Can Gaze Inform Egocentric Action Recognition?

Zehua Zhang
David Crandall
zehzhang@indiana.edu
djcran@indiana.edu
Indiana University Bloomington
Bloomington, Indiana, USA

Michael J. Proulx
Sachin S. Talathi
Abhishek Sharma
michaelproulx@fb.com
stalathi@fb.com
abhishek.sharma@fb.com
Reality Labs Research, Meta
Redmond, Washington, USA

ABSTRACT

We investigate the hypothesis that gaze-signal can improve egocentric action recognition on the standard benchmark, EGTEA Gaze++ dataset. In contrast to prior work where gaze-signal was only used during training, we formulate a novel neural fusion approach, Cross-modality Attention Blocks (CMA), to leverage gaze-signal for action recognition during inference as well. CMA combines information from different modalities at different levels of abstraction to achieve state-of-the-art performance for egocentric action recognition. Specifically, fusing the video-stream with optical-flow with CMA outperforms the current state-of-the-art by 3%. However, when CMA is employed to fuse gaze-signal with video-stream data, no improvements are observed. Further investigation of this counter-intuitive finding indicates that small spatial overlap between the network's attention-map and gaze ground-truth renders the gaze-signal uninformative for this benchmark. Based on our empirical findings, we recommend improvements to the current benchmark to develop practical systems for egocentric video understanding with gaze-signal.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**; **Neural networks**.

KEYWORDS

egocentric action recognition, deep neural networks, gaze, attention

ACM Reference Format:
Zehua Zhang, David Crandall, Michael J. Proulx, Sachin S. Talathi, and Abhishek Sharma. 2022. Can Gaze Inform Egocentric Action Recognition?. In *ETRA '22: 2022 Symposium on Eye Tracking Research and Applications, June 08–11, 2022, Seattle, WA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

We live in a complex multi-sensory environment and leverage multiple sources of information, such as vision, speech, touch, motion and/or smell, to effectively accomplish our daily tasks. Arguably, an advanced artificial intelligence system (AIS) should be able to mimic our ability to match human-level performance [Zhang et al. 2018, 2020b, 2019]. Among other tasks, egocentric video understanding is an important application in AR/VR. There have been prior attempts at fusing RGB video-stream, optical-flow and gaze-signal to improve egocentric action-recognition [Furnari and Farinella 2020; Kazakos et al. 2019; Li et al. 2020]. In addition, there is strong evidence for the usefulness of gaze-signal in egocentric action recognition from investigations by neuroscience research streams [Borji and Itti 2014; Henderson et al. 2013; Iqbal and Bailey 2004; Yarbus 2013]. Based on these findings, it is natural to expect that the current AIS would leverage gaze-signals along with RGB video-stream for egocentric action-recognition. However, none of the existing approaches leverage gaze-signals for egocentric action-recognition during inference because they all report degraded performance with gaze's inclusion, at the inference stage, due to noise in the gaze-signal [Huang et al. 2020; Li et al. 2020; Sudhakaran et al. 2019].

In order to resolve this contradiction between neuroscience and data-driven AIS observations, we first attempt to improve the current multi-modal fusion strategies that employ simple concatenation or element-wise summation/multiplication of high-level neural-network features. We hypothesize that this simple schema for fusion has several limitations, which can lead to inferior performance. For example, these simple fusion approaches lack the ability to capture the long-range spatio-temporal relationships which not only exist within the same modality, i.e. video-stream data, but also across different modalities, i.e., video-stream, optical flow, and gaze. Intuitively, such long-range dependencies contain useful information for feature fusion in egocentric videos. For example, in a video clip corresponding to the action of "taking a plate", fusion of features corresponding to hands and plates across all the frames within the same modality would be desirable. It would also be beneficial to perform cross-modality fusion by merging optical flow features of hands with appearance features of plates, and vice versa. Unfortunately, fusion approaches that simply concatenate or add features cannot afford adaptive feature updating, which has been shown to be effective in many single-modal learning tasks [Bertasius et al. 2021; Carion et al. 2020; Dosovitskiy et al. 2020; Vaswani et al. 2017; Wang et al. 2018]. As a result, models with simple fusion rely heavily on deep CNN-based neural-network architectures that

ETRA '22, June 08–11, 2022, Seattle, WA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9252-5/22/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

achieve long-range spatio-temporal fusion and adaptive feature update by first concatenating/adding/multiplying features and then stacking multiple CNN layers, which is not efficient.

