Audio-Visual Action Prediction with Soft-Boundary in Egocentric Videos

Luchuan Song Jing Bi Chao Huang Chenliang Xu
University of Rochester
{lsong11@ur., jing.bi@, chuang65@ur., chenliang.xu@}rochester.edu

Abstract

This paper proposes a novel framework for accurate and stable action prediction in egocentric videos. The diverse nature of the captured audio and visual features from different domains, such as speech, environmental sounds, and visual objects, can cause significant inconsistencies in action prediction, exacerbated by the use of fixed classification boundaries in previous methods. To address this issue, we introduce a learnable classification module that adaptively adjusts the boundaries for each instance, resulting in more precise and stable action predictions. We validate the effectiveness of our approach on a new egocentric microwave operation dataset recorded with Microsoft HoloLens2, which includes both audio and visual streams. Our experimental results demonstrate that our approach achieves superior accuracy and stability compared to existing methods in predicting actions in real-world egocentric audio-visual videos.

1. Introduction

Wearable computing devices, such as AR glasses, provide promising avenues for enhancing user-environment interaction. A fundamental challenge in achieving better interaction is accurately understanding and predicting user behavior. Relying on a single modality is insufficient to handle the complexity and diversity of real-world scenarios. For instance, identifying a microwave oven is heating through the video stream alone can be challenging as the sound dominates the scene. As a result, capturing multimodal information is crucial to achieve accurate user behavior prediction.

In recent years, there has been a significant focus on leveraging both audio and visual information for improving egocentric video tasks. For instance, Owens *et al*. [11] leverage audio to aid the visual related tasks.

Egocentric videos are commonly captured in unconstrained environments, resulting in large intra-class distance due to variability in lighting, camera motion, and perspectives. For instance, the same action (such as open microwave) captured from different angles can have significantly different visual features, making it difficult to classify the action.

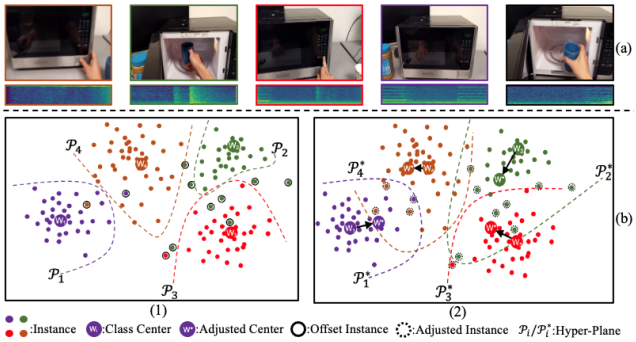

Figure 1. Overview of our proposed method. (a) We record and annotate the start time and end time of each action with the AR glass. (b) Our proposed IAM adapts to each instance point by making adjustments to the hyper-plane and classification center based on the features of different instances. This approach creates a soft boundary between classes.

We formulate the action recognition in egocentric videos as a multi-classification task, where we optimize an *n*-fold classification hyper-plane $\mathcal{P} \in \{\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_n\}$ based on the *n* class centers. Each category center $[\mathbf{w}_1, \mathbf{w}_2, ... \mathbf{w}_n]$ is the center of clustering of corresponding features in a high-dimensional space, and these centers are optimized during training to achieve optimal classification performance. However, the final optimized classification hyper-plane based on SoftMax in previous methods is determined solely by the class centers and cannot accurately classify in-the-wild frames. Specifically, instances near the optimized classification hyper-plane tend to be more ambiguous for classification, which can result in misclassification and inconsistencies in recognizing actions. For instance, the proximity of the action of 'microwave heating' to the hyper-plane can cause it to be misclassified as 'press the button'. To the best of our knowledge, this issue has not been previously addressed in audio-visual works for action recognition.

