Context in Human Action through Motion Complementarity

Eadom Dessalene, Michael Maynord, Cornelia Fermüller, Yiannis Aloimonos
University of Maryland, College Park
College Park, MD 20742, USA
{edessale,maynord,fermulcm,jyaloimo@umd.edu}

Abstract

Motivated by Goldman's Theory of Human Action - a framework in which action decomposes into 1) base physical movements, and 2) the context in which they occur - we propose a novel learning formulation for motion and context, where context is derived as the complement to motion. More specifically, we model physical movement through the adoption of Therbligs, a set of elemental physical motions centered around object manipulation. Context is modeled through the use of a contrastive mutual information loss that formulates context information as the action information not contained within movement information. We empirically prove the utility brought by this separation of representation, showing sizable improvements in action recognition and action anticipation accuracies for a variety of models. We present results over two object manipulation datasets: EPIC Kitchens 100, and 50 Salads.

1. Introduction

Goldman's Theory of Action [14] addresses the problem of action categorization, across levels of context. Taking inspiration from Goldman's theory, we model base movement as the core of action, and model context to exclude base movement.

To illustrate, consider the physical act of moving an object to one's mouth, and 4 scenarios where this occurs. These scenarios share the same base motion, but differ in context: 1) moving an apple to one's mouth in a kitchen, 2) moving a microphone to one's mouth on a stage, 3) moving a drink to one's mouth at a coffee shop, 4) moving a toothbrush to one's mouth in the bathroom. In these scenarios modeling of context separate from physical motion will provide information relevant to understanding the action being performed. Therefore, we model context as that which is relevant to action but not contained within base movement.

Fundamentally the nature of body movement and context differ. Body movement is well-defined and constrained within the physical embodiment of the actor, whereas context extends beyond the motion of the actor. However, most action models do not differentiate between the two. We propose the separation of the physical aspects of action from its contextual aspects.

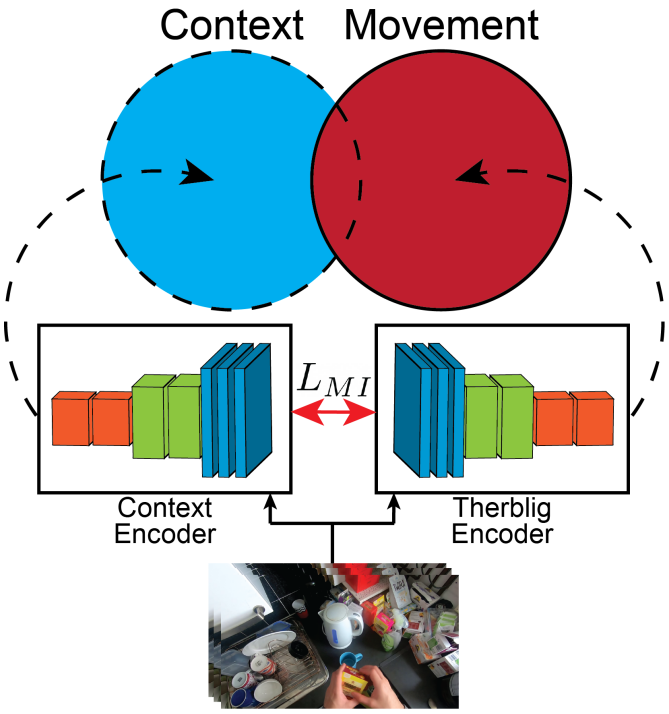

Figure 1. Conceptual illustration of our proposed framework, which models motion representations separate from contextual representations, where these representations are constrained to be complementary through a contrastive loss formulation L_{MI}. We employ two streams: A Context Encoder and a Therblig Encoder. The Therblig Encoder maps video to representations of movement. The Context Encoder models representations of action complementary to representations of movement produced by the Therblig Encoder. Together, they capture information pertaining to the relevant aspects of action.

Figure 2. Example sequences from the EPIC Kitchens dataset [12] where contextual information plays a large role in the interpretation of action. Sequences listed from top to bottom: (A) the tray is emptied after the removal of each object inside, and so the action becomes *empty the dish rack*, (B) the asparagus is wet prior to its placement in the drainer, and so the action becomes *drain water from asparagus*, (C) the stove is turned on throughout the flipping of the food inside the pan, and so the action becomes *cook the mix inside the pan*. In each of these examples, understanding the base movement being performed is insufficient to arrive at a full understanding of the high level action being performed. Only when context is incorporated does the full nature of the action become apparent.