Motivated by the shortcomings of the existing approaches for fusing multi-modal spatio-temporal signals, here we propose Cross-modality Attention (CMA) blocks as a novel approach. CMA leverages a self-attention mechanism [Vaswani et al. 2017] to first infer associations between cross-modal pairs of spatial-temporal signals to aggregate information from all the modalities. This step is followed by updating the features from each modality to incorporate the aggregated information through learned relation matrices. This design leads to a very flexible module that can be inserted into any place within neural networks and can be used to fuse any number of modalities.

We conducted experiments with the proposed CMA architecture for egocentric action-recognition, on the EGTEA-Gaze++ dataset [Li et al. 2020], to show it's efficacy for fusing multi-modal information (video, gaze, optical-flow). With the same standard backbone of I3D [Carreira and Zisserman 2017; Wang et al. 2018], models with CMA that fuse RGB-video with optical flow outperform all state-of-the-art methods as well as model variants with Non-Local Blocks [Wang et al. 2018] by significant margins. Motivated by this observation, we tried fusing gaze-signal with RGB-video using CMA to assess whether CMA-fusion is effective in tackling the reported issue of noise in gaze-signals. We observed that the action-recognition accuracy did not deteriorate with gaze as input, unlike all the existing approaches. This indicates that perhaps CMA can effectively learn to tackle the noise in gaze-signals. However, we did not observe any improvements to action-recognition accuracy. Further analysis led to the discovery that the CMA module is completely ignoring the gaze-signal input for final inference. Our investigations into this issue revealed that the spatial intersection of the raw gaze-point with an *information-theoretic optimal-attention map* is extremely low for all the training/testing samples, which essentially forces the network to treat raw gaze-signal as noise. This finding is consistent with the prior body of work [Li et al. 2020; Min and Corso 2021], and the implications offer additional insights into the reason behind this phenomenon. These findings should be used to guide the data collection protocols in the future and provide a sanity-check for leveraging gaze signal for attention.

In summary, we made these important contributions:

(1) We introduced a novel attention-based multi-modal fusion approach for spatio-temporal signal fusion and improved the current state-of-the-art on the standard benchmark for video and optical-flow fusion.
(2) We obtain information-theoretic optimal attention-maps for action-recognition tasks and show that the ground-truth gaze-fixation points for EGTEA Gaze++ dataset have extremely low intersection with the aforementioned attention maps, therefore, rendering the gaze-signals ineffective.
(3) Based on our observations, we discuss further improvements to egocentric dataset collection protocols.

2 RELATED WORK

In this section, we first provide a short summary of current approaches for egocentric action-recognition followed by a short review of popular multi-modal fusion approaches for the same.

Egocentric Video Action Recognition: The egocentric action-recognition task is gaining popularity due to its importance for AR experience. Spriggs et al. [Spriggs et al. 2009] leveraged video and wearable sensor data for daily activity segmentation and recognition. Kitani et al. [Kitani et al. 2011] proposed an unsupervised method for ego-action learning using a global motion descriptor. Fathi et al. [Fathi et al. 2012] demonstrated an ability to simultaneously recognize actions and gaze with a probabilistic generative model. While these early work relied on hand-designed features, recent methods are usually deep neural network models [Huang et al. 2020; Kapidis et al. 2019; Kazakos et al. 2019; Li et al. 2020; Lu et al. 2019; Ma et al. 2016; Min and Corso 2021; Ryoo et al. 2015; Singh et al. 2016; Sudhakaran et al. 2019; Sudhakaran and Lanz 2018]. Following from their previous work [Fathi et al. 2012], Li et al. [Li et al. 2020] proposed an I3D-based model that jointly estimates gaze and classifies actions. The estimated gaze is then used for re-weighting the features that are fed to the action recognition head. Min et al. [Min and Corso 2021] adopted a similar workflow but employed a residual connection for feature updating. Sudhakaran et al. [Sudhakaran et al. 2019] proposed to incorporate soft attention to LSTM models. In contrast to these methods, our work improves egocentric video action recognition by proposing a novel Cross-modality Attention Block that can better fuse information from each different modality of input signal streams.

