An Egocentric Action Anticipation Framework via Fusing Intuition and Analysis

Tianyu Zhang[1,2], Weiqing Min[1,2], Ying Zhu[2], Yong Rui[3], Shuqiang Jiang[1,2]
[1] Key Lab of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China
[3] Lenovo Group, Beijing, 100085, China
tianyu.zhang@vipl.ict.ac.cn;{minweiqing,sqjiang}@ict.ac.cn;zhuying161@mails.ucas.ac.cn;yongrui@lenovo.com

ABSTRACT

In this paper, we focus on egocentric action anticipation from videos, which enables various applications, such as helping intelligent wearable assistants understand users' needs and enhance their capabilities in the interaction process. It requires intelligent systems to observe from the perspective of the first person and predict an action before it occurs. Owing to the uncertainty of future, it is insufficient to perform action anticipation relying on visual information especially when there exists salient visual difference between past and future. In order to alleviate this problem, which we call visual gap in this paper, we propose one novel Intuition-Analysis Integrated (IAI) framework inspired by psychological research, which mainly consists of three parts: Intuition-based Prediction Network (IPN), Analysis-based Prediction Network (APN) and Adaptive Fusion Network (AFN). To imitate the implicit intuitive thinking process, we model IPN as an encoder-decoder structure and introduce one procedural instruction learning strategy implemented by textual pre-training. On the other hand, we allow APN to process information under designed rules to imitate the explicit analytical thinking, which is divided into three steps: recognition, transitions and combination. Both the procedural instruction learning strategy in IPN and the transition step of APN are crucial to improving the anticipation performance via mitigating the visual gap problem. Considering the complementarity of intuition and analysis, AFN adopts attention fusion to adaptively integrate predictions from IPN and APN to produce the final anticipation results. We conduct experiments on the largest egocentric video dataset. Qualitative and quantitative evaluation results validate the effectiveness of our IAI framework, and demonstrate the advantage of bridging visual gap by utilizing multi-modal information, including both visual features of observed segments and sequential instructions of actions.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413964

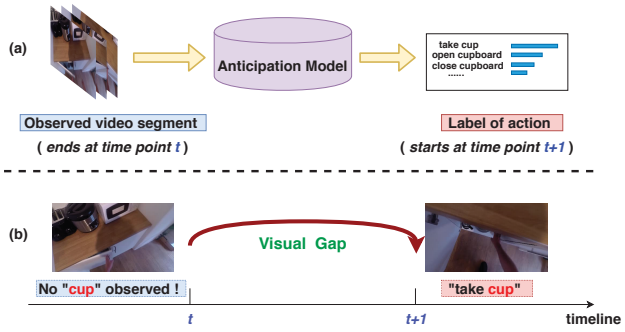

Figure 1: (a) An instance of egocentric action anticipation. The model takes an observed video segment x_t as its input and outputs the label y_{t+1} of the action which occurs one unit of time later. (b) The visual gap problem in egocentric action anticipation. Humans can often easily infer what will happen even though there exists salient visual difference between past and future. How to endow the model with this capability deserves exploration.

KEYWORDS

Egocentric Action Anticipation; Visual Gap; Intuition; Analysis; Adaptive Fusion

ACM Reference Format:
Tianyu Zhang[1,2], Weiqing Min[1,2], Ying Zhu[2], Yong Rui[3], Shuqiang Jiang[1,2]. 2020. An Egocentric Action Anticipation Framework via Fusing Intuition and Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413964

1 INTRODUCTION

Imagine such a scenario, someone whose hands are injured wants to cook when he/she is alone at home, it would be very helpful if an intelligent system (e.g., a wearable exoskeleton robot) can guess the user's intentions and needs [15], predict the user's next action in advance, and thus assist the latter to execute this action. This kind of ability to act as the first person is closely related to the task of egocentric (i.e., First-Person Vision [4]) action anticipation, which has attracted increasing attention in recent years [5, 9, 10, 20].

As shown in Fig. 1 (a), the task can be described as follows: given an observed video segment x_t which ends at time point t, the anticipation model is expected to output the label y_{t+1} of the action that starts at time point $t + 1$, where 1 represents one unit of time. There is no doubt that the task of predicting the future

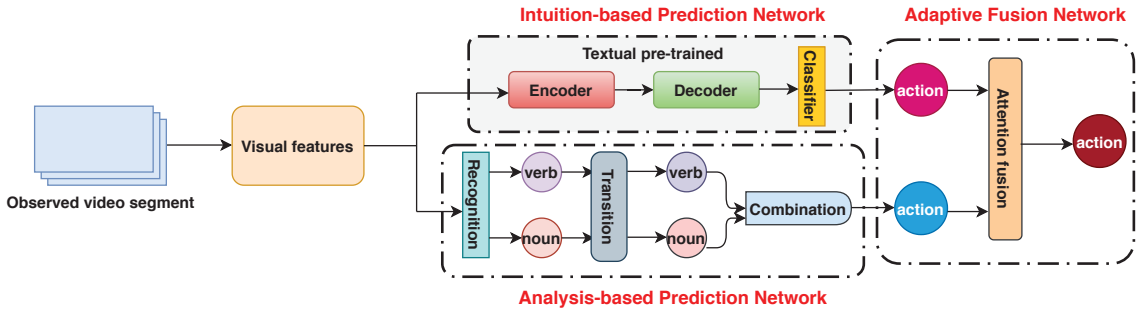

Figure 2: Overview of our proposed Intuition-Analysis Integrated (IAI) framework, which mainly consists of three sub-networks named Intuition-based Prediction Network (IPN), Analysis-based Prediction Network (APN), and Adaptive Fusion Network (AFN).

has a wide range of research significance as well as application value, including human-robot interaction and augmented/virtual reality. However, it is quite difficult for the model to perform action anticipation on account of the time distance between x_t and y_{t+1}. To be more specific, it is insufficient to address the task relying on visual information especially when there exists salient difference between past and future in terms of visual information. As illustrated in Fig 1 (b), the action "take cup" may not be correctly predicted if the object "cup" has not appeared in the observed video segment. In this paper, we call this phenomenon as "visual gap". The visual gap problem results from the inherent uncertainty of future [34] and adds challenges to the task of egocentric action anticipation. Since humans can often easily infer what will happen even though it is difficult to capture correlation between past and future at visual level, it is worth exploring how to enable the model to make predictions like humans.