To model base movements, we adopt the Therblig formulation from [13] - a set of elemental motion primitives that divides all motion into three categories: 1) motion required for performing an operation, 2) extraneous motion slowing down the performing of operation, and 3) motions that do not perform an operation. Therbligs were originally used to analyze brick-laying work [4], but have since been applied to represent and understand a wide variety of domains, e.g. assembly lines [21], education [8], surgery [18], etc. We adopt densely labeled sets of Therblig annotations over several video datasets from [13] - we use these annotations in the modeling of physical movement. We extract features f_i through the modeling of Therbligs.

However, contextual information extends well beyond motion information. Within the realm of action, context includes objects, timings, location, activity, etc. Context information is bound by information pertaining to that of the action being performed.

To give intuition to the role of context in understanding actions, Figure 2 lists examples of how the context accompanying the base physical actions performed defines the actions. Many actions (e.g. remove, insert, cook, search, etc) can only be understood through the context in which they occur.

Without the understanding of context, the only actions to be understood are those that are defined at the level of body movement. For example, EPIC Kitchens, at most only 33% of the 95 verbs stay at the level of agent movement, with interpretations of various levels of context required for the understanding of all other actions.

Just as the action of moving a finger overlaps with the action of shooting a gun, actions in Figure 2 exhibit semantic overlap (e.g. flipping vegetables overlaps with cooking vegetables). The contextual content of overlapping actions enables the modeling of action beyond body movement action. That is, by removing the body movement from action descriptions, the context of action can be derived.

Existing approaches to action understanding typically model aspects of context individually (e.g. grasp type, body formation, object timings, long-term activity, etc). Rather than model each element of context individually, we model context holistically through *complementarity* (that is, by removing the body movement from action descriptions, the context of action can be derived). We extract contextual information c_i by means of complementarity with f_i. To do this, we adopt the estimation of mutual information as a regularizer during training to minimize the information shared between f_i and c_i. Simultaneously, context features c_i are used to capture information relevant to the understanding of action through training towards downstream tasks in action understanding.

More specifically, we introduce a novel loss formulation that minimizes the information shared between Therblig features and contextual features. This loss formulation also maximizes the information shared between contextual features and action in the form of categorical cross-entropy. Through this approach we are able to minimize mutual information between context features and movement features, and achieve complementarity between context and movement with respect to action. We perform a final concate-

nation of the features and demonstrate their superiority to models which do not differentiate between the two.

The primary contributions of our work are as follows:

- A contrastive formulation for the modeling of context in a way that achieves complementarity with movement, implemented within a two-stream archicture.
- Novel use of Therbligs in the derivation of context.
- Empirical support for utility of context in action understanding over two popular datasets (EPIC Kitchens 100 [12] and 50 Salads [27]) and four popular video architectures.

The rest of this paper is structured as follows: in Section 2 we cover related work; in Sections 3 and 4 we detail methods; in Section 5 we present experiments; in Section 6 we provide discussion; and in Section 7 we conclude.

2. Related Works

2.1. Therbligs

Taking a rigorous approach to analyzing the movements of workers for the minimization of unnecessary motion and simplification of necessary motion, Frank B. Gilbreth introduced a set of elemental motions he termed Therbligs [4] (Gilbreth spelled backwards). Therbligs are centered on the *eyes* and *hands*. Gilbreth categorized all motion into three categories: 1) motion required for performing an operation, 2) extraneous motion slowing down the performing of an operation, and 3) motions that do not perform an operation. Therbligs were originally used to analyze brick-laying work [4], but have since been applied to represent and understand a wide variety of domains, i.e. assembly lines [21], education [8], surgery [18], etc.

We leverage Therblig annotations released in [13] over the EPIC Kitchens and 50 Salads datasets. These annotations allow us to model the sequence of physical movements involved in a given video of human activity in isolation from broader contextual cues. Therbligs differ from other sub-action ontologies by 1) resolving temporal ambiguity by means of contact, 2) having a simple, logically consistent data collection process enabled through the imposing of commonsense rules, and 3) being flexible in application to a wide variety of datasets within the realm of object manipulation without relying on domain expertise. See [13] for details.