Multi-Modal Fusion: Although multi-modal fusion is an important problem, which has been studied a lot in other fields [Chen et al. 2017; Djuric et al. 2020; Feichtenhofer et al. 2016; Ku et al. 2018; Liang et al. 2019, 2020; Luo et al. 2018; Simonyan and Zisserman 2014; Zhang et al. 2020a], it has not received much attention for egocentric video understanding. Most methods adopt simple concatenation or element-wise summation / multiplication for fusing video and optical-flow. In contrast, Kazakos et al. [Kazakos et al. 2019] fused RGB clips, optical flows and audio signals within a range of temporal offsets for egocentric video action recognition. All of these existing approaches do not use gaze-signal as input and report loss in accuracy if they do. Some recent approaches also leverage attention-transformers for multi-modal learning, for example: [Gheini et al. 2021] uses cross-attention to avoid fine-tuning for language translation models; [Mohla et al. 2020] uses attention from Lidar and content from spectral imaging to combine them for image-segmentation; and [Ye et al. 2019] uses attention-transformers to segment out the object described in the form of text from a given image. CMA, on the other hand, infers the spatio-temporal relationships across different modalities by combining information from all the modalities via attention-transformers [Vaswani et al. 2017] and adaptively updates features for each modality to disseminate the global information from all the modalities. The closest technique to CMA is presented in [Bhatti et al. 2021], where Electrocardiogram (ECG) and Electrodermal Activity (EDA) signals are fused together by a transformer-like module for emotion-recognition. The main difference from CMA is in their use of asymmetric-attention modules, which requires domain-knowledge to decide where attention comes from; CMA doesn't require such knowledge.

3 CROSS-MODALITY ATTENTION BLOCKS (CMA)

As the core technical contribution of this work, CMA is designed to capture cross-modality relationships across multiple spatio-temporal input signal streams in the feature space at each spatial and temporal location of the feature map followed by updating the representations of each modality adaptively according to the relationship matrix.

CMA builds on the attention-transformer networks [Vaswani et al. 2017] and hence familiarity with them would be necessary to follow the details of CMA. The input to a CMA block are the feature representations, $\{F_1, F_2, ..., F_n\}$, where $F_i \in R^{T\times H\times W\times C_i}$, $i = 1, 2, ..., n$ for the given modalities, $\{M_1, M_2, ..., M_n\}$. Here, T, H, W, C denote the number of frames (or time-steps), height, width, and channels, respectively. For spatio-temporal data, such as video-clips, the features are 4-D tensors, with one convolution feature map for each time-step, see Figure 1.

Differing from the traditional self-attention [Vaswani et al. 2017; Wang et al. 2018] networks, we first aggregate *global* information from all the features through a learned mapping Γ:

$$\tilde{F} = \Gamma(F_1, F_2, ..., F_n) \tag{1}$$

to yield $\tilde{F} \in R^{T\times H\times W\times C}$, respectively. We note that the proposed CMA architecture requires the spatio-temporal dimensions of F_is and $\tilde{F}$ to be the same.

Once we have obtained the aggregated feature representation, $\tilde{F}$, we reshape it to 2D-tensors of shape $THW \times C$. It is done to extract the per time-step per spatial-location feature representation, $\tilde{f}_{t,x,y} = \tilde{F}[t, x, y, :]$, which simply translates into representing an image-region centered at the 2D gaze-location, (x, y) in the t^{th} frame of the input-video. Now, we follow the attention-computation mechanism of attention-networks and for each modality, M_i, compute the modality-specific attention matrix, $A_i \in R^{THW\times THW}$, which is described below.

To estimate A_i, a modality-specific key matrix $K_i \in R^{THW\times C_A}$ and a query matrix $Q_i \in R^{THW\times C_A}$ are first computed by applying a single linear projection layer on $\tilde{F}'$:

$$K_i = h_i(\tilde{F}') \tag{2}$$

$$Q_i = g_i(\tilde{F}') \tag{3}$$

where h_i and g_i are modality-specific linear projection layers with learnable weight and bias.