Considering that action prediction from an egocentric perspective is directly and tightly related to human psychology, we turn our attention to this field. Inspired by relevant research [6, 12, 25], we propose an Intuition-Analysis Integrated (IAI) framework to imitate humans in performing egocentric action anticipation. These psychological studies suggest that: 1) intuition and analysis are two modes of cognition, which are served by interacting but independent systems; 2) intuition is a subconscious, habitual process while analysis is conscious and tends to combine information using organized principles; 3) intuition is neither the opposite of rationality nor a random process of guessing. Instead, it can outperform rational analysis in tackling many practical problems. Therefore, integrating both intuitive and analytical thinking is probably the most appropriate way to make predictions.

As shown in Fig. 3, our framework mainly consists of three parts: Intuition-based Prediction Network (IPN), Analysis-based Prediction Network (APN) and Adaptive Fusion Network (AFN). For IPN, since intuitive mode is automatic, holistic and cannot be easily articulated explicitly, we model it as an encoder-decoder structure connected with a classifier, an analogous black-box process that handles action as a whole. The encoder understands the past visual information and the decoder produces the future action information to be classified. More importantly, intuition draws from knowledge stored in the subconscious, which is named as tacit knowledge [6]. Considering the visual gap problem, we further introduce procedural instruction learning implemented by textual pre-training to enable IPN to acquire tacit knowledge in advance. Particularly, before sending visual features of x_t to IPN, we replace them with the sequential instructions of the action executed before y_{t+1} to pre-train the parameters of IPN, making IPN store considerable tacit knowledge. For APN, we design a set of rules to implement analysis-based prediction. Note that egocentric actions are usually represented as *(verb, noun)* pairs [5, 9], we divide the analytical process into three steps. First, we identify the current verb v_t and noun n_t based on the observation. Second, we transit from v_t and n_t to the future v_{t+1} and n_{t+1} separately depending upon a Markov assumption. The transition step is necessary in mitigating the visual gap problem via capturing statistical correlations between past and future. Finally, we combine v_{t+1} and n_{t+1} into y_{t+1}. This combination is guided by the prior knowledge acquired from prior statistics on co-occurrence probability between verbs and nouns, which can greatly avoid generating invalid combinations. Furthermore, since both intuition and analysis are indispensable in solving complex problems [25], AFN integrates results from IPN and APN into the final anticipation result via attention fusion, which enables our IAI framework to adaptively determine to give more credit to intuition or analysis by assigning different weights for them.

Considering kitchens witness extensive common daily activities, including taking containers, washing dishes and cooking recipes, we evaluate our model on the largest egocentric video dataset EPIC-Kitchens in the kitchen environment [5]. The experimental results show the effectiveness of our proposed method. Overall, our key contributions are summarized as follows:

- To the best of our knowledge, we are the first to jointly introduce both intuition and analysis to imitate humans for egocentric action anticipation, which is expected to provide methodological basis for future research in the multimedia.
- We propose one Intuition-Analysis Integrated (IAI) framework for egocentric action anticipation, where an encoder-decoder structure is adopted for intuition and a three-step pipeline of recognition-transition-combination is designed for analysis. Furthermore, sequential instructions of actions and prior knowledge from co-occurrence statistics are explored to improve the anticipation performance.

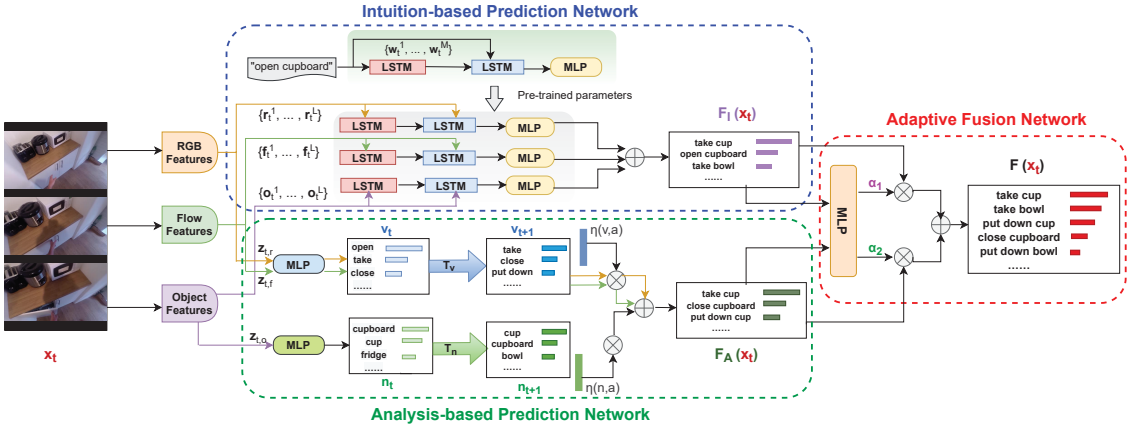

Figure 3: The concrete architecture of our IAI framework where the three sub-networks are explicitly modeled. Given x_t, outputs from IPN and APN (i.e., $F_I(x_t)$ and $F_A(x_t)$) are integrated to obtain the final anticipation result $F(x_t)$ via AFN.

- We validate the effectiveness of our IAI framework on the largest egocentric video dataset EPIC-Kitchens. The qualitative results also demonstrate that our IAI framework is capable of bridging visual gap between past and future.

2 RELATED WORK

Egocentric Video Analysis Different from videos recorded in third person vision where cameras are fixed in the environment, egocentric videos are captured from the subjects' own point of view [23] and explored for different purposes [4], such as object recognition and tracking [8], hand grasp analysis and gesture recognition [11, 36], action and activity analysis [19, 24]. Among these, action analysis plays a connecting role between lower-level tasks (e.g. object recognition, hand grasp analysis) and higher-level applications such as human-robot interaction [16] and health care [3]. While more attention is paid to egocentric action recognition [19, 24, 31], some recent works have been devoted to action anticipation in first person vision [9, 10], which requires understanding the past and making assumptions about the future. In addition, similar to third person vision, egocentric video data can also be collected with multi-sensor wearable devices [4] in different scenes, such as parks [7], library exhibitions [29] and kitchens [5]. In this paper, we conduct egocentric video analysis for action anticipation on the largest egocentric video dataset EPIC-Kitchens [5] in the daily kitchen environment.