2.2. Mutual Information in Computer Vision

Though mutual information (MI) has remained challenging to estimate, there has been significant interest in its estimation over several applications in computer vision. Mutual information minimization amounts to minimizing the correlation between variables x and y. Mutual information has been a useful regularizer in reducing non-task related redundancy in representation learning for downstream tasks in computer vision and has been applied towards image classification [3] [24], semantic segmentation [32], saliency detection [31], object detection [30], question generation [20], etc. [7] prove that, generally, minimizing standard cross-entropy loss amounts to maximizing the mutual information. We apply this finding along with the mutual information minimization formulation of [31] and [32] to achieve complementarity of representation.

Symbol	Name	Description
	Grasp (**G**)	When the worker's hand grabs the object
	Release (**R**)	The releasing of the object when it reaches its destination
	Hold (**H**)	The retention of an object such that it undergoes no movement while in operation
	Move (**M**)	Moving a loaded hand to the point of release, use, hold or (pre)position
	Reach (**Re**)	Moving an empty hand from the point of release
	Use (**U**)	When an object is being operated as intended
	Orient (**O**)	Changing the orientation of object while keeping it in roughly the same position

Figure 3. Listed above are the Therbligs we employ, their symbolic illustrations, and brief descriptions of their usage.

2.3. Context in Action Understanding

There exist a wide variety of approaches that incorporate a wide variety of contextual cues for downstream tasks in action understanding. Previous approaches that represent context through additional learning objectives use region proposals localizing people and/or surrounding objects [10] [16], high-level representations of activity [25], longer-term temporal awareness [5] [11], body and hand pose [15] [2], etc. To our knowledge, we are the first to represent contextual features as that which represent information complementary to information represented by movement features. This formulation aims to represent context in the broadest sense possible than modeling context as that which is strictly spatial, temporal, or knowledge-based.

2.4. Two-Stream Video Architectures

Two-stream architectures for the understanding of action began with [26], which used RGB and optical flow frames to represent spatial and temporal information in separate streams of input. Many approaches have followed this trend over different base architectures and action understanding tasks [23] [28] [29] [22]. While these approaches incorporate streams capturing different representations, our primary contribution which differentiates our work from other two-stream approaches is the enforcement of complementarity, modeled through use of a mutual-information loss formulation. Furthermore, two-stream approaches to action understanding rely on optical flow as motion. Flow does not contain much of the semantic information pertaining to base physical movement that is contained within the Therbligs.

3. Therblig Dataset

For video segment s_i with T frames, the Therblig representation t_i is a sequence of N Therblig tuples for every 100-frame chunk of s_i. Each Therblig annotation t_i is a sequence of the form $(v_0, o_0), ...(v_{N-1}, o_{N-1})$, where $v_j \in V$ and $V = \{\phi, Re, M, G, R, U, O, H\}$. In other words, v_j indicates the Therblig verb and each o_j indicates the noun of the object of interaction. See Figure 2 for example Therblig sequences shown side-by-side with their corresponding contexts and actions and Figure 3 for full descriptions of each element of V (not included is ϕ, corresponding to an empty sequence where no Therbligs occur). See [13] for more on the Therblig dataset we utilize throughout the experiments conducted in this paper.

4. Methods

We introduce a two-stream architecture where one stream models physical movement and another stream models context. This offers a division of our architecture into two primary components; a component (**Therblig Encoder**) mapping video to representations of movement and another component (**Context Encoder**) mapping video to representations of context. The base architectures of both encoders are derived directly from existing video architectures, with slight alterations performed to the Therblig Encoder (detailed in Section 4.1). We adopt the same backbone architectures between the Therblig Encoder and the Context Encoder, and both encoders are initialized from the same set of pre-trained weights over Kinetics-600.

Of the four models over which we experiment, models #1 and #2 (I3D, MoViNet) are built on 3D Convolutions, and #3 and #4 (ViViT, TimeSFormer) are variants of Transformer models applied towards the modeling of video.

We adopt a two-stage training process, where the Therblig Encoder is trained initially over our dataset of Therblig annotations and frozen during the training of the Context Encoder. Figure 4 shows an overview of our framework. The Therblig encoder produces features f_i, feeding f_i to a 2-layer GRU which predicts Therblig sequence $\hat{t_i}$. The Therblig Encoder is trained solely off the categorical cross-entropy loss L^t_{CE} between the sequence of Therblig predictions $\hat{t_i}$ and the ground truth Therblig sequence t_i. As such, the Therblig Encoder captures only information pertaining to motion.