A_i can be then computed based on any similarity metrics $\mathcal{S}(\cdot, \cdot)$, followed by proper normalization with $C(\cdot)$:

$$A_i = C(\mathcal{S}(K_i, Q_i)) \tag{4}$$

There are multiple alternatives for $\mathcal{S}(\cdot, \cdot)$ and $C(\cdot)$ [Wang et al. 2018]. One possible way is embedded Gaussian, where $\mathcal{S}(K_i, Q_i) = Q_i K_i^T$ and $C(\cdot)$ is softmax normalization over each row of $\mathcal{S}(K_i, Q_i)$.

In order to adaptively update F_i with A_i, we first reshape F_i to $F_i' \in R^{THW\times C_i}$ and project each feature vector to a new space with dimension C_V with trainable linear projection $f_i(\cdot)$, resulting in a value matrix $V_i = f_i(F_i') \in R^{THW\times C_V}$. The entry at the m_{th} row and n_{th} column of A_i describe the relationship between the m_{th} and the n_{th} spatial-temporal location of modality i and therefore the feature updating can be performed as:

$$V_i' = A_i V_i \tag{5}$$

V_i' is then projected back to the feature space of dimension C_i by another trainable linear projection $\phi(\cdot)$, $V_i'' = \phi(V_i') \in R^{THW\times C_i}$, and reshaped to $V_i''' \in R^{T\times H\times W\times C_i}$, which will be finally added to F_i to form a residual connection for feature updating:

$$F_i'' = F_i + V_i''' \tag{6}$$

In summary, we learn modality-specific spatio-temporal attention mappings derived from the aggregated information, which affords fusing information across different modalities and effectively disseminating them to the individual modalities re-weighted through spatio-temporal attention scores. The proposed CMA blocks can be inserted at any layer in the model for multi-modal fusion. Besides, it can be used to fuse features from any number of modalities. CMA is a generalization of Non-Local Networks [Wang et al. 2018] for more than one input stream.

4 RESULTS

We conducted experiments to assess the efficacy of CMA towards egocentric action-recognition on the open-source EGTEA-Gaze++ [Li et al. 2020] dataset because it comes with egocentric video and gaze-directions, and optical-flow can be computed from the videos. In this section, we show results of using CMA for fusing video with optical-flow, and video with gaze-signal, followed by additional analysis to understand the behaviour of CMA block for the latter. The EGTEA-Gaze++ dataset [Li et al. 2020] contains 10321 videos belonging to one of the 106 action classes. A few examples of the action-categories are "cut onion", "wash pot", "open fridge" etc. These videos are recorded by first-person cameras capturing meal preparation scenarios in a kitchen with per-frame gaze-directions projected in the image-space. See [Li et al. 2020] for more details on the data-capture protocol.

4.1 RGB Clips and Optical Flows

We first injected CMA into a two-stream Inception-V1 [Szegedy et al. 2015] and I3D [Carreira and Zisserman 2017] with RGB clips and optical flows as input. Optical flows were obtained and preprocessed following [Li et al. 2020]. RGB clip input is also normalized accordingly. The input to each stream is a 32-frame sequence sampled every two frames. For clips that don't have enough frames to fill-up 32 frames, the last frame is repeated. Γ is an element-wise summation and we adopt embedded Gaussian for $\mathcal{S}(\cdot, \cdot)$ and $C(\cdot)$. The consensus head is global average pooling of features from both streams and element-wise summation, followed by a dropout layer of ratio 0.7 and a linear classification layer for final action recognition.

Following the protocol from [Li et al. 2020], the I3D backbone is initialized with weights from a pre-trained model on Kinetics-400 [Kay et al. 2017] and CMA is initialized according to [Wang et al. 2018]. A cross-entropy loss between the output action probability, obtained by softmax operation, and the ground-truth action labels is minimized during training. The initial learning rate is 0.03 and

