EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition

Evangelos Kazakos[1] Arsha Nagrani[2] Andrew Zisserman[2] Dima Damen[1]
[1]Visual Information Lab, University of Bristol [2]Visual Geometry Group, University of Oxford

Abstract

We focus on multi-modal fusion for egocentric action recognition, and propose a novel architecture for multi-modal temporal-binding, i.e. the combination of modalities within a range of temporal offsets. We train the architecture with three modalities – RGB, Flow and Audio – and combine them with mid-level fusion alongside sparse temporal sampling of fused representations. In contrast with previous works, modalities are fused before temporal aggregation, with shared modality and fusion weights over time. Our proposed architecture is trained end-to-end, outperforming individual modalities as well as late-fusion of modalities.

We demonstrate the importance of audio in egocentric vision, on per-class basis, for identifying actions as well as interacting objects. Our method achieves state of the art results on both the seen and unseen test sets of the largest egocentric dataset: EPIC-Kitchens, on all metrics using the public leaderboard.

1. Introduction

With the availability of multi-sensor wearable devices (e.g. GoPro, Google Glass, Microsoft Hololens, MagicLeap), egocentric audio-video recordings have become popular in many areas such as extreme sports, health monitoring, life logging, and home automation. As a result, there has been a renewed interest from the computer vision community on collecting large-scale datasets [8, 35] as well as developing new or adapting existing methods to the first-person point-of-view scenario [9, 17, 21, 32, 44, 46].

In this work, we explore audio as a prime modality to provide complementary information to visual modalities (appearance and motion) in egocentric action recognition. While audio has been explored in video understanding in general [2, 3, 5, 6, 11, 23, 27–29, 34] the egocentric domain in particular offers rich sounds resulting from the interactions between hands and objects, as well as the close proximity of the wearable microphone to the undergoing action. Audio is a prime discriminator for some actions (e.g. 'wash', 'fry') as well as objects within actions (e.g. 'put plate' vs 'put bag'). At times, the temporal progression (or change) of sounds can separate visually ambiguous actions (e.g. 'open tap' vs 'close tap'). Audio can also capture actions that are out of the wearable camera's field of view, but audible (e.g. 'eat' can be heard but not seen). Conversely, other actions are *sound-less* (e.g. 'wipe hands') and the wearable sensor might capture irrelevant sounds, such as talking or music playing in the background. The opportunities and challenges of incorporating audio in egocentric action recognition allow us to explore new multi-sensory fusion approaches, particularly related to the potential *temporal asynchrony* between the action's appearance and the discriminative audio signal – the main focus of our work.

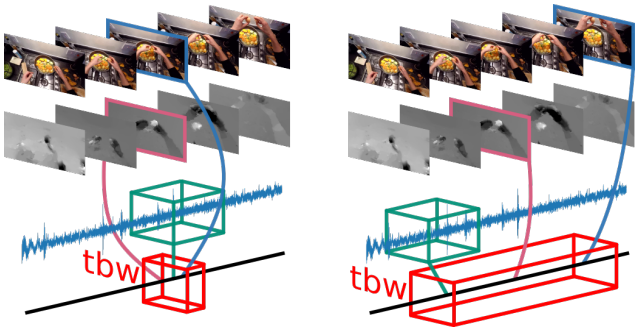

Figure 1: As the width of the temporal binding window increases (left to right), modalities (appearance, motion and audio) are fused with varying temporal shifts.

While several multi-modal fusion architectures exist for action recognition, current approaches perform temporal aggregation *within* each modality *before* modalities are fused [22, 42] or embedded [23]. Works that do fuse inputs before temporal aggregation, e.g. [10], do so with inputs synchronised across modalities. In Fig. 1, we show an example of 'breaking an egg into a pan' from the EPIC-Kitchens dataset. The distinct sound of cracking the egg, the motion of separating the egg and the change in appearance of the egg occur at different frames/temporal positions within the video. Approaches that fuse modalities with synchronised input would thus be limited in their ability

to learn such actions. In this work, we explore fusing inputs within a Temporal Binding Window (TBW) (Fig 1), allowing the model to train using asynchronous inputs from the various modalities. Evidence in neuroscience and behavioural sciences points at the presence of such a TBW in humans [30, 41]. The TBW offers a "range of temporal offsets within which an individual is able to perceptually bind inputs across sensory modalities" [39]. This is triggered by the gap in the biophysical time to process different senses [25]. Interestingly, the width of the TBW in humans is heavily task-dependant, shorter for simple stimuli such as flashes and beeps and intermediate for complex stimuli such as a hammer hitting a nail [41].

Combining our explorations into audio for egocentric action recognition, and using a TBW for asynchronous modality fusion, our contributions are summarised as follows. First, an end-to-end trainable mid-level fusion Temporal Binding Network (TBN) is proposed[1]. Second, we present the first audio-visual fusion attempt in egocentric action recognition. Third, we achieve state-of-the-art results on the EPIC-Kitchens public leaderboards on both seen and unseen test sets. Our results show (i) the efficacy of audio for egocentric action recognition, (ii) the advantage of mid-level fusion within a TBW over late fusion, and (iii) the robustness of our model to background or irrelevant sounds.

2. Related Work

We divide the related works into three groups: works that fuse visual modalities (RGB and Flow) for action recognition (AR), works that fuse modalities for egocentric AR in particular, and finally works from the recent surge in interest of audio-visual correspondence and source separation.