Action Anticipation The task of action anticipation aims at predicting an action before it is actually executed [17]. One characteristic "predicting an action before it occurs " distinguishes this task from early action recognition where an ongoing action should be predicted based on partially observed videos [28]. Another characteristic "predicting a short-duration action " also makes it different from long-term activity prediction which requires anticipating multiple actions happening in the future for quite a long time [1] as well as the task of forecasting the first-person trajectory [27].

Compared with action anticipation in third person vision [16–18, 34], fewer works focus on egocentric action anticipation. Damen *et al.* [5] utilized action recognition models for action anticipation in the kitchen scenario. Miech *et al.* [20] proposed a transitional model to establish an interpretable relationship between past and future, which is related to our proposed Analysis-based Prediction Network. Different from [20], we process verbs and nouns separately at first and combine them afterwards instead of handling them as a whole. Furnari *et al.* [9] considered the task as a multi-label classification problem and studied the design of loss functions. They further made use of three kinds of visual features by sending them into Rolling-Unrolling LSTMs [10], which can be viewed as a special case of our proposed Intuition-based Prediction Network. In order to bridge visual gap between past and future, we consider multi-modal information, including both visual features of videos and sequential instructions of actions, to perform action anticipation via integrating intuition and analysis.

3 PROPOSED FRAMEWORK

In this section, we will detail the Intuition-Analysis Integrated (IAI) framework. Given the observed video segment x_t, the task of predicting the label of action which starts one unit of time later (i.e., y_{t+1}) can be treated as a supervised classification problem [9], where each x_t is a training sample with an annotated ground-truth label $\bar{y}_{t+1}$. Therefore, our goal is to find a function $F(\cdot)$ which leads $y_{t+1} = F(x_t)$ to approximate $\bar{y}_{t+1}$ as close as possible. To implement this, as illustrated in Fig. 3, we first design $F_I(\cdot)$ and $F_A(\cdot)$ respectively for Intuition-based Prediction Network (IPN) and Analysis-based Prediction Network (APN). Then Adaptive Fusion Network (AFN) takes $F_I(x_t)$ and $F_A(x_t)$ as its inputs and yields $F(x_t)$ as the final anticipation result.

3.1 Intuition-based Prediction Network

Given an input video segment x_t, we firstly sample a snippet with L frames and extract visual features of each sampled frame. In our framework, similar to [10], we consider three kinds of visual features, namely RGB features $\{\mathbf{r}_t^1, \ldots, \mathbf{r}_t^L\}$, flow features $\{\mathbf{f}_t^1, \ldots, \mathbf{f}_t^L\}$ and object features $\{\mathbf{o}_t^1, \ldots, \mathbf{o}_t^L\}$, where $\mathbf{r}_t^i$, $\mathbf{f}_t^i$ and $\mathbf{o}_t^i$ are fixed-length feature vectors of the i-th frame. As discussed earlier, the key of intuition is to store tacit knowledge. Considering LSTMs

are good at capturing temporal dependencies in sequential data, we employ them to implement this process. As shown in Fig. 3, the process of handling three types of visual features is identical. Taking RGB features for example, we sent $\{\mathbf{r}_t^1, \ldots, \mathbf{r}_t^L\}$ into the first LSTM to encode the past information, where each hidden state $[\mathbf{h}_t^i]_e (i = 1, \ldots, L)$ can be computed as:

$$\begin{aligned} [\mathbf{h}_t^i]_e &= \text{LSTM}_{\theta_e}(\mathbf{r}_t^i, [\mathbf{h}_t^{i-1}]_e) \\ [\mathbf{h}_t^0]_e &= \mathbf{0} \end{aligned} \tag{1}$$

where LSTM_{θ_e} represents the encoder with learnable parameters θ_e. The last hidden state $[\mathbf{h}_t^L]_e$ is sent to the second LSTM as its initial hidden state:

$$\begin{aligned} [\mathbf{h}_t^i]_d &= \text{LSTM}_{\theta_d}(\mathbf{r}_t^i, [\mathbf{h}_t^{i-1}]_d) \\ [\mathbf{h}_t^0]_d &= [\mathbf{h}_t^L]_e \end{aligned} \tag{2}$$

where LSTM_{θ_d} plays a role of decoder with learnable parameters θ_d and its last hidden state $[\mathbf{h}_t^L]_d$ is viewed as the overall representation of decoded information. Then we apply a MLP to compute action scores $\mathbf{s}_{t,r}$ with respect to RGB features:

$$\mathbf{s}_{t,r} = \Phi_{\theta_s}([\mathbf{h}_t^L]_d) \tag{3}$$

where $\Phi_{\theta_s}(\cdot)$ stands for a MLP with learnable parameters θ_s. Similarly, we can obtain action scores $\mathbf{s}_{t,f}$ and $\mathbf{s}_{t,o}$ with respect to flow and object features by replacing $\{\mathbf{r}_t^1, \ldots, \mathbf{r}_t^L\}$ with $\{\mathbf{f}_t^1, \ldots, \mathbf{f}_t^L\}$ and $\{\mathbf{o}_t^1, \ldots, \mathbf{o}_t^L\}$ based on Eq. (1)-(3). Action scores from each visual modality are fused to obtain the predicted result $F_I(x_t)$ as in [10].

Procedural Instruction Learning via Textual Pre-training Considering the visual gap problem where visual information is insufficient to build up a connection between x_t and y_{t+1}, we introduce textual pre-training to learn procedural instructions. For $\bar{y}_{t+1}$, suppose its previous annotated action label $\bar{y}_t$ is composed of M words and $\mathbf{w}_t^j (j = 1, \ldots, M)$ is the embedded word vector of the j-th word, we pre-train IPN by treating $\{\mathbf{w}_t^1, \ldots, \mathbf{w}_t^M\}$ as a training sample and $\bar{y}_{t+1}$ as its ground-truth label. All learnable parameters (i.e., θ_e, θ_d and θ_s) can be trained in advance as follows:

$$\begin{aligned} [\mathbf{h}_t^j]_e &= \text{LSTM}_{\theta_e}(\mathbf{w}_t^j, [\mathbf{h}_t^{j-1}]_e) \\ [\mathbf{h}_t^j]_d &= \text{LSTM}_{\theta_d}(\mathbf{w}_t^j, [\mathbf{h}_t^{j-1}]_d) \\ F_I(x_t) &= \Phi_{\theta_s}([\mathbf{h}_t^M]_d) \end{aligned} \tag{4}$$

where j ranges from 1 to M and $[\mathbf{h}_t^0]_e = \mathbf{0}, [\mathbf{h}_t^0]_d = [\mathbf{h}_t^M]_e$. By learning from procedure instructions in textual form, IPN is expected to store relevent tacit knowledge ahead, which guides it to make intuitive predictions when observing the corresponding video.