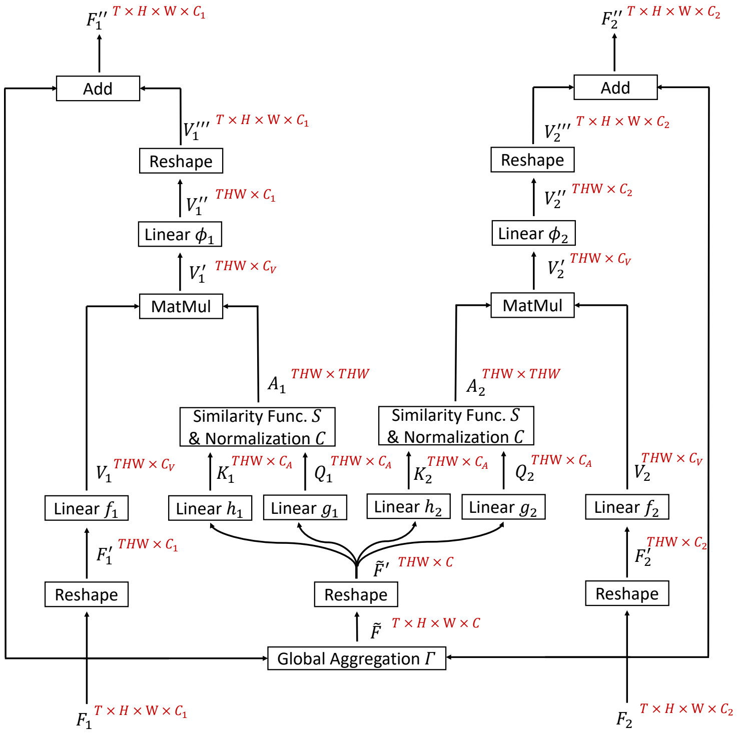

Figure 1: Illustration of a CMA block with two modalities as the input. The dark red text besides a variable indicates its dimension. Here, T, H, W, C denote the number of frames (or time-steps), height, width, and channels, respectively.

the batch size is 64. During inference, 10 clips are taken from each video and the action probabilities of all the clips are averaged to get the final action recognition score. Each clip has 32 frames sampled every two frames. The starting timestamps of each clip are evenly sampled from 0 to $L - 64$, L is the total length of the video. When a video contains fewer frames than needed, the last frame will be repeated and 10 identical clips are used for inference.

We compare CMA-models against other state-of-the-art methods and Non-local Blocks [Wang et al. 2018] with regards to average instance accuracy and mean-class accuracy over 3 train/test splits, see [Li et al. 2020] for split information. The results are reported in Table 1. We observe that I3D with 1 CMA block at different network depths outperforms other state-of-the-art methods by a large margin in terms of both instance accuracy and mean-class accuracy. For a fair comparison, we use the same backbone of Inception-V1 [Szegedy et al. 2015] I3D [Carreira and Zisserman 2017] and the same input modalities as Min et al. [Min and Corso 2021], which is the current state-of-the-art. Moreover, [Min and Corso 2021] requires gaze-signal during training, while our results only use video and optical-flow, both during training and inference.

We further note that the performance gain by placing 1 CMA-block after ConvBlock_4 is higher than placing 1 CMA after ConvBlock_5. We hypothesize that the degradation in performance could be due to the large size receptive field for features in ConvBlock_5, which leads to coarse spatial-temporal attention-mapping by CMA and therefore is less beneficial. We also compared with model variants that have Non-local Blocks [Wang et al. 2018] inserted to the same place of both streams. Models with CMA outperform variants with Non-local Blocks significantly, supporting the effectiveness of CMA for multi-modal learning.

4.2 RGB Clips and Gaze

Unlike the optical-flow stream, which could be computed from the video-stream and available at every time-step, gaze is not available for all the frames. Whenever the gaze is available, it’s represented as a 2D Gaussian heat-map with its center at the gaze-point and its variance σ equal to 40 pixels. When there is no valid gaze available, we use a 2D uniform distribution to indicate to the network that the gaze can be anywhere with uniform probability. The RGB stream is processed by a ResNet50 I3D as used in [Wang et al. 2018] initialized with weights pretrained on Kinetics-400 [Kay et al. 2017]. The input RGB clips are normalized by subtracting the mean and dividing by standard deviation. The network for the gaze-stream processing is randomly initialized I3D with half of the depth (number of layers) and one fourth of the width (channel number) as the RGB stream. We hypothesize that such a 3D-convolution based gaze stream can eliminate noise from the raw gaze data and hallucinate gaze for those frames without valid gaze recorded based on gaze in adjacent frames besides processing gaze and extracting information. The element-wise summation operation in Γ and the consensus head is replaced by concatenation for this experiment. Other details are the same as in Sec. 4.1.

Table 1: Experiments on egocentric action recognition tasks on EGTEA-Gaze++ with input modalities of RGB clips and optical flows. I3D with 1 CMA significantly outperforms other state-of-the-art methods as well I3D with Non-local blocks at the same place of both streams in terms of both instance accuracy and mean class accuracy.