Visual Fusion for AR: By observing the importance of spatial and temporal features for AR, two-stream (appearance and motion) fusion has become a standard technique [10, 36, 42]. *Late fusion*, first proposed by Simonyan and Zisserman [36], combines the streams' independent predictions. Feichtenhofer *et al*. [10] proposed *mid-level fusion* of the spatial and temporal streams, showing optimal results by combining the streams after the last convolutional layer. In [7], 3D convolution for spatial and motion streams was proposed, followed by late fusion of modalities. All these approaches do not model the temporal progression of actions, a problem addressed by [42]. Temporal Segment Networks (TSN) [42] perform sparse temporal sampling followed by temporal aggregation (averaging) of softmax scores across samples. Each modality is trained independently, with late fusion of modalities by averaging their predictions. Follow-up works focus on pooling for temporal aggregation, still training modalities independently [13, 45]. Modality fusion before temporal aggregation was proposed in [18], where the appearance of the current frame is fused with 5 uniformly sampled motion frames, and vice versa, using two temporal models (LSTM). While their motivation is similar to ours, their approach focuses on using predefined asynchrony offsets between two modalities. In contrast, we relax this constraint and allow fusion from any random offset within a temporal window, which is more suitable for scaling up to many modalities.

Fusion in Egocentric AR: Late fusion of appearance and motion has been frequently used in egocentric AR [8, 24, 38, 40], as well as extended to additional streams aimed at capturing egocentric cues [21, 37, 38]. In [21], the spatial stream segments hands and detects objects. The streams are trained jointly with a triplet loss on objects, actions and activities, and fused through concatenation. [37] uses head motion features, hand masks, and saliency maps, which are stacked and fed to both a 2D and a 3D ConvNet, and combined by late fusion. All previous approaches have relied on small-scale egocentric datasets, and none utilised audio for egocentric AR.

Audio-Visual Learning: Over the last three years, significant attention has been paid in computer vision to an underutilised and readily available source of information existing in video: the audio stream [2, 3, 5, 6, 11, 23, 27–29, 34]. These fall in one of four categories: i) *audio-visual representation learning* [2, 5, 6, 23, 28, 29], ii) *sound-source localisation* [3, 28, 34], iii) *audio-visual source separation* [11, 28] and (iv) *visual-question answering* [1]. These approaches attempt fusion [2, 28] or embedding into a common space [3, 6, 26]. Several works sample the two modalities with temporal shifts, for learning better synchronous representations [16, 28]. Others sample within a 1s temporal window, to learn a correspondence between the modalities, e.g. [2, 3]. Of these works, [16, 28] note this audio-visual representation learning could be used for AR, by pretraining on the self-supervised task and then fine-tuning for AR.

Fusion for AR using three modalities (appearance, motion and audio) has been explored in [43], employing late-fusion of predictions, and [19, 20] using attention to integrate local features into a global representation. Tested on UCF101, [43] shows audio to be the least informative modality for third person action recognition (16% accuracy for audio compared to 80% and 78% for spatial and motion). A similar conclusion was made for other third-person datasets (AVA [12] and Kinetics [19, 20]).

In this work, we show audio to be a competitive modality for egocentric AR on EPIC-Kitchens, achieving comparable performance to appearance. We also demonstrate that audio-visual modality fusion in egocentric videos improves the recognition performance of both the action and the accompanying object.

[1] Code at: http://github.com/ekazakos/temporal-binding-network

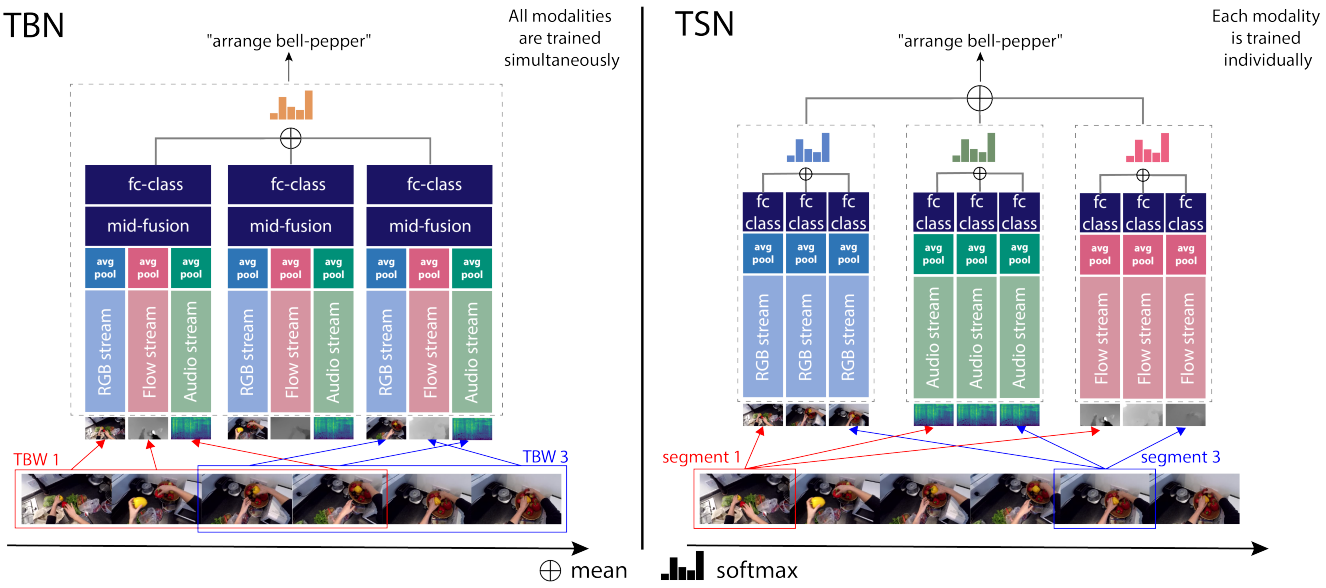

Figure 2: **Left:** our proposed Temporal Binding Network (TBN). Modalities are sampled within a TBW, and modality-specific weights (same colour) are shared amongst different inputs. Modalities are fused with mid-level fusion and trained jointly. Predictions from multiple TBWs, possibly overlapping, are averaged. **Right:** TSN [42] with an additional audio stream performing *late* fusion. Modalities are trained independently. Note that while in TSN a prediction is made for each modality, TBN produces a single prediction per TBW after fusing all modality representations. Best viewed in colour.