3.2 Analysis-based Prediction Network

Compared with intuition, analysis-based prediction prefers to perform more interpretable operations under given rules. We divide the analytical process into three steps, namely recognition, transition, and combination guided by prior knowledge.

3.2.1 Recognition. As shown in Fig. 3, the whole analyical process can be regarded as processing information through two branches, which we call verb-based branch and noun-based branch. This is because an egocentric action is generally represented as a *(verb, noun)* pair [5, 9], and either the number of verb classes N_v or noun classes N_n is much smaller than action classes N_a. To reduce computation complexity, we handle nouns and verbs separately at first and leverage prior knowledge to combine them afterwards. Considering object features convey considerable noun information while RGB and flow features capture appearance and motion indicating verb-relevant information, we utilize object features for the noun-based branch and RGB/flow features for the verb-based branch. Before sending object features to the noun-based branch, we first obtain an overall representation $\mathbf{z}_{t,o}$ by aggregating the frame-level features $\{\mathbf{o}_t^1, \ldots, \mathbf{o}_t^L\}$ over time:

$$\mathbf{z}_{t,o} = f(\mathbf{o}_t^1, \ldots, \mathbf{o}_t^L) \tag{5}$$

where f is a temporal aggregation function and we utilize LSTM for f in our experiments. Similarly, we also obtain $\mathbf{z}_{t,r}$ and $\mathbf{z}_{t,f}$ from $\{\mathbf{r}_t^1, \ldots, \mathbf{r}_t^L\}$ and $\{\mathbf{f}_t^1, \ldots, \mathbf{f}_t^L\}$ before sending RGB and flow features to the verb-based branch in an identical manner. Then we use MLPs $\Phi_v(\cdot)$ and $\Phi_n(\cdot)$ to separately recognize the verb $\mathbf{v}_t$ and the noun $\mathbf{n}_t$:

$$\mathbf{v}_t = \Phi_v(\mathbf{z}_{t,r/f}) \qquad \mathbf{n}_t = \Phi_n(\mathbf{z}_{t,o}) \tag{6}$$

where $\mathbf{v}_t^j$ denotes the j-th ($j = 1, \ldots, N_v$) element of $\mathbf{v}_t$ and its value $P(\mathbf{v}_t^j)$ represents the probability that $\mathbf{v}_t$ belongs to the j-th verb class. The similar meaning also goes for $\mathbf{n}_t$.

3.2.2 Transition. Considering that the prediction of the next action depends upon the recognition results, we model the sequence of performed actions as a Markov process. Different from [20] which regarded an action as a whole, we process its verb and noun separately. Therefore, at the time point $t + 1$, the probability that the verb $\mathbf{v}_{t+1}$ belongs to the j-th class can be computed as:

$$P(\mathbf{v}_{t+1}^j) = \sum_{i=1}^{N_v} P(\mathbf{v}_{t+1}^j | \mathbf{v}_t^i) P(\mathbf{v}_t^i) \tag{7}$$

where $P(\mathbf{v}_{t+1}^j | \mathbf{v}_t^i)$ stands for the conditional probability between the past and the future verb. In other words, $P(\mathbf{v}_{t+1}^j | \mathbf{v}_t^i)$ is actually at the (i, j) position of Markov transition matrix $\mathbf{T}_v$ which captures the explicit correlation between the past and the future verb. Similarly, the transition process from the noun $\mathbf{n}_t$ to $\mathbf{n}_{t+1}$ is computed as:

$$P(\mathbf{n}_{t+1}^j) = \sum_{i=1}^{N_n} P(\mathbf{n}_{t+1}^j | \mathbf{n}_t^i) P(\mathbf{n}_t^i) \tag{8}$$

where $P(\mathbf{v}_{t+1}^j | \mathbf{v}_t^i)$ is at the (i, j) position of transition matrix $\mathbf{T}_n$.

3.2.3 Combination. Suppose the k-th class action a^k is composed of the i-th class verb v^i and the j-th class verb n^j, which can be represented as $a^k = (v^i, n^j)$, it is conceivable that we cannot obtain $P(y_{t+1} = a^k)$ by simply multiplying $P(\mathbf{v}_{t+1}^i)$ and $P(\mathbf{n}_{t+1}^j)$ because many *(verb,noun)* pairs are invalid. Therefore, we propose a way to combine $\mathbf{v}_{t+1}$ and $\mathbf{n}_{t+1}$ guided by prior knowledge:

$$P(a^k) = \frac{\exp\left(\eta(v^i, a^k) P(\mathbf{v}_{t+1}^i) + \eta(n^j, a^k) P(\mathbf{n}_{t+1}^j)\right)}{\sum_{k=1}^{N_a} \exp\left(\eta(v^i, a^k) P(\mathbf{v}_{t+1}^i) + \eta(n^j, a^k) P(\mathbf{n}_{t+1}^j)\right)} \tag{9}$$

where $\eta(v^i, a^k)$ represents the prior knowledge that given the verb category v^i, how possible it is to predict the action category a^k, which can be computed as:

$$\eta(v^i, a^k) = \frac{\sum_{c=1}^{N} I_c(a^k)}{\sum_{c=1}^{N} I_c(v^i)} \tag{10}$$

where N is the total number of training samples, and $I_c(a^k)$ is an indicator function reflecting whether a^k is in the c-th training sample. The calculation method also applies to $\eta(n^j, a^k)$. In this way, we can avoid generating invalid predictions to a large degree. Therefore, we can obtain $y_{t+1} = F_A(x_t)$ with its k-th element $P(y_{t+1} = a^k)$ finally computed based on Eq. (9).