Method	Instance Acc (%)	Mean Class Acc (%)
Li et al. [Li et al. 2020]	-	55.03
Sudhakaran et al. [Sudhakaran and Lanz 2018]	60.76	-
LSTA [Sudhakaran et al. 2019]	61.86	-
MCN [Huang et al. 2020]	-	55.63
Kapidis et al. [Kapidis et al. 2019]	66.59	59.44
Lu et al. [Lu et al. 2019]	68.60	60.54
Min et al. [Min and Corso 2021]	69.58	62.84
I3D [Carreira and Zisserman 2017] with 1 Non-local Block [Wang et al. 2018] after ConvBlock_4 of each stream	72.09	64.22
I3D [Carreira and Zisserman 2017] with 1 Non-local Block [Wang et al. 2018] after ConvBlock_5 of each stream	71.32	63.43
I3D [Carreira and Zisserman 2017] with 1 CMA after ConvBlock_4	**72.95**	**64.65**
I3D [Carreira and Zisserman 2017] with 1 CMA after ConvBlock_5	**71.86**	**63.99**

Table 2: Experiments on egocentric action recognition tasks on EGTEA with input modalities of RGB clips and gaze.

Stream(s)	CMA	Instance Acc (%)
RGB	-	64.34
RGB & Gaze	1 after Res5	64.89
RGB & Gaze	1 after Res4 & 1 after Res5	66.12

As the results in Table 2 show, when compared to models with only a RGB stream, the performance seems to be improved by using gaze as additional input and leveraging CMA for multi-modal fusion. We also found that adding more than one CMA block can slightly improve the performance. In order to tease apart the improvement offered due to CMA block for attention-modeling and gaze-signal as input, we dropped gaze-information from all the frames and fed 2D uniform distribution with each frame during testing. We hoped to observe some deterioration in the accuracy, which would translate into the improvement coming from gaze. Surprisingly, however, the performance didn't change at all! Further analysis of the neural activation coming from the gaze-network towards the final classification-head confirmed that the model simply learns to ignore gaze signals altogether. Essentially, CMA block simply reduces to a Non-Local network block on video-stream, which is not the result we expected. We hypothesize that gaze signals were ignored because many recorded gaze points, though valid and accurate, do not fall within the regions that are related to the action being performed. This can happen in many situations; for example, when we have a reliable memory about the environment and therefore do not need to stare at the object that is being manipulated, or when we are very experienced with the task at hand and therefore our gaze might just precede the hand, if at all. Under such situations, gaze-signal can actually distract the network attention away from the optimal location.

In order to verify our hypothesis, we first computed the Class-Activation Map [Zhou et al. 2016], CAM for short, for each frame corresponding to the known ground-truth action category. CAM obtains the input image-regions, through back-propagation from the target all the way to the input image, that are strongly correlated with the correct action category. Intuitively, CAM can be thought of as class-specific saliency map indicating the regions that contribute the most towards the correct ground-truth. We confirm the above intuition by using the CAM saliency-map for weighted pooling before the action-classification head and obtain 95% instance accuracy, which clearly validates our claim. Since the CAM saliency-map is a pixel-wise score, $s_{[x,y]} \in [0,1]$, we first threshold it at different score-levels, t, to obtain binary masks, $CAM^t_{[x,y]} = \mathbf{1}(CAM_{[x,y]} \geq t)$. Finally, we compute the fraction of times the ground-truth gaze falls within the CAM binary-mask for the best model, an event we refer to as a *hit*, that we obtained in Sec.4.1 (I3D with 1 CMA after ConvBlock_4 with RGB and flows as input). The results of hit-rate with different score-thresholds are shown in Table 3. From the table, we can clearly observe that even at $t = 0.5$, the hit-rate is merely 48.6%, which means that the ground-truth gaze will not result in reasonable attention 50% of the time. This analysis, for the first time, explains the counter-intuitive observation of decreasing accuracy with the use of gaze-signal as input, which was widely reported by multiple works in the past on this dataset [Li et al. 2020; Min and Corso 2021]. While our CMA block could not overcome this problem, we hope that the insights uncovered in this article could be useful for setting the future directions of research in this area.