To address this challenge, we propose an instance adaptive module to incorporate instance-specific features into the hyper-plane, enabling it to fast adapt to the input instance. To evaluate the effectiveness of our proposed method, we conduct real world experiments with on the Microsoft

HoloLens2 (HL2), one of the most popular AR glass devices. The difference in perspective between the HL2 dataset, which is more focused on the direction of eye gaze, and public egocentric datasets presents a significant challenge in using models trained on public datasets for real world testing. To overcome this limitation, we manually recorded a small dataset includes 140 videos and the corresponding audio with HL2. Our dataset focuses on the task of operating a microwave oven for heating, and we collected videos and corresponding synchronous audios of multiple different users wearing HL2, repeating seven different actions such as 'open microwave', 'close microwave', 'place cup', 'take out cup', 'press button', 'heating', and 'alarm', and manually annotated the start and end times of these actions as labels. Our experiments demonstrate that our proposed method achieves excellent performance in real-world scenarios, highlighting the effectiveness of our approach.

Our contributions are summarized as follows: 1) We propose an instance adaptive module that dynamically adjusts the position of the hyper-plane according to the input instance, addressing the challenge of classification inconsistencies caused by large intra-class distances between real-world frames and dataset samples. 2) We introduce a small dataset recorded using the Hololens2 device, which contains videos of operating a microwave oven from a different egocentric perspective than previous datasets. Our dataset also includes full action annotations, enabling more real-life evaluation and development of egocentric audio-visual algorithm for practical applications.

2. Related Work

Action Recognition in Egocentric Videos. Egocentric videos provide various intrinsic cues that can be used for action recognition. For instance, hand gestures and movement have been utilized for action recognition. Yan *et al*. [13] utilized eye gaze to model the attention area. Other works have emphasized the importance of objects present in the scene and those related to the tasks for action recognition. Aboubakr *et al*. [2] modeled object state transitions to infer actions. In contrast to previous works, we only utilize video and corresponding audio to verify our instance adaptive module.

Egocentric Video Datasets. In recent years, several egocentric video datasets have been introduced to facilitate research on action recognition in first-person vision. EPIC-Kitchens [4] contains over 11 hours of egocentric video data captured in a kitchen environment. Activities of Daily Living (ADL) [6] consists of egocentric video data captured during six different activities of daily living, Ego4D [8] is a large-scale egocentric video dataset with multiple modalities and ground-truth annotations for action recognition, object detection, and hand-object interactions. These datasets have been useful for various person vision tasks, but they do not specifically target the vision challenges posed by using AR glasses. To address this gap, we recorded a dataset focusing on person vision using AR glasses in real-world scenarios.

Egocentric Audio and Visual Learning. The frequent and abrupt changes in camera view in egocentric video can make purely visual predictions unstable and inaccurate for action recognition. To address this limitation, several works have focused on incorporating additional sensory modalities, such as audio, to improve the stability and accuracy of the predictions. Owens *et al*. [11] has shown that incorporating audio information can help mitigate the instability of purely visual predictions. Recently, there has been a growing interest in learning semantic correspondences between visual and audio features for prediction. Akbari *et al*. [3] have all explored this approach. In our work, we follow these popular approaches and perform feature fusion on aligned audio and video streams

3. Method and Datasets

In this section, we introduce our proposed instance adaptive module (IAM) in detail and provide an overview of our HoloLens2 dataset.To extract the visual and audio features, we choose the simplest ResNet50 and CNN14 [9] backbone modules, respectively, and classify the prediction results together with the IAM.

3.1. Instance Adaptive Module

The instability of action recognition in egocentric video arises from the large intra-class distance between features within the same action class, resulting in ambiguity in instances that lie close to the classification hyperplane. Previous methods have employed classification centers trained to cover all training data, which helps to generalize to unseen data, but often results in poor performance in ambiguous cases. To address this issue, the hyperplane needs to dynamically adjust to each instance. In this work, we propose the Instance Adaptive Module (IAM), which extends the adaptive hyperplane idea from face forgery recognition [12], to higher-dimensional hyperplanes for action recognition in egocentric video. The IAM adaptively adjusts the classification center based on the instance itself, enabling robust and efficient predictions in the wild.