3. The Temporal Binding Network

Our goal is to find the optimal way to fuse multiple modality inputs while modelling temporal progression through sampling. We first explain the general notion of temporal binding of multiple modalities in Sec 3.1, then detail our architecture in Sec 3.2.

3.1. Multimodal Temporal Binding

Consider a sequence of samples from one modality in a video stream, $m_i = (m_{i1}, m_{i2}, \cdots, m_{iT/r_i})$ where T is the video's duration and r_i is the modality's framerate (or frequency of sampling). Input samples are first passed through unimodal feature extraction functions f_i. To account for varying representation sizes and frame-rates, most multi-modal architectures apply pooling functions G to each modality in the form of average pooling or other temporal pooling functions (e.g. maximum or VLAD [15]), before attempting multimodal fusion.

Given a pair of modalities m_1 and m_2, the final class predictions for a video are hence obtained as follows:

$$y = h\big(G(f_1(m_1)), G(f_2(m_2))\big) \tag{1}$$

where f_1 and f_2 are unimodal feature extraction functions, G is a temporal aggregation function, h is the multimodal fusion function and y is the output label for the video. In such architectures (e.g. TSN [42]), modalities are temporally aggregated for a prediction before different modalities are fused; this is typically referred to as 'late fusion'.

Conversely, multimodal fusion can be performed at *each* time step as in [10]. One way to do this would be to synchronise modalities and perform a prediction at *each* timestep. For modalities with matching frame rates, synchronised multi-modal samples can be selected as (m_{1j}, m_{2j}), and fused according to the following equation:

$$y = h\big(G(f_{sync}(m_{1j}, m_{2j}))\big) \tag{2}$$

where f_{sync} is a multimodal feature extractor that produces a representation for each time step j, and G then performs temporal aggregation over all time steps. When frame rates vary, and more importantly so do representation sizes, only approximate synchronisation can be attempted,

$$y = h\big(G(f_{sync}(m_{1j}, m_{2k}))\big) \quad : k = \lceil \frac{jr_2}{r_1} \rceil \tag{3}$$

We refer to this approach as 'synchronous fusion' where synchronisation is achieved or approximated.

In this work, however, we propose fusing modalities within temporal windows. Here modalities are fused within a range of temporal offsets, with all offsets constrained to lie within a finite time window, which we henceforth refer

to as a temporal binding window (TBW). Formally,

$$y = h\big(G(f_{tbw}(m_{1j}, m_{2k}))\big) \quad : k \in [\lceil \frac{jr_2}{r_1} - b \rceil, \lceil \frac{jr_2}{r_1} + b \rceil] \tag{4}$$

where f_{tbw} is a multimodal feature extractor that combines inputs within a binding window of width $\pm b$. Interestingly, as the number of modalities increases, say from two to three modalities, the TBW representation allows fusion of modalities each with different temporal offsets, yet within the same binding window $\pm b$:

$$y = h\big(G(f_{tbw}(m_{1j}, m_{2k}, m_{3l})\big) : k \in [\lceil \frac{jr_2}{r_1} - b \rceil, \lceil \frac{jr_2}{r_1} + b \rceil] \\ : l \in [\lceil \frac{jr_3}{r_1} - b \rceil, \lceil \frac{jr_3}{r_1} + b \rceil] \tag{5}$$

This formulation hence allows a large number of different inputs combinations to be fused. This is different from proposals that fuse inputs over predefined temporal differences (e.g. [18]). Sampling within a temporal window allows fusing modalities with various temporal shifts, *up to* the temporal window width $\pm b$. This: 1) enables straightforward scaling to multiple modalities with different frame rates, 2) allows training with a variety of temporal shifts, accommodating, say, different speeds of action performance and 3) provides a natural form of data augmentation.

With the basic concept of a TBW in place, we now describe our proposed audio-visual fusion model, TBN.

3.2. TBN with Sparse Temporal Sampling

Our proposed TBN architecture is shown in Fig 2 (left). First, the action video is divided into K segments of equal width. Within each segment, we select a random sample of the first modality $\forall k \in K : m_{1k}$. This ensures the temporal progression of the action is captured by sparse temporal sampling of this modality, as with previous works [42, 45], while random sampling within the segment offers further data for training. The sampled m_{1k} is then used as the centre of a TBW of width $\pm b$. The other modalities are selected randomly from within each TBW (Eq. 5). In total, the input to our architecture in both training and testing is $K \times M$ samples from M modalities.

Within each of the K TBWs, we argue that the complementary information in audio and vision can be better exploited by combining the internal representations of each modality before temporal aggregation, and hence we propose a *mid-level* fusion. A ConvNet (per modality) extracts *mid-level* features, which are then fused through *concatenating* the modality features and feeding them to a fully-connected layer, making multi-modal predictions per TBW. We backpropagate all the way to the inputs of the ConvNets. Fig 3 details the proposed TBN block. The predictions, for each of these unified multimodal representations, are then aggregated for video-level predictions. In the proposed architecture, we train all modalities simultaneously. The convolutional weights for each modality are shared over the K segments. Additionally, mid-level fusion weights and class prediction weights are also shared across the segments.

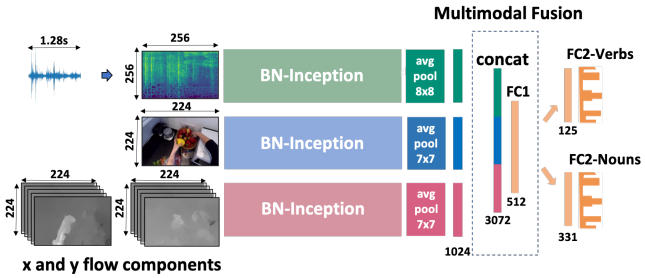

Figure 3: A single TBN block showing architectural details and feature sizes. Outputs from multiple TBN blocks are averaged as shown in Fig. 2. We model the problem of learning both verbs and nouns as a multi-task learning problem, by adding two output FC layers, one that predicts verbs and the other nouns (as in [8]). Best viewed in colour.

