on.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis In *CVPR*, 2014. 3

[2] G. A. Miller. WordNet: a lexical database for English. In *Communications of the ACM*, 38(11), 1995. 3

[3] K. W. Church, P. Hanks. Word association norms, mutual information, and lexicography. In *Computational linguistics*, 16(1), 1990. 3

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.04805*, 2018. 4

[5] F. Schroff, D. Kalenichenko, J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 4

[6] R. Girdhar, D. Ramanan. Attentional pooling for action recognition. In *NIPS*, 2017. 6

[7] G. Gkioxari, R. Girshick, J. Malik. Contextual action recognition with r* cnn. In *NIPS*, 2015. 6

[8] T. Honderich. The Oxford companion to philosophy. In *OUP Oxford*, 2005. 2

[9] C. Lu, H. Su, Y. Li, Y. Lu, L. Yi, C.-K. Tang, L. J. Guibas. Beyond holistic object recognition: Enriching image understanding with part states. In *CVPR*, 2018. 2

[10] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar and others. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 1, 3

[11] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 1, 3

[12] Y. W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human-object interactions. In *WACV*, 2018. 1, 3, 5

[13] Y. W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015. 1, 3, 5, 6, 7

[14] H.-S. Fang, J. Cao, Y.-W. Tai, and C. Lu. Pairwise body-part attention for recognizing human-object interactions. In *ECCV*, 2018. 3, 6, 7

[15] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 1, 3

[16] S. Gupta and J. Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 3

[17] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*, 2018. 3

[18] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016. 1

[19] A. Mallya and S. Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In *ECCV*, 2016. 6

[20] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1

[21] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y.-F. Wang, C. Lu. Transferable Interactiveness Prior for Human-Object Interaction Detection. In *arXiv preprint arXiv:1811.08264*, 2018. 5

[22] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. v. d. Hengel. Care about you: towards large-scale human-centric visual relationship detection. *arXiv preprint arXiv:1705.09892*, 2017. 3

[23] C. Gao, Y. Zou, and J.-B. Huang. iCAN: Instance-Centric Attention Network for Human-Object Interaction Detection. *arXiv preprint arXiv:1808.10437*, 2018. 5

Appendices

Some samples from the proposed Part State Library, each part state consists of a state description and a corresponding cropped part region:

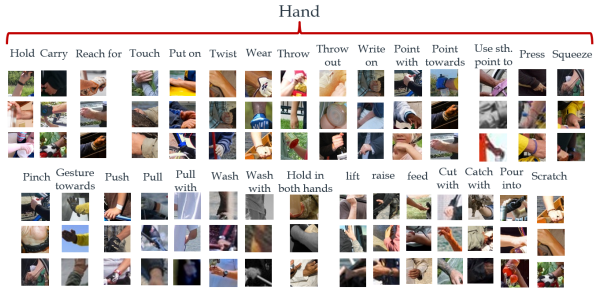


Figure 12. Some "hand" states.

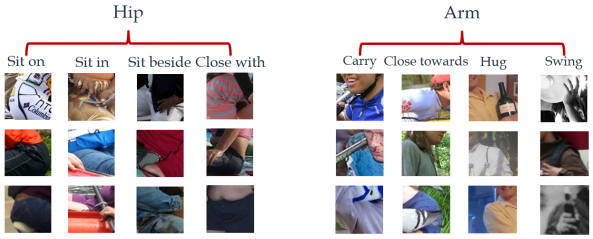


Figure 13. Some "hip" and "arm" states.

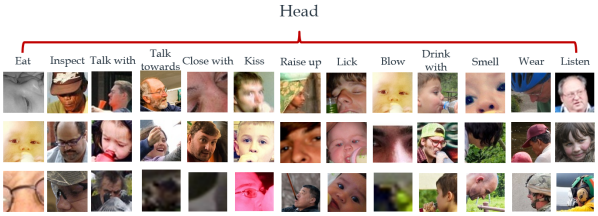


Figure 14. Some "head" states.

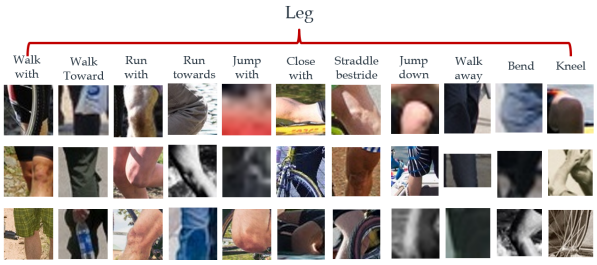


Figure 15. Some "leg" states.

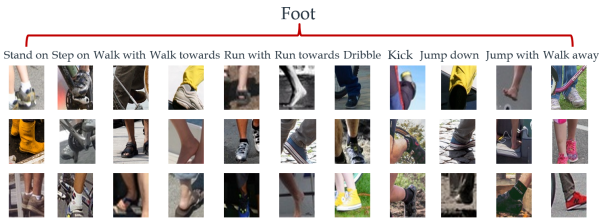


Figure 16. Some "foot" states.