3.3 Adaptive Fusion Network

Based on psychological studies [12, 25], intuition-based prediction might be more useful than analysis-based prediction in some cases and vice versa, which means neither results from IPN (i.e., $F_I(x_t)$) nor APN (i.e, $F_A(x_t)$) could be neglected. Therefore, in order to further improve the anticipation performance, we adopt an adaptive integration strategy to effectively exploit both $F_I(x_t)$ and $F_A(x_t)$. To be more specific, inspired by previous study on attention mechanism [2], we first send $F_I(x_t)$ and $F_A(x_t)$ to a MLP $\Phi(\cdot)$ to compute attention scores for them:

$$s_1 = \Phi\left(F_I(x_t)\right) \qquad s_2 = \Phi\left(F_A(x_t)\right) \tag{11}$$

where s_1 and s_2 are attention scores indicating the relative importance of intuition or analysis for the final prediction. Then we obtain fusion weights by normalizing the attention scores:

$$\alpha_1 = \frac{\exp(s_1)}{\exp(s_1) + \exp(s_2)} \qquad \alpha_2 = \frac{\exp(s_2)}{\exp(s_1) + \exp(s_2)} \tag{12}$$

where α_1 and α_2 are fusion weights. This operation ensures that α_1 and α_2 sum to one, which considers the complementarity of intuition and analysis in making predictions. Finally, we obtain the fused anticipation result $F(x_t)$ by integrating the two parts with a linear combintion:

$$F(x_t) = \alpha_1 F_I(x_t) + \alpha_2 F_A(x_t) \tag{13}$$

In this way, our proposed framework can make predictions by taking advantage of both intuition and analysis at the same time. Particularly, it is capable of dynamiclly assigning fusion weights for IPN and APN, which means it can adaptively choose to give more trust to intuition-based prediction or analysis-based prediction.

4 EXPERIMENT

4.1 Dataset and Evaluation Metrics

We evaluate our proposed framework on the largest dataset in first person vision: EPIC-Kitchens [5], which consists of 55 hours of videos collected by 32 participants performing daily activities in their native kitchen environments. The dataset is very close to real-world data in that extensive and various cooking-related actions such as "take cup" and "wash carrot" are recorded from an egocentric perspective. There are 125 unique verb classes and 352 unique noun classes in total while the number of unique action categories is 2,513. Similar to [10], we randomly select 232 videos with 23,493 segments for training and 40 videos with 4,979 segments for validation. Test results are obtained from the evaluation server.

Top-1 accuracy and Top-5 accuracy are adopted as evaluation metrics. For simplicity, we make Top1@N, Top1@V and Top1@A denote Top-1 accuracy with respect to nouns, verbs and actions, and similar meaning for Top5@N, Top5@V and Top5@A.

Table 1: The effect of procedural instruction learning via textual pre-training on IPN (%).

Method	Top1@A	Top5@A
Visual(rgb)	12.28	29.38
Textual+Visual(rgb)	12.50	29.76
Improvement	+0.22	+0.38
Visual(flow)	6.79	17.54
Textual+Visual(flow)	7.16	18.09
Improvement	+0.37	+0.55
Visual(obj)	9.44	28.13
Textual+Visual(obj)	9.68	28.62
Improvement	+0.24	+0.49
Visual(rgb+flow+obj)	13.57	32.73
Textual+Visual(rgb+flow+obj)	13.90	33.19
Improvement	+0.33	+0.46

4.2 Experimental Setup

Our task requires predicting the label of action starting at time point $t + 1$ based on the observed video segment that ends at time point t, where 1 represents one second. Similar to [10], we employ three visual modalities (i.e., RGB, flow and object features). Specifically, both RGB frames and pre-computed optical flows provided by [5] are sent to the BN-Inception CNN [14] to obtain 1024-dimensional feature vectors. The object features are extracted by sending each frame to the Faster R-CNN object detector [26] with a ResNet-101 backbone [13] to obtain one 352-dimensional feature vector where each element represents the existence probability of each noun in the dataset. We sample a snippet with 11 frames for each input. For textual pre-training in IPN, we use fixed-length vectors to embed the input words whose dimensions are equal to the corresponding visual modality. We set the hidden state size of LSTM_{θ_e} and LSTM_{θ_d} to 1024. The MLPs we use consist of three fully connected layers with ReLU non-linearities. Our model is trained by minimizing the standard cross entropy via the SGD optimizer with momentum of 0.9 and a batch size of 128. The learning rate is set to 0.01 initially and decayed by a factor of 10 after 50 epochs. To regularize the training and avoid overfitting, dropout with retain probability 0.8 is used. All of our experiments are trained for 200 epochs.

4.3 Effect of Procedural Instruction Learning

To evaluate the effect of procedural instruction learning implemented by textual pre-training in IPN, we report four groups of comparative results in Table 1. For the first group, Visual(rgb) denotes only RGB features are used to train IPN while Textual+Visual(rgb) means textual annotations of actions are utilized to pre-train IPN. Similar meaning goes for the second and third groups. For the last group, Visual(rgb+flow+obj) represents all three visual features are

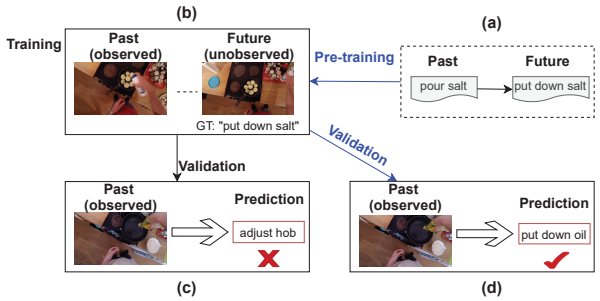

Figure 4: The improvement brought by procedural instruction learning via textual pre-training.

used to train IPN, which is consistent with descriptions in Section 3.1. From Table 1, we can see that: (1) utilizing three visual features outperforms using single one; (2) introducing procedural instruction learning via textual pre-traning brings consistent improvement to simply using visual features to train IPN.

We further show an example to visualize this improvement in Fig. 4. As shown from (b)→(c), relying only on visual information may lead to the wrong prediction. Intrinsically, the time distance between past and future makes visual context discontinuous at temporal level whereas textual context can be unaffected. Therefore, seeing from (a)→(b)→(d), by introducing procedural instruction learning via textual pre-training, IPN is more likely to produce the accurate result, which can be interpreted as storing tacit knowledge for intuition in advance.

4.4 Ablation Studies on APN

For APN, it consists of three steps: recognition, transition and combination. To eliminate any interference between modalities, we first conduct ablation stuides with respect to each step by maintaining the input visual modality as object features. In addition, we also evaluate the performance of different combinations used for APN.