To avoid biasing the fusion towards longer or shorter action lengths, we calculate the window width b relative to the action video length. Our TBW is thus of variable width, where the width is a function of the length of the action. We note again that b can be set independently of the number of segments K, allowing the temporal windows to overlap. This is detailed in Sec. 4.1.

Relation to TSN. In Fig 2, we contrast the TBN architecture (left) to an extended version of the TSN architecture (right). The extension is to include the audio modality, since the original TSN only utilises appearance and motion streams. There are two key differences: first, in TSN each modality is temporally aggregated independently (across segments), and the modalities are only combined by late fusion (e.g. the RGB scores of each segment are temporally aggregated, and the flow scores of each segment are temporally aggregated, individually). Hence, it is not possible to benefit from combining modalities *within* a segment which is the case for TBN. Second, in TSN, each modality is trained independently first after which predictions are combined in inference. In the TBN model instead, all modalities are trained simultaneously, and their combination is also learnt.

4. Experiments

Dataset: We evaluate the TBN architecture on the largest dataset in egocentric vision: EPIC-Kitchens [8], which contains 39,596 action segments recorded by 32 participants performing non-scripted daily activities in their native kitchen environments. In EPIC-Kitchens, an action is defined as a combination of a *verb* and a *noun*, *e.g.* 'cut

cheese'. There are in total 125 verb classes and 331 noun classes, though these are heavily-imbalanced. The test set is divided in two splits: Seen Kitchens (S1) where sequences from the same environment are in both training, and Unseen Kitchens (S2) where the complete sequences for 4 participants are held out for testing. Importantly, EPIC-Kitchens sequences have been captured using a head-mounted GoPro with the audio released as part of the dataset. No previous baseline on using audio for this dataset is available.

4.1. Implementation Details

RGB and Flow: We use the publicly available RGB and computed optical flow with the dataset [8].

Audio Processing: We extract 1.28s of audio, convert it to single-channel, and resample it to 24kHz. We then convert it to a log-spectrogram representation using an STFT of window length 10ms, hop length 5ms and 256 frequency bands. This results in a 2D spectrogram matrix of size 256×256, after which we compute the logarithm. Since many egocentric actions are very short (< 1.28s), we extract 1.28s of audio from the untrimmed video, allowing the audio segment to extend beyond the action boundaries.

Training details: We implement our model in PyTorch [31]. We use Inception with Batch Normalisation (BN-Inception) [14] as a base architecture, and fuse the modalities after the average pooling layer. We chose BN-Inception as it offers a good compromise between performance and model-size, critical for our proposed TBN that trains all modalities simultaneously, and hence is memory-intensive. Compared to TSN, the three modalities have 10.78M, 10.4M and 10.4M parameters, with only one modality in memory during training. In contrast, TBN has 32.64M paramaters.

We train using SGD with momentum [33], a batch size of 128, a dropout of 0.5, a momentum of 0.9, and a learning rate of 0.01. Networks are trained for 80 epochs, and the learning rate is decayed by a factor of 10 at epoch 60. We initialise the RGB and the Audio streams from ImageNet. While for the Flow stream, we use stacks of 10 interleaved horizontal and vertical optical flow frames, and use the pre-trained Kinetics [7] model, provided by the authors of [42].

Note that our network is trained end-to-end for all modalities and TBWs. We train with $K = 3$ segments over the $M = 3$ modalities, with $b = T$, allowing the temporal window to be as large as the action segment. We test using 25 evenly spaced samples for each modality, as with the TSN basecode for direct comparison.

4.2. Results

This section is organised as follows. First, we show and discuss the performance of single modalities, and compare them with our proposed TBN, with a special focus on the efficacy of the audio stream. Second, we compare different mid-level fusion techniques. And finally, we investigate the effect of the TBW width on both training and testing.

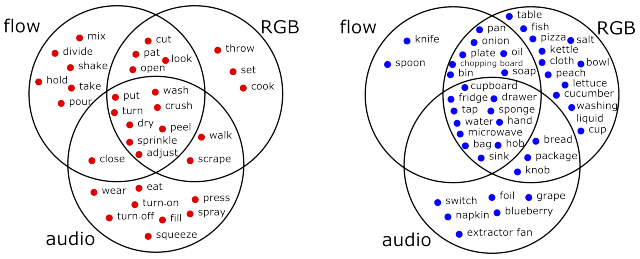

Figure 4: Verb (left) and noun (right) classes' performances using single modalities for top-performing 32 verb and 41 noun classes, using single modality accuracy. For each, we consider whether the accuracy is high for Flow, Audio or RGB, or for two or all of these modalities. It can be clearly seen that noun classes can be predicted with high accuracy using RGB alone, whereas for many verbs, Flow and Audio are also important modalities.

Single-modality vs multimodal fusion performance: We examine the overall performance of each modality individually in Table 1. Although it is clear that RGB and optical flow are stronger modalities than audio, an interesting find is that audio performs comparably to RGB on some of the metrics (e.g. top-1 verb accuracy), signifying the relevance of audio on recognising egocentric actions. While as expected optical flow outperforms RGB in **S2**, interestingly for **S1**, the RGB and Flow modalities perform comparatively, and in some cases RGB performs better. This matches the expectation that Flow is more invariant to the environment.