The pipeline of our network structure is shown in Fig 2, a two stream network to extract audio and video features separately, we select the audio length is 1 second and random sample 5 frames in the corresponding video as the input. For the audio stream, we use a 14-layers CNN [9] to extract the audio feature, which is pretrained on Audioset [7]. To capture pixel-level semantic features for visual features, we employ ResNet50 as a feature extractor which is pretrained on COCO [10] segmentation datasets. The input to IAM $\mathbf{x}$ is the output of the two stream feature extractor, specifically, the features from both streams are fused and then

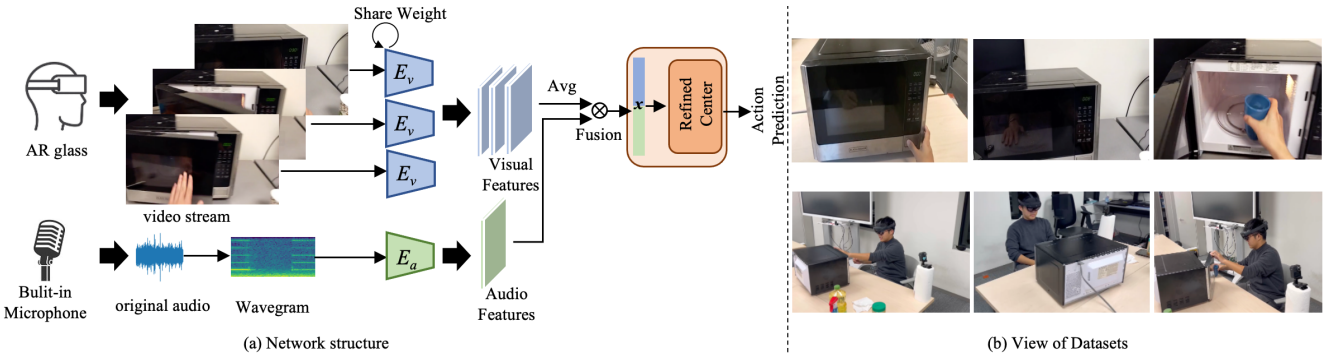

Figure 2. (a) The pipeline structure of our methods. The AR glass captures the video stream, the built-in microphone captures the audio, two stream input are encoded via E_v and E_a. E_v is a weight-sharing structure that encodes each frame separately. (b) The upper pair is the perspective of the AR glass, the lower is the perspective of the 3rd-person camera, the researcher wears the HL2 to record the videos.

subjected to a max-pooling operation to obtain $\mathbf{x}$. The $\mathbf{x}$ is a embedding feature vector of each instance. And we assume the corresponding action label as y (*e.g.* 'open microwave', 'close microwave', 'heating', 'press button' *etc.*). Then the conditional probability output (classification score) $P(Y=y|\mathbf{x})$ by a deep neural network can be estimated via the SoftMax operator after FC layer (the normalized form):

$$P(Y=y|\mathbf{x}) = \frac{\exp(\tau \frac{\mathbf{w}_y^\top}{\|\mathbf{w}_y^\top\|_2} \frac{\mathbf{x}}{\|\mathbf{x}\|_2})}{\sum_j^N \exp(\tau \frac{\mathbf{w}_j^\top}{\|\mathbf{w}_j^\top\|_2} \frac{\mathbf{x}}{\|\mathbf{x}\|_2})}, \tag{1}$$

where τ is a scaling factor and $[\mathbf{w}_1, \cdots, \mathbf{w}_n] \in \mathbb{R}^{d\times n}$ is the weight tensor of the last fully-connected layer. N denotes the number of classes (N is 7 in our microwave task). d is the dimension of embeddings. The classification center for each category can be represented by $[\mathbf{w}_1, \cdots, \mathbf{w}_n]$, where n is the number of categories. To classify an instance embedding $\mathbf{x}$, we compute its cosine distance with each $\mathbf{w}_i$. Previous approaches [3] commonly use a fixed positive margin, as shown in Equ 1, to classify all instances. However, as shown in Fig 1,some instances, such as the dotted points, may be located near the classification boundary and cannot be well-separated by a fixed margin. To address this challenge, we propose the instance-adapted classification centers (IAM) in Equ 2, which enables the dynamic adjustment of the margin to either positive or negative. This approach creates more flexible decision boundaries that can better adapt to difficult instances, resulting in improved classification performance.