Rationality of handling verbs and nouns separately. To verify the rationality of the two-branch process, which handles verbs and nouns separately in analysis-based prediction, we compare this method with three baselines:

- Single branch (V): Single-branch process where only verbs are recognized, transited and guided to actions by $\eta(v, a)$.
- Single branch (N): Single-branch process where only nouns are recognized, transited and guided to actions by $\eta(n, a)$.

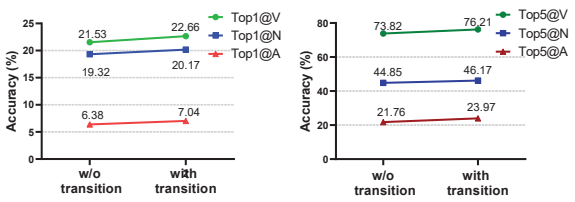

Figure 5: The necessity of transition in APN.

Table 2: The rationality of processing verbs and nouns separately (%).

Method	Top1@V	Top1@N	Top1@A	Top5@V	Top5@N	Top5@A
Single branch (V)	20.16	8.25	3.97	71.94	42.56	18.78
Single branch (N)	19.07	12.79	5.13	70.08	44.31	20.83
Single branch (A)	**23.92**	19.61	6.88	75.82	45.23	23.71
Two branches (V+N)	22.66	**20.17**	**7.04**	**76.21**	**46.17**	**23.97**

- Single branch (A): Single-branch process where actions are recognized and transited as a whole.

Table 2 lists the performance of different branches. We can see that our adopted two-branch process outperforms handling verbs or nouns only. This is easy to understand because whether a verb or a noun is an indispensable part of an action. Furthermore, the performance of processing verbs and nouns separately is also better than processing actions as a whole, which is reflected in both Top-1 and Top-5 anticipation accuracy on actions (+0.16% and +0.28%). This is mainly because both N_v (125) and N_n (352) are in a much smaller scale than N_a (2513), which makes addressing verbs and nouns in a separate way actually lower the per-iteration complexity by cutting down on the number of parameters.

Necessity of transition. To demonstrate the necessity of transition process in analysis-based prediction, we maintain the steps of recognition and combination but remove the transition step from APN, which means the recognized verbs and nouns are directly combined guided by prior knowledge. As shown in Fig. 5, the transition process based on Markov assumption brings considerable improvement (+0.66% and +2.21% with respect to Top-1 and Top-5 accuracy) to the action anticipation results. This confirms that the transition process is necessary in mitigating the visual gap problem via establishing a powerful relationship between past and future.

Superiority of combination guided by prior knowledge. To assess the role of combination guided by prior knowledge, we compare our method with a baseline that simply multiplies the probabilities of verbs and nouns to obtain the probability of actions. We call it as naive combination, which can be formalized as $P(y_{t+1} = a^k) = P(v^i_{t+1})P(n^j_{t+1})$ based on the description in Section 3.2.3. For a fair comparison, we maintain the process of recognition and transition. It can be observed from Fig. 6 that combining verbs

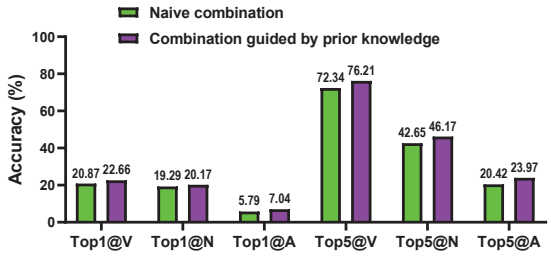

Figure 6: The superiority of combination guided by prior knowledge.

Table 3: Evaluation on feature combinations used for APN (%).

Method	Top1@V	Top1@N	Top1@A	Top5@V	Top5@N	Top5@A
V(R/F/O)+N(R/F/O)	28.19	18.51	10.09	76.41	44.03	26.12
V(R/F/O)+N(O))	29.38	19.23	10.24	**77.69**	43.65	26.91
V(R/O)+N(R/F/O)	28.76	19.33	10.15	75.87	43.86	25.90
V(F/O)+N (R/F/O)	27.13	17.57	9.98	72.65	43.51	24.95
V(R/F)+N(R/F/O)	28.94	19.57	10.43	76.19	44.11	26.28
V(R/O)+N(O)	29.04	19.52	10.71	77.43	44.29	**27.41**
V(F/O)+N(O)	29.01	**20.39**	10.49	76.05	44.40	26.81
V(R/F)+N(O)	**30.39**	19.94	**10.87**	77.16	**44.74**	27.37

and nouns guided by prior knowledge definitely improves both Top-1 and Top-5 anticipation accuracy on actions (+1.25% and +3.55%) as well as verbs (+1.79% and 3.87%) and nouns (+0.88% and +3.52%). It is understandable that this way of combination fully considers the co-occurrence probability between verbs and nouns.

Evaluation on feature combinations. To evaluate the performance of different feature combinations used for APN, we list several cases in Table 3. As shown in the last case, V(R/F) represents RGB and flow features are used to process verb-based branch while N(O) represents object features are used to process noun-based branch. Similar meaning also goes for other cases. From Table 3, we see that the combination of V(R/F) and N(O) outperforms other combinations under many metrics, even though some metrics can not surpass. It is understandable because object features are extracted based on noun information while RGB and flow features capture appearance and motion that indicate verb-relevant information.

4.5 Effect of Adaptive Fusion

In order to evaluate the strategy of attention fusion which adaptively integrates anticipation results from IPN and APN, we compare it with two traditional fusion strategies. One is average fusion that simply averages the anticipation results from two parts while the other is maximum fusion that takes the larger one as the final result. We denote them as Fusion-Avg and Fusion-Max in Table 4. As shown in Table 4, using adaptive fusion (Fusion-Ada) further improves the performance of both IPN (+0.17% and +0.18%) and APN (+3.20% and +5.00%) with respect to Top-1 and Top-5 accuracy on actions, which demonstrates the complementarity between intuition and analysis. We also see that attention fusion outperforms the other two fusion strategies. Considering that egocentric action anticipation is a highly complex and uncertain task, it is essential that our method can dynamically adjust the fusion weights of intuition and analysis instead of generating final results under fixed principles, which means our proposed framework is more flexible in making predictions based on different situations.

4.6 Performance Comparison among Models

As shown in table 5, we present the performance comparison between our proposed method and the following methods: 2SCNN [30], TSN [35], TSN+MCE [9], Miech *et al.* [20] and RULSTM [10] on the unseen test test. From table 5, we can see that our method exceeds the two action recognition models (i.e., 2SCNN and TSN)

Table 4: The effect of adaptive fusion in our framework (%).