To obtain a better analysis of how these modalities perform, we examine the accuracy of *individual* verb and noun classes on **S1**, using single modalities. Fig 4 plots top-performing verb and noun classes, into a Venn diagram. For each class, we consider the accuracy of individual modalities. If all modalities perform comparably (within 0.15), we plot that class in the intersection of the three circles. On the other hand, if one modality is clearly better than the others (more than 0.15), we plot the class in the outer part of the modality's circle. For example, for the verb 'close', we have per-class accuracy of 0.23, 0.47 and 0.42 for RGB, Flow and Audio respectively. We thus note that this class performs best for two modalities: Flow and Audio, and plot it in the intersection of these two circles.

From this plot, many verb and noun classes perform comparably for all modalities (e.g. 'wash', 'peel' and 'fridge', 'sponge'). This suggests all three modalities contain useful information for these tasks. A distinctive difference, however, is observed in the importance of individual modalities for verbs and nouns. Verb classes are strongly related to the temporal progression of actions, making Flow

		Top-1 Accuracy			Top-5 Accuracy			Avg Class Precision			Avg Class Recall		
		VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION
S1	RGB	45.68	36.80	19.86	85.56	64.19	41.89	61.64	34.32	09.96	23.81	31.62	08.81
	Flow	55.65	31.17	20.10	85.99	56.00	39.30	48.83	26.84	09.02	27.58	24.15	07.89
	Audio	43.56	22.35	14.21	79.66	43.68	27.82	32.28	19.10	07.27	25.33	18.16	06.17
	TBN (RGB+Flow)	60.87	42.93	30.31	89.68	68.63	51.81	**61.93**	39.68	18.11	39.99	38.37	16.90
	TBN (All)	**64.75**	**46.03**	**34.80**	**90.70**	**71.34**	**56.65**	55.67	**43.65**	**22.07**	**45.55**	**42.30**	**21.31**
S2	RGB	34.89	21.82	10.11	74.56	45.34	25.33	19.48	14.67	04.77	11.22	17.24	05.67
	Flow	48.21	22.98	14.48	77.85	45.55	29.33	23.00	13.29	05.63	19.61	16.09	07.61
	Audio	35.43	11.98	06.45	69.20	29.49	16.18	22.46	09.41	04.59	18.02	09.79	04.19
	TBN (RGB+Flow)	49.61	25.68	16.80	78.36	50.94	32.61	30.54	20.56	09.89	21.90	20.62	11.21
	TBN (All)	**52.69**	**27.86**	**19.06**	**79.93**	**53.78**	**36.54**	**31.44**	**21.48**	**12.00**	**28.21**	**23.53**	**12.69**

Table 1: Comparison of our fusion method to single modality performance. For both splits, the fusion outperforms single modalities. For the seen split, the RGB and Flow modalities perform comparatively, whereas for the unseen split the Flow modality outperforms RGB by a large margin. Audio is comparable to RGB on top-1 verb accuracy for both splits.

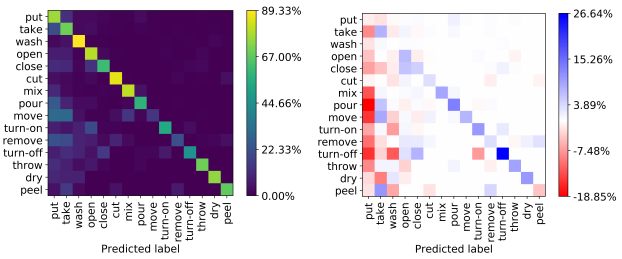

Figure 5: Confusion matrix for the largest-15 verb classes, with audio (left), as well as the difference to the confusion matrix without audio (right).

		All			RGB+Flow		
	TBN	VERB	NOUN	ACTION	VERB	NOUN	ACTION
S1	irrelevant	61.37	46.46	32.63	57.28	42.55	27.73
	rest	65.28	45.97	35.14	61.44	42.99	30.72
S2	irrelevant	47.32	23.36	15.30	44.41	20.45	12.39
	rest	57.21	31.66	22.22	54.00	30.09	20.52

Table 2: Comparing top-1 accuracy of All modalities (left) to RGB+Flow (right). Actions are split in segments with 'irrelevant' background sounds, and the 'rest' of the test set.

more important for verbs than nouns. Conversely, noun classes can be predicted with high accuracy using RGB alone. Audio, on the other hand, is important for both nouns and verbs, particularly for some verbs such as 'turn-on', and 'spray'. For nouns, Audio tends to perform better for objects with distinctive sounds (e.g. 'switch', 'extractor fan') and materials that sound when manipulated (e.g. 'foil').

In Table. 1, we compare single modality performance to the performance over the three modalities. Single modalities are trained as in TSN, as TBN is designed to bind multiple modalities. We find that the fusion method outperforms single modalities, and that audio is a significantly informative modality across the board. Per-class accuracies, for individual modalities as well as for TBN trained on all three modalities, can be seen in Figure 6. The advantage of the fusion method is more pronounced for verbs (where we expect motion and audio to be more informative) than nouns, and more for particular noun classes than others, such as 'pot', 'kettle', 'microwave', and particular verb classes eg. 'spray' (fusion 0.54, RGB 0.09, Flow 0, Audio 0.3). This suggests that the mixture of complementary and redundant information captured in a video is highly dependant on the action itself, yielding the fusion method to be more useful for some classes than for others. We also note that the fusion method helps to significantly boost the performance of the tail classes (Fig. 6, right and table in supplementary), where individual modality performance tends to suffer.

Efficacy of audio: We train TBN only with the visual modalities (RGB+Flow) and the results can be seen in Table 1. An increase of 5% (S1) and 4% (S2) in top-5 action recognition accuracy with the addition of audio demonstrates the importance of audio for egocentric action recognition. Fig 5 shows the confusion matrix with the utilisation of audio for the largest-15 verb classes (in **S1**). Studying the difference (Fig 5 right) clearly demonstrates an increase (blue) in confidence along the diagonal, and a decrease (red) in confusion elsewhere.