4.3 Advancing the Value of Gaze for Egocentric Video Understanding

The findings we report here draw attention to a number of improvements that could be made to the current benchmark. One issue was that many gaze points did not fall into a region of interest defined by the action. This does not mean that gaze is uninformative per se, but that another factor such as expertise or the nature of the task [Hadnett-Hunter et al. 2019] could influence whether gaze needs to focus directly on the task or not. For example, a well-practiced task might be possible to carry out with peripheral vision [Rosenholtz 2016], whereas a difficult or unfamiliar task might be more reliant on direct gaze rather than peripheral vision and covert attention [Matthis et al. 2018]. Another issue could be the dynamics of eye gaze and hand movements, such that the temporal co-registration of gaze might lead the hands at different intervals depending on the nature of the action being performed, such that a shifting selection window might be required [de Vries et al. 2018; Land 2006].

Table 3: The variation of ground-truth gaze-point hit-rate w.r.t. the threshold for score-level in CAM maps.

Threshold	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
Frequency (%)	48.6	48.1	47.7	47.2	46.7	46.2	45.5	44.7	43.7	42.0

How might a benchmark, such as the EGTEA-Gaze++ dataset, be improved to address such issues? The classes of action-categories provided and the nature of the person performing actions in each video could provide additional classifications to support a more nuanced analysis of gaze. For example, some tasks might be more dangerous or difficult than others (say, cutting versus washing) and the tight coupling between gaze and the action should likely reflect that; expertise migth be key as well. A future dataset that explicitly included tasks that vary in difficulty would be crucial, as well as complementary data on the expertise of the person carrying out the task. The latter could be provided by having actions recorded more than once by the same person, such that changes in interaction and the value of gaze might change over time and with experience.

5 CONCLUSION

We investigated whether gaze-signals can be informative for egocentric video understanding tasks, particularly when used during inference. We proposed Cross-modality Attention Blocks (CMA) for multi-modal fusion for solving egocentric video action recognition tasks to leverage gaze-signals and assess whether that is informative. We show that CMA-fused RGB video-stream and optical flow outperform state-of-the-art methods and Non-local Blocks of similar configuration. We also applied CMA to RGB clips and gaze signal fusion, but our results and analyses indicate that the spatial incoherence between the ground-truth gaze-data with regards to class-attention-maps for correct classification is too strong and forces the CMA block to treat ground-truth gaze as noise, therefore, learning to ignore gaze-signals altogether. Determining whether, and when, gaze-signals might provide an informative contribution to models of egocentric video understanding tasks remains a challenge for continuing research.

REFERENCES

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding? *arXiv preprint arXiv:2102.05095* (2021).

Anubhav Bhatti, Behnam Behinaein, Dirk Rodenburg, Paul Hungler, and Ali Etemad. 2021. Attentive Cross-modal Connections for Deep Multimodal Wearable-based Emotion Recognition. *CoRR* abs/2108.02241 (2021).

Ali Borji and Laurent Itti. 2014. Defending Yarbus: Eye movements reveal observers' task. *Journal of vision* 14, 3 (2014), 29–29.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*. Springer, 213–229.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. 2017. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1907–1915.

S de Vries, R Huys, and PG Zanone. 2018. Keeping your eye on the target: eye–hand coordination in a repetitive Fitts' task. *Experimental Brain Research* 236, 12 (2018), 3181–3190.

Nemanja Djuric, Henggang Cui, Zhaoen Su, Shangxuan Wu, Huahua Wang, Fang-Chieh Chou, Luisa San Martin, Song Feng, Rui Hu, Yang Xu, et al. 2020. Multixnet: Multiclass multistage multimodal motion prediction. *arXiv preprint arXiv:2006.02000* (2020).

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

Alireza Fathi, Yin Li, and James M Rehg. 2012. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*. Springer, 314–327.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1933–1941.

Antonino Furnari and Giovanni Farinella. 2020. Rolling-Unrolling LSTMs for Action Anticipation from First-Person Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 1–1. https://doi.org/10.1109/TPAMI.2020.2992889

Mozhdeh Gheini, Xiang Ren, and Jonathan May. 2021. Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation. arXiv:2104.08771 [cs.CL]

Jacob Hadnett-Hunter, George Nicolaou, Eamonn O'Neill, and Michael Proulx. 2019. The effect of task on visual attention in interactive virtual environments. *ACM Transactions on Applied Perception (TAP)* 16, 3 (2019), 1–17.