Method	Top1@V	Top1@N	Top1@A	Top5@V	Top5@N	Top5@A
IPN	32.72	22.36	13.90	78.65	50.17	33.19
APN	30.39	19.94	10.87	77.16	44.74	27.37
Fusion-Avg	33.07	22.24	13.97	79.01	**50.86**	33.23
Fusion-Max	33.31	22.17	13.68	78.91	49.77	33.11
Fusion-Ada	**33.33**	**22.44**	**14.07**	**79.15**	50.55	**33.37**

Table 5: Performance of our method and compared approaches (%).

Method	Top1@V	Top1@N	Top1@A	Top5@V	Top5@N	Top5@A
2SCNN [30]	25.23	9.97	2.29	68.66	27.38	9.35
TSN [35]	25.30	10.41	2.39	68.32	29.50	9.63
TSN+MCE [9]	21.27	9.90	5.57	63.33	25.50	15.71
Miech *et al.* [20]	**28.37**	12.43	7.24	69.96	32.20	19.29
RULSTM [10]	27.01	**15.19**	8.16	69.55	34.38	21.20
Ours-IPN	27.24	14.58	8.06	69.17	34.21	20.21
Ours-APN	24.07	14.65	7.27	68.62	34.45	18.33
Ours-IAI	27.89	14.89	**8.57**	**70.06**	**35.51**	**21.41**

by a large margin, which demonstrates that methods used for action recognition are not applicable to action anticipation. Our method also outperforms two powerful models (i.e., TSN+MCE and Miech *et al.*) as well as the state-of-the-art results (i.e., RULSTM), which further verifies the effectiveness of our method.

4.7 Qualitative Results

Fig. 7 shows some qualitative examples of anticipation results obtained by our proposed framework. For each example, we show Top-5 predictions with respect to IPN, APN and the whole IAI framework. The bar next to each predicted reuslt visualizes the corresponding probability. Specifically, the second and third column show the results obtained by IPN and APN separately while the forth column shows the weighted sum of intuition-based and analysis-based predicted results. Particularly, the correct one is highlighted in bold and underlined.

For the first example, even if the anticipation object "mug" has not appeared in the observation, both IPN and APN can predict the ground-truth action "get mug" with certain probabilities. Through adaptive fusion, the IAI framework is able to make an accurate prediction, which shows advantages in mitigating the visual gap problem. The second and third examples also demonstrate that our framework can further improve the anticipation performance of both intuition-based prediction and analysis-based prediction.

It should be noted that our proposed framework fails in the last example, a more complex case that is uncertain and intractable in essence. To be more specific, the observed segment reflects the action of opening the fridge which indicates something unobserved is likely to be taken from the fridge. However, different from the first example, the situation in the fridge is much more complicated than that in the cupboard because the fridge always stores all kinds of food and drinks while the cupboard only contains limited kinds of containers such as cups and bowls. As a result, it is still difficult to

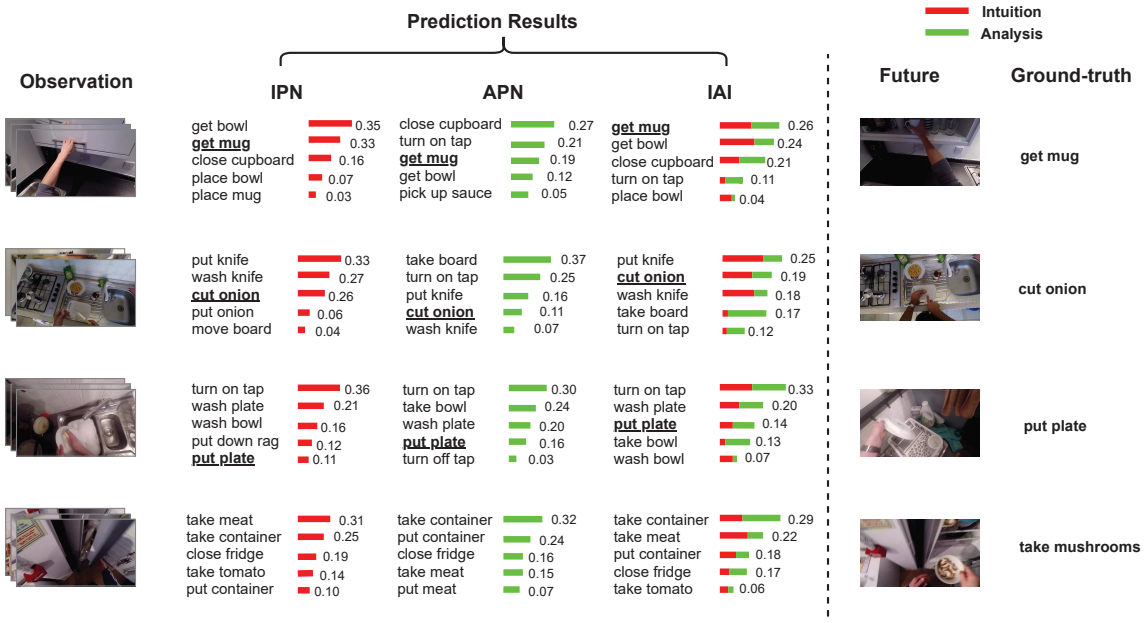

Figure 7: Qualitative examples of the egocentric action anticipation results obtained by our IAI framework.

infer which object will be taken from the fridge from the perspective of either intuition or analysis. As shown in the last example, the model tends to anticipate commonly stored objects in the fridge such as "meat" and "tomato" instead of the "mushrooms" that exist at a much smaller frequency. This case reveals that the current information we utilize is still very limited, regardless of the visual form or the textual form.

4.8 Discussions

In our method, we utilize multi-modal information by introducing textual modality to mitigate the visual gap problem which is intractable relying only on visual information. However, the current available textual information only contains annotations of short-duration actions, making it difficult to have a comprehensive understanding of the ongoing activities happening in one special scenario. Particularly, in the EPIC-Kitchens from the kitchen scenario, it is worth introducing more information from cooking domain [21], such as ingredients, sequential cooking instructions as well as food attributes [22]. For example, if the ingredient information of the being-prepared dishes is available, the model is more likely to make a correct prediction in the last case of Fig. 7. Furthermore, in order to embed more food-related information, the anticipation task should also be extended from predicting a single short-duration action to predicting a long-term activity consisting of multiple actions in the future, similar to [1].