Audio with irrelevant sounds: In the recorded videos for EPIC-Kitchens, background sounds irrelevant to the observed actions have been captured by the wearable sensor. These include music or TV playing in the background, ongoing washing machine, coffee machine or frying sounds while actions take place. To quantify the effect of these sounds, we annotated the audio in the test set, and report that 14% of all action segments in S1, and 46% of all action segments in S2 contain other audio sources. We refer to these as actions containing 'irrelevant' sounds, and independently report the results in Table 2. The table shows that the model's accuracy increases consistently when audio is incorporated, even for the 'irrelevant' segments. Both models (All and RGB+Flow) show a drop in performance for

		Top-1 Accuracy			Top-5 Accuracy			Avg Class Precision			Avg Class Recall		
		VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION
S1	Concatenation	**64.75**	**46.03**	**34.80**	**90.70**	**71.34**	**56.65**	55.67	**43.65**	**22.07**	45.55	**42.30**	**21.31**
	Context gating [22]	63.77	44.33	33.47	90.04	69.09	54.10	**57.31**	42.20	21.72	**45.63**	41.53	20.20
	Gating fusion [4]	61.52	43.54	31.61	89.54	68.42	52.57	52.07	39.62	18.39	42.55	39.77	18.66
S2	Concatenation	**52.69**	**27.86**	19.06	**79.93**	**53.78**	**36.54**	**31.44**	21.48	12.00	**28.21**	**23.53**	12.69
	Context gating [22]	52.65	27.35	**19.16**	79.25	52.00	36.40	30.82	**23.16**	11.72	23.39	25.03	12.58
	Gating fusion [4]	50.16	27.25	18.41	78.80	50.84	34.04	28.42	22.42	**12.34**	23.92	24.15	**13.14**

Table 3: Comparison of mid-level fusion techniques for the TBN architecture.

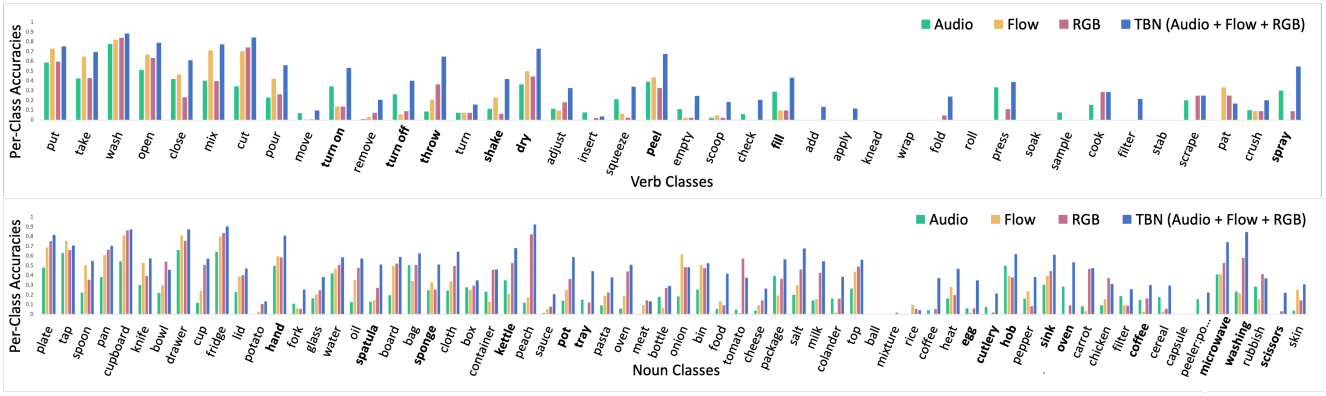

Figure 6: Per-class accuracies for the **S1** test set for verbs (top) and nouns (bottom) for fusion and single modalities. We select verb classes with more than 10 samples, and noun classes with more than 30 samples. The classes are presented in the order of number of samples per class, from left to right. For most classes the fusion method provides significant performance gains over single modality classification (largest improvements shown in bold). Best viewed in colour.

'irrelevant' S2 (comparing to 'rest'), validating that irrelevant sounds are not the source of confusion, but that this set of action segments is more challenging even in the visual modalities. This demonstrates the robustness of our network to noisy and unconstrained audio sources.

Comparison of fusion strategies: As Fig 2 indicates, TBN performs mid-level fusion on the modalities within the binding window. Here we describe three alternative mid-level fusion strategies, and then compare their performances.

(i) Concatenation, where the feature maps of each modality are concatenated, and a fully-connected layer is used to model the cross-modal relations.

$$f_{tbw}^{concat} = \phi(W[m_{1j}, m_{2k}, m_{3l}] + b) \quad (6)$$

where ϕ is a non-linear activation function. When used within TBWs, shared weights f_{tbw} are to be learnt between modalities within a range of temporal shifts.

(ii) *Context gating* was used in [22], aiming to recalibrate the strength of the activations of different units with a self-gating mechanism:

$$f_{tbw}^{context} = \sigma(Wh + b_z) \circ h \quad (7)$$

where $\circ$ is element-wise multiplication. We apply context gating on top of our multi-modal fusion with concatenation, so h in (7) is equivalent to (6).

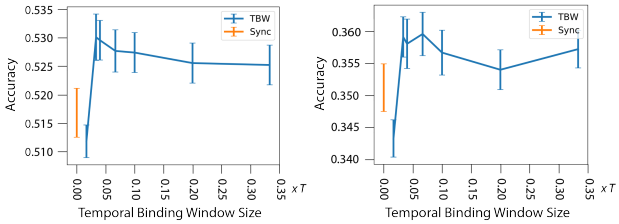

Figure 7: Effect of TBW width for verbs (left) and nouns (right) in the **S1** test set.