The architecture of IAM is illustrated in Fig 2, each category is initialized to represent an FC layers, which we denote as $\mathbf{w}_i$. Then the corresponding Batch Normalization and ReLU are utilized to extract bias embeddings $\mathbf{b}_i$. Our IAM SoftMax is formalized as:

$$P(Y=y|\mathbf{x}) = \frac{\exp(\tau \frac{\mathbf{w}_y^\top + \mathbf{b}_y^\top(\mathbf{x})}{\|\mathbf{w}_y^\top + \mathbf{b}_y^\top(\mathbf{x})\|_2} \frac{\mathbf{x}}{\|\mathbf{x}\|_2})}{\sum_j^N \exp(\tau \frac{\mathbf{w}_j^\top + \mathbf{b}_j^\top(\mathbf{x})}{\|\mathbf{w}_j^\top + \mathbf{b}_j^\top(\mathbf{x})\|_2} \frac{\mathbf{x}}{\|\mathbf{x}\|_2})}, \tag{2}$$

we provide further insights into how the IAM adjusts the classification centers in Equ 3,

$$\begin{aligned} T_{\text{Norm}} &= \frac{\mathbf{w}^\top}{\|\mathbf{w}^\top\|_2}\frac{\mathbf{x}}{\|\mathbf{x}\|_2}, T_{\text{Bias}} = \frac{\mathbf{b}^\top(\mathbf{x})}{\|\mathbf{b}^\top(\mathbf{x})\|_2}; \\ T_{\text{IAM}} &= \frac{\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})}{\|\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})\|_2}\frac{\mathbf{x}}{\|\mathbf{x}\|_2} \\ &= \frac{\mathbf{w}^\top}{\|\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})\|_2}\frac{\mathbf{x}}{\|\mathbf{x}\|_2} + \frac{\mathbf{b}^\top}{\|\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})\|_2}\frac{\mathbf{x}}{\|\mathbf{x}\|_2} \\ &= \frac{\left\|\mathbf{w}^\top\right\|_2}{\|\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})\|_2}\left(\frac{\mathbf{w}^\top}{\|\mathbf{w}^\top\|_2}\frac{\mathbf{x}}{\|\mathbf{x}\|_2}\right) + \\ &\quad \frac{\left\|\mathbf{b}^\top(\mathbf{x})\right\|_2}{\|\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})\|_2}\left(\frac{\mathbf{b}^\top(\mathbf{x})}{\|\mathbf{b}^\top(\mathbf{x})\|_2}\frac{\mathbf{x}}{\|\mathbf{x}\|_2}\right) \\ &= \frac{\left\|\mathbf{w}^\top\right\|_2}{\|\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})\|_2}T_{\text{Norm}} + \frac{\left\|\mathbf{b}^\top(\mathbf{x})\right\|_2}{\|\mathbf{w}^\top + \mathbf{b}^\top(\mathbf{x})\|_2}T_{\text{Bias}} \\ &= \alpha T_{\text{Norm}} - \epsilon. \end{aligned} \tag{3}$$

As shown in Equ 3, The parameter ϵ in our framework can be seen as an adaptive margin that is specific to each instance, which is paramerized with multiple fully connected layers.

3.2. Dataset Overview

With focus on online action recognition with AR glass, we employed HL2 to simultaneously sample audio at 16 KHZ while transmitting video at 30 FPS [5]. As depicted in Fig 1(b), the dataset recording process involved the researcher wearing the HL2 device and performing actions based on rendered instructions within the HL2 display. The resulting dataset has a more natural perspective that better aligns with human vision than previous egocentric datasets recorded with cameras. Our dataset comprises 150 video clips, each of which has been annotated at the millisecond level using VIA [1]. For each video clip, we segmented the start and end times for each action and assigned a corresponding action labels. The video clips are randomly divided into 80% for training and 20% for testing.