John M Henderson, Svetlana V Shinkareva, Jing Wang, Steven G Luke, and Jenn Olejarczyk. 2013. Predicting cognitive state from eye movements. *PloS one* 8, 5 (2013), e64937.

Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. 2020. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing* 29 (2020), 7795–7806.

Shamsi T Iqbal and Brian P Bailey. 2004. Using eye gaze patterns to identify user tasks. In *The Grace Hopper Celebration of Women in Computing*, Vol. 4. 2004.

Georgios Kapidis, Ronald Poppe, Elsbeth van Dam, Lucas Noldus, and Remco Veltkamp. 2019. Multitask learning to improve egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. 2019. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5492–5501.

Kris M Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. 2011. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR 2011*. IEEE, 3241–3248.

Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. 2018. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1–8.

Michael F Land. 2006. Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research* 25, 3 (2006), 296–324.

Yin Li, Miao Liu, and James M. Rehg. 2020. In the Eye of the Beholder: Gaze and Actions in First Person Video. arXiv:2006.00626 [cs.CV]

Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. 2019. Multi-task multi-sensor fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7345–7353.

Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. 2020. Pnpnet: End-to-end perception and prediction with tracking in the loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11553–11562.

Minlong Lu, Danping Liao, and Ze-Nian Li. 2019. Learning Spatiotemporal Attention for Egocentric Action Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

Wenjie Luo, Bin Yang, and Raquel Urtasun. 2018. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 3569–3577.

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.

Jonathan Samir Matthis, Jacob L Yates, and Mary M Hayhoe. 2018. Gaze and the control of foot placement when walking in natural terrain. *Current Biology* 28, 8 (2018), 1224–1233.

Kyle Min and Jason J. Corso. 2021. Integrating Human Gaze Into Attention for Egocentric Activity Recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1069–1078.

Satyam Mohla, Shivam Pande, Biplab Banerjee, and Subhasis Chaudhuri. 2020. FusAtNet: Dual Attention Based SpectroSpatial Multimodal Fusion Network for Hyperspectral and LiDAR Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Ruth Rosenholtz. 2016. Capabilities and limitations of peripheral vision. *Annual Review of Vision Science* 2 (2016), 437–457.

Michael S Ryoo, Brandon Rothrock, and Larry Matthies. 2015. Pooled motion features for first-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 896–904.

Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199* (2014).

Suriya Singh, Chetan Arora, and CV Jawahar. 2016. First person action recognition using deep learned descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2620–2628.

Ekaterina H Spriggs, Fernando De La Torre, and Martial Hebert. 2009. Temporal segmentation and activity classification from first-person sensing. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 17–24.

Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. 2019. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9954–9963.

Swathikiran Sudhakaran and Oswald Lanz. 2018. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. *arXiv preprint arXiv:1807.11794* (2018).

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7794–7803.

Alfred L Yarbus. 2013. *Eye movements and vision*. Springer.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-Modal Self-Attention Network for Referring Image Segmentation. *CoRR* abs/1904.04745 (2019).

Zehua Zhang, Sven Bambach, Chen Yu, and David J Crandall. 2018. From Coarse Attention to Fine-Grained Gaze: A Two-stage 3D Fully Convolutional Network for Predicting Eye Gaze in First Person Video. In *British Machine Vision Conference (BMVC)*.

Zhishuai Zhang, Jiyang Gao, Junhua Mao, Yukai Liu, Dragomir Anguelov, and Congcong Li. 2020a. Stinet: Spatio-temporal-interactive network for pedestrian detection and trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11346–11355.

Zehua Zhang, Ashish Tawari, Sujitha Martin, and David Crandall. 2020b. Interaction Graphs for Object Importance Estimation in On-road Driving Videos. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 8920–8927.

Zehua Zhang, Chen Yu, and David Crandall. 2019. A Self Validation Network for Object-Level Human Attention Estimation. In *Advances in Neural Information Processing Systems*. 14702–14713.

B. Zhou, A. Khosla, Lapedriza. A., A. Oliva, and A. Torralba. 2016. Learning Deep Features for Discriminative Localization. *CVPR* (2016).