(iii) *Gating fusion* was introduced in [4], where a gate neuron takes as input the features from all modalities to learn the importance of one modality w.r.t. all modalities.

$$h_i = \phi(W_i m_{ij} + b_i) \quad \forall i \quad (8)$$

$$z_i = \sigma(W_{zi}[m_{1j}, m_{2k}, m_{3l}] + b_{z_i}) \quad \forall i \quad (9)$$

$$f_{tbw}^{gating} = z_1 \circ h_1 + z_2 \circ h_2 + z_3 \circ h_3, \quad (10)$$

In Table 3, we compare the various fusion strategies. We find that the simplest method, concatenation (Eq. 6) generally outperforms more complex fusion approaches. We believe this shows modality binding within a temporal binding window to be robust to the mid-level fusion method.

		Top-1 Accuracy			Top-5 Accuracy			Avg Class Precision			Avg Class Recall		
		VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION
S1	Attention Clusters [19]	40.39	19.37	11.09	78.13	41.73	24.36	21.17	09.65	02.50	14.89	11.50	03.41
	[8] (from leaderboard)	48.23	36.71	20.54	84.09	62.32	39.79	47.26	35.42	11.57	22.33	30.53	09.78
	Ours (TSN [42] w. Audio)	55.49	36.27	23.95	87.04	64.17	44.26	53.85	30.94	13.55	30.60	29.82	11.11
	Ours (TBN, Single Model)	64.75	46.03	34.80	90.70	71.34	56.65	55.67	43.65	22.07	45.55	42.30	21.31
	Ours (TBN, Ensemble)	**66.10**	**47.89**	**36.66**	**91.28**	**72.80**	**58.62**	**60.74**	**44.90**	**24.02**	**46.82**	**43.89**	**22.92**
S2	Attention Clusters [19]	32.37	11.95	05.60	69.89	31.82	15.74	17.21	03.86	01.84	11.59	07.94	02.64
	[8] (from leaderboard)	39.40	22.70	10.89	74.29	45.72	25.26	22.54	15.33	06.21	13.06	17.52	06.49
	Ours (TSN [42] w. Audio)	46.61	22.50	13.05	78.19	48.59	29.13	28.92	15.48	06.47	21.58	16.61	07.55
	Ours (TBN, Single Model)	52.69	27.86	19.06	79.93	53.78	36.54	31.44	21.48	**12.00**	**28.21**	23.53	12.69
	Ours (TBN, Ensemble)	**54.46**	**30.39**	**20.97**	**81.23**	**55.69**	**39.40**	**32.57**	**21.68**	10.96	27.60	**25.58**	**13.31**

Table 4: Results on the EPIC-Kitchens for seen (S1) and unseen (S2) test splits. At the time of submission, our method outperformed all previous methods on all metrics, and in particular by 11%, 5% and 4% on top-1 verb, noun and action classification on S1. Our method achieved second ranking in the 2019 challenge. Screenshots of the leaderboard at submission and challenge conclusion are in the supplementary material.

The effect of TBW width: Here, we investigate the effect of the TBW width in training and testing. We varied TBW width in training with $b \in \{\frac{T}{6}, \frac{T}{3}, T\}$, by training three TBN models for each respective window width. We noted little difference in performance. As changing b in training is expensive and performance is subject to the particular optimisation run, we opt for a more conclusive test by focusing on varying b in testing for a single model.

In testing, we vary $b \in \{\frac{T}{60}, \frac{T}{30}, \frac{T}{25}, \frac{T}{15}, \frac{T}{10}, \frac{T}{5}, \frac{T}{3}\}$. This corresponds, in average, to varying the width of TBW on the **S1** test set between 60ms and 1200ms. We additionally run with synchrony $b \sim 0$. In each case we sample a *single TBW*, to solely assess the effect of the window size. We repeat this experiment for 100 runs and report mean and standard deviation in Fig. 7, where we compare results for verb and noun classes separately. The figure shows that best performance is achieved for $b \in \left[\frac{T}{30}, \frac{T}{20}\right]$, that is on average $b \in [120ms \pm 190ms, 180ms \pm 285ms]$. TBWs of smaller width show a clear drop in performance, with synchrony comparable to $b = \frac{T}{60}$. Note that the 'Sync' baseline provides only approximate synchronisation of modalities, as modalities have different sampling rates (RGB 60fps, flow 30fps, audio 24000kHz). The model shows a degree of robustness for larger TBWs.

Note that in Fig. 7, we compare widths on a single temporal window in testing. When we temporally aggregate multiple TBWs, the effect of the TBW width is smoothed, and the model becomes robust to TBW widths.

Comparison with the state-of-the-art: We compare our work to the baseline results reported in [8] in Table 4 on all metrics. First we show that a late fusion with an additional audio stream, outperforms the baseline on top-1 verb accuracy by 7% on S1 and also 7% on S2. Second, we show that our TBN single model, improves these results even further (9%, 10% and 11% on top-1 verb, noun and action accuracy on S1, and 6%, 5% and 6% on S2 respectively). Finally we report results of an Ensemble of five TBNs, where each one is trained with a different TBW width. The ensemble shows additional improvement of up to 3% on top-1 metrics.

We compare TBN with Attention Clusters [19], a previous effort to utilise RGB, Flow, and Audio for action recognition, using *pre-extracted features*. We use the authors available implementation, and fine-tuned features (TSN, BN-Inception), from the global avg pooling layer (1024D), to provide a fair comparison to TBN, and follow the implementation choices from [19]. The method from [19] performs significantly worse than the baseline, as pre-extracted video features are used to learn attention weights.

At the time of submission, our TBN Ensemble results demonstrated an overall improvement over all state-of-the-art, published or anonymous, by 11% on top-1 verb for both S1 and S2. Our method was also ranked 2nd in the 2019 EPIC-Kitchens Action Recognition challenge. Details of the public leaderboard are provided in supplementary.