3.3. Experiment and Results

In-the-wild Evaluation. Our method's strength lies in its robustness and accuracy for testing in real-world scenarios.

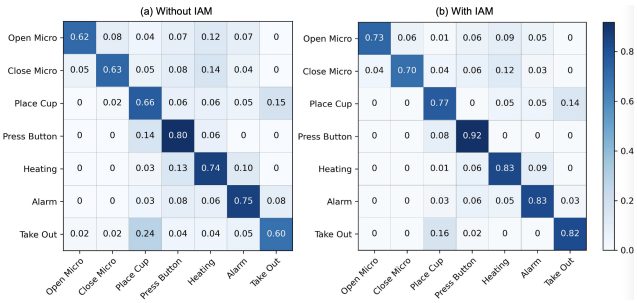

Figure 3. The confusion matrix for each label in evaluation. The higher accuracy with darker color. Comparison of (a) and (b) shows that IAM reduces confusion and improves accuracy for each action label.

Actions	*wo* IAM	Our
Open Micro	0.62	**0.73**
Close Micro	0.63	**0.70**
Place Cup	0.66	**0.77**
Press Button	0.80	**0.92**
Heating	0.74	**0.83**
Alarm	0.75	**0.83**
Take Out	0.60	**0.82**

Table 1. Evaluation results of the baseline method (with SoftMax shown in Equ 1 and the proposed method. Our method improves the classification accuracy for all actions labels.

We captured the stream from HL2 and push it to recognition model to enable real-time processing of audio and video synchronously without delays.

Implementation details. We divide videos in the datasets to clips of 1 second. Each segmented video clip includes 30 frames. The optimizer is AdamW and the learning rate sets to 1e-3. We train our model for 20 epochs.

Baseline and Metrics. We choose the normalized softmax as the baseline method and compare with the proposed IAM in our work. We compare the performance of different methods using accuracy as an intuitive metric.

3.4. Comparison Results

Quantitative Comparison Results. Table 1 shows the quantitative comparison results. We can find that with IAM, the accuracy of all action labels has been greatly improved. Furthermore, by analyzing the confusion matrix in Fig 3, we can find that our accuracy was significantly improved for some challenging cases, such as 'heating'. When using IAM, our model was able to more accurately predict the action of heating without random misclassifications.

Qualitative Results. We have included our results in a demo video in a following anonymous link[1], which we highly recommend viewing for a quantitative demonstration of the results.

[1]https://files.catbox.moe/az7l3l.mp4

4. Acknowledgments

This work has been partially supported by the Defense Advance Research Projects Agency (DARPA) under Contract HR00112220003. The content of the information does not necessarily reflect the position of the Government, and no official endorsement should be inferred.

References

[1] Vgg image annotator. https://www.robots.ox.ac.uk/~vgg/software/via/. 3

[2] Nachwa Aboubakr, James L Crowley, and Rémi Ronfard. Recognizing manipulation actions from state-transformations. *arXiv preprint arXiv:1906.05147*, 2019. 2

[3] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021. 2, 3

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 2

[5] Juan C Dibene and Enrique Dunn. Hololens 2 sensor streaming. *arXiv preprint arXiv:2211.02648*, 2022. 3

[6] Peter F Edemekong, Deb Bomgaars, Sukesh Sukumaran, and Shoshana B Levy. Activities of daily living. 2019. 2

[7] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 2

[8] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[9] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. 2

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[11] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 801–816. Springer, 2016. 1, 2

[12] Luchuan Song, Zheng Fang, Xiaodan Li, Xiaoyi Dong, Zhenchao Jin, Yuefeng Chen, and Siwei Lyu. Adaptive face forgery detection in cross domain. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 467–484. Springer, 2022. 2

[13] Yan Yan, Elisa Ricci, Gaowen Liu, and Nicu Sebe. Recognizing daily activities from first-person videos with multi-task clustering. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part IV 12*, pages 522–537. Springer, 2015. 2