Moreover, our proposed IAI framework is inspired by studies on the cognitive modes of humans. Admittedly, this kind of simulation is rough and premature due to the lack of deeper understanding of the field of psychology. Overall, our work aims to construct a basic framework, which is expected to expand the boundaries of relevant research. Therefore, many questions are worth exploration in the future. For example, the task of action anticipation depends upon the understanding of the past video information so that studies on related tasks such as video caption [33] and video question answering [32] may offer significance for reference. More intrinsically, how to model the process of intuition-based prediction and analysis-based prediction as well as integrate them in a more powerful manner requires deeper investigations into psychology. Furthermore, whether the framework is also applicable to other tasks (especially first person vision related) needs further study.

5 CONCLUSIONS

In this paper, we have proposed one framework which adaptively integrates intuition and analysis to imitate humans in performing egocentric action anticipation. Our work can be extended in the following three directions: 1) introducing more multi-modal information from specific domain, such as ingredients, cooking instructions and food attributes from cooking domain in the kitchen environment; 2) extending the task from predicting a single short-duration action to predicting a long-term activity; 3) optimizing the design of intuition-based prediction and analysis-based prediction via deeper studies on psychology as well as other video understanding tasks.

ACKNOWLEDGMENTS

This work was supported by National Key Research and Development Project of New Generation Artificial Intelligence of China, under Grant 2018AAA0102500, in part by the National Natural Science Foundation of China under Grant 61532018, 61972378, U1936203, in part by the Lenovo Outstanding Young Scientists Program.

REFERENCES

[1] Yazan Abu Farha, Alexander Richard, and Juergen Gall. 2018. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5343–5352.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations*.

[3] Andrea Bandini and José Zariffa. 2020. Analysis of the hands in egocentric vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[4] Alejandro Betancourt, Pietro Morerio, Carlo S Regazzoni, and Matthias Rauterberg. 2015. The evolution of first person vision methods: A survey. *IEEE Transactions on Circuits and Systems for Video Technology* 25, 5 (2015), 744–760.

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

[6] Seymour Epstein. 1998. Cognitive-experiential self-theory. In *Advanced personality*. 211–238.

[7] Alircza Fathi, Jessica K Hodgins, and James M Rehg. 2012. Social interactions: A first-person perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1226–1233.

[8] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. 2017. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation* 49 (2017), 401–411.

[9] Antonino Furnari, Sebastiano Battiato, and Giovanni Maria Farinella. 2018. Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation. In *Proceedings of the European Conference on Computer Vision*. 389–405.

[10] Antonino Furnari and Giovanni Maria Farinella. 2019. What Would You Expect? Anticipating Egocentric Actions with Rolling-Unrolling LSTMs and Modality Attention. In *Proceedings of the IEEE International Conference on Computer Vision*. 6251–6260.

[11] Iván González-Díaz, Jenny Benois-Pineau, Jean-Philippe Domenger, and Aymar de Rugy. 2018. Perceptually-guided Understanding of Egocentric Video Content: Recognition of Objects to Grasp. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. 434–441.

[12] Robert M Hamm. 1988. Clinical intuition and clinical analysis: expertise and the cognitive continuum. *Professional judgment: A reader in clinical decision making* (1988), 78–105.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[14] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*. 448–456.

[15] Takeo Kanade and Martial Hebert. 2012. First-person vision. *Proc. IEEE* 100, 8 (2012), 2442–2453.

[16] Hema S Koppula and Ashutosh Saxena. 2015. Anticipating human activities using object affordances for reactive robotic response. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 1 (2015), 14–29.

[17] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. 2014. A hierarchical representation for future action prediction. In *Proceedings of the European Conference on Computer Vision*. Springer, 689–704.

[18] Kang Li, Jie Hu, and Yun Fu. 2012. Modeling complex temporal composition of actionlets for activity prediction. In *Proceedings of the European conference on computer vision*. Springer, 286–299.

[19] Minghuang Ma, Haoqi Fan, and Kris M Kitani. 2016. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1894–1903.

[20] Antoine Miech, Ivan Laptev, Josef Sivic, Heng Wang, Lorenzo Torresani, and Du Tran. 2019. Leveraging the Present to Anticipate the Future in Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2915–2922.

[21] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. 2019. A survey on food computing. *Comput. Surveys* 52, 5 (2019), 1–36.

[22] Weiqing Min, Shuqiang Jiang, Shuhui Wang, Jitao Sang, and Shuhuan Mei. 2017. A delicious recipe analysis framework for exploring multi-modal recipes with various attributes. In *Proceedings of the ACM on Multimedia Conference*. 402–410.

[23] Thi-Hoa-Cuc Nguyen, Jean-Christophe Nebel, Francisco Florez-Revuelta, et al. 2016. Recognition of activities of daily living with egocentric vision: A review. *Sensors* 16, 1 (2016), 72.

[24] Hamed Pirsiavash and Deva Ramanan. 2012. Detecting activities of daily living in first-person camera views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2847–2854.

[25] Jean E Pretz. 2008. Intuition versus analysis: Strategy and experience in complex everyday problem solving. *Memory & cognition* 36, 3 (2008), 554–566.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. 91–99.

[27] Nicholas Rhinehart and Kris M Kitani. 2017. First-person activity forecasting with online inverse reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 3696–3705.

[28] Michael S Ryoo. 2011. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Proceedings of the IEEE International Conference on Computer Vision*. 1036–1043.

[29] Giuseppe Serra, Marco Camurri, Lorenzo Baraldi, Michela Benedetti, and Rita Cucchiara. 2013. Hand segmentation for gesture recognition in ego-vision. In *Proceedings of the 3rd ACM international workshop on Interactive multimedia on mobile & portable devices*. 31–36.

[30] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*. 568–576.

[31] Ekaterina H Spriggs, Fernando De La Torre, and Martial Hebert. 2009. Temporal segmentation and activity classification from first-person sensing. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Workshops*. 17–24.

[32] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4631–4640.

[33] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*. 4534–4542.

[34] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 98–106.

[35] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference on Computer Vision*. Springer, 20–36.

[36] Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. 2018. Egogesture: a new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia* 20, 5 (2018), 1038–1050.