5. Conclusion

We have shown that the TBN architecture is able to flexibly combine the RGB, Flow and Audio modalities to achieve an across the board performance improvement, compared to individual modalities. In particular, we have demonstrated how audio is complementary to appearance and motion for a number of classes; and the pre-eminence of appearance for noun (rather than verb) classes. The performance of TBN significantly exceeds TSN trained on the same data; and provides state-of-the-art results on the public EPIC-Kitchens leaderboard.

Further avenues for exploration include a model that learns to adjust TBWs over time, as well as implementing class-specific temporal binding windows.

Acknowledgements Research supported by EPSRC LOCATE (EP/N033779/1), GLANCE (EP/N013964/1) & Seebibyte (EP/M013774/1). EK is funded by EPSRC Doctoral Training Partnership, and AN by a Google PhD Fellowship.

References

[1] Huda Alamri, Chiori Hori, Tim K. Marks, Dhruv Batra, and Devi Parikh. Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7. In *DSTC7 at AAAI2019 Workshop*, 2018. 2

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017. 1, 2

[3] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *ECCV*, September 2018. 1, 2

[4] John Arevalo, Thamar Solorio, Manuel Montes-y Gmez, and Fabio A. Gonzlez. Gated multimodal units for information fusion. In *ICLRW*, 2017. 7

[5] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *NIPS*, 2016. 1, 2

[6] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *CoRR*, abs/1706.00932, 2017. 1, 2

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2, 5

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 1, 2, 4, 5, 8

[9] Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, and Walterio Mayol-Cuevas. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*, 2014. 1

[10] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 1, 2, 3

[11] Ruohan Gao and Kristen Grauman. 2.5D visual sound. In *CVPR*, 2019. 1, 2

[12] Rohit Girdhar, Joo Carreira, Carl Doersch, and Andrew Zisserman. A better baseline for ava. In *ActivityNet Workshop at CVPR*, 2018. 2

[13] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *CVPR*, 2017. 2

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 5

[15] Herve Jegou, Matthijs Douze, Cordelia Schmid, and Patrick Perez. Aggregating local descriptors into a compact image representation. In *CVPR*, 2010. 3

[16] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NIPS*, pages 7763–7774. 2018. 2

[17] Yong J. Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 1

[18] Weiyao Lin, Yang Mi, Jianxin Wu, Ke Lu, and Hongkai Xiong. Action recognition with coarse-to-fine deep feature integration and asynchronous fusion. *AAAI*, 2018. 2, 4

[19] Xiang Long, Chuang Gan, Gerard de Melo, Jiajun Wu, Xiao Liu, and Shilei Wen. Attention clusters: Purely attention based local feature integration for video classification. In *CVPR*, June 2018. 2, 8

[20] Xiang Long, Chuang Gan, Gerard Melo, Xiao Liu, Yandong Li, Fu Li, and Shilei Wen. Multimodal keyless attention fusion for video classification. In *AAAI Conference on Artificial Intelligence*, 2018. 2

[21] Minghuang Ma, Haoqi Fan, and Kris M. Kitani. Going deeper into first-person activity recognition. In *CVPR*, 2016. 1, 2

[22] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *CoRR*, abs/1706.06905, 2017. 1, 7

[23] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a Text-Video Embedding from Imcomplete and Heterogeneous Data. In *arXiv*, 2018. 1, 2

[24] Davide Moltisanti, Michael Wray, Walterio Mayol-Cuevas, and Dima Damen. Trespassing the boundaries: Labeling temporal bounds for object interactions in egocentric video. In *ICCV*, 2017. 2

[25] Pierre Mgevand, Sophie Molholm, Ashabari Nayak, and John J. Foxe. Recalibration of the multisensory temporal window of integration results from changing task demands. *PLOS ONE*, 8, 2013. 2

[26] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable PINs: Cross-modal embeddings for person identity. *ECCV*, 2018. 2

[27] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *CVPR*, 2018. 1, 2

[28] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 1, 2

[29] Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *ECCV*, 2016. 1, 2

[30] CesareaV. Parise, Charles Spence, and Marc O. Ernst. When correlation implies causation in multisensory integration. *Current Biology*, 22(1):46 – 49, 2012. 2

[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 5

[32] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 1

[33] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999. 5

[34] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018. 1, 2

[35] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *CVPR*, 2018. 1

[36] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*. 2014. 2

[37] Suriya Singh, Chetan Arora, and C. V. Jawahar. First person action recognition using deep learned descriptors. In *CVPR*, 2016. 2

[38] Sibo Song, Vijay Chandrasekhar, Bappaditya Mandal, Liyuan Li, Joo-Hwee Lim, Giduthuri Sateesh Babu, Phyo Phyo San, and Ngai-Man Cheung. Multimodal multi-stream deep learning for egocentric activity recognition. In *CVPRW*, 2016. 2

[39] Ryan A. Stevenson, Magdalena M. Wilson, Albert R. Powers, and Mark T. Wallace. The effects of visual training on multisensory temporal processing. *Experimental Brain Research*, 225(4):479–489, 2013. 2

[40] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. In *BMVC*, 2018. 2

[41] Mark T. Wallace and Ryan A. Stevenson. The construct of the multisensory temporal binding window and its dysregulation in developmental disabilities. *Neuropsychologia*, 64:105 – 123, 2014. 2

[42] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 1, 2, 3, 4, 5, 8

[43] Zuxuan Wu, Yu-Gang Jiang, Xi Wang, Hao Ye, and Xiangyang Xue. Multi-stream multi-class fusion of deep networks for video classification. In *ACM International Conference on Multimedia*, 2016. 2

[44] Ryo Yonetani, Kris M. Kitani, and Yoichi Sato. Recognizing micro-actions and reactions from paired egocentric videos. In *CVPR*, 2016. 1

[45] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018. 2, 4

[46] Yipin Zhou and Tamara L Berg. Temporal perception and prediction in ego-centric video. In *ICCV*, 2015. 1