After training the Therblig Encoder for 30 epochs, we begin the training of the Context Encoder. The modeling of contextual cues for purposes of action understanding is by no means novel. However, our modeling of context goes beyond the modeling of individual contextual characteristics in that we represent context very broadly - contextual information is defined as that which corresponds to the information carried in the action not contained within the base physical movements performed. We achieve this by simultaneously 1) minimizing the information shared between the produced context representations c_i and movement representations f_i by means of a mutual information loss term L_{MI}, and 2) maximizing the information shared between context representations c_i and action labels by means of a categorical cross-entropy loss term L^a_{CE}.

We keep the weights of the Therblig Encoder fixed during the training of the Context Encoder - this prevents the Therblig Encoder from learning representations unrelated to physical movement through the L^a_{CE} loss term applied to the Context Encoder, and allows the Context Encoder to capture our notion of context. Incorporating information beyond motion information during the training of the Therblig Encoder would hinder the ability of the Context Encoder to model context during its training. As such, during the training of the Therblig Encoder we abstract away object category using a sequential ordering schema (e.g. a Therblig sequence corresponding to *Reach for [cup], Grasp [cup], Use [faucet]* is predicted as *Reach for [0], Grasp [0], Use [1]*) - that is, object information is not given to the Therblig encoder during training.

The outputs of the Therblig Encoder f_i and the Context Encoder c_i are concatenated and fed to final classification layer(s) (Action Head) to produce action class likelihoods $\hat{a}$.

4.1. Therblig Encoder

The modeling of Therbligs requires an architecture capable of capturing the fine-grained aspects of motion, below the verb-level of action. However, video understanding architectures are typically designed for the purposes of predicting a single high-level action class likelihood representing the entirety of the video input clip and are not immediately applicable for prediction of low-level sequential movements.

With this in mind, we perform a small adjustment to each

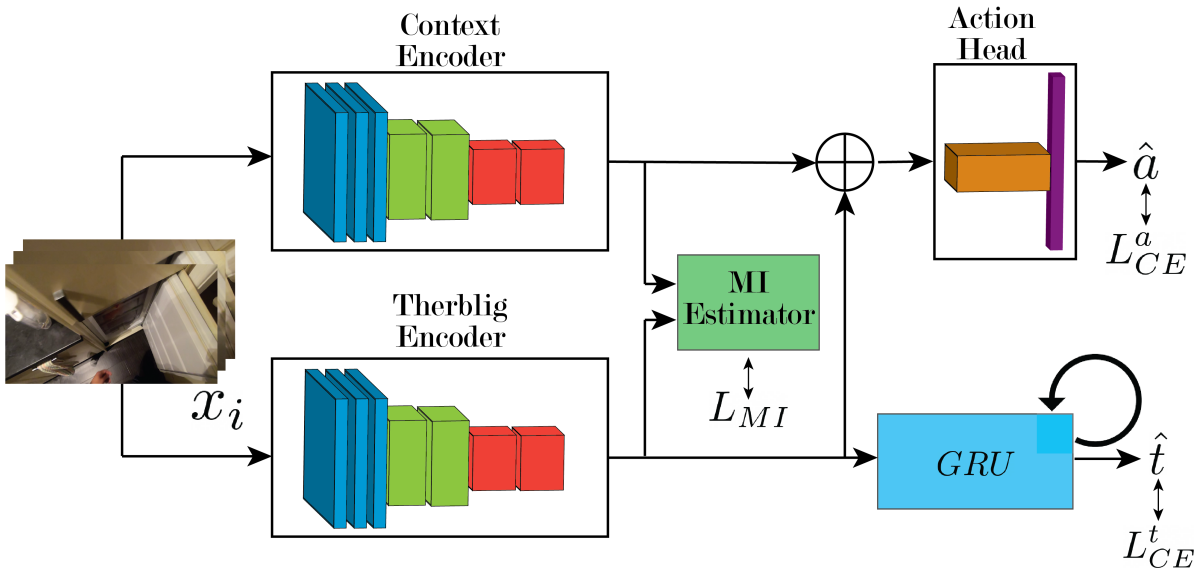

Figure 4. Our proposed two-stream architecture, where input video is simultaneously fed to our proposed Therblig Encoder and Context Encoder. The two streams are trained separately over different sources of supervision: The Therblig Encoder consists of a 2-layer GRU stacked on top of a backbone video architecture, where the loss function consists of the Categorical Cross-Entropy between ground truth Therblig sequence t and predicted Therblig sequence $\hat{t}$. The Context Encoder consists of an identical backbone video architecture producing context features c, where the loss function comprises of the mutual information estimate between c and t and the categorical cross-entropy between ground truth action annotations a and predicted actions $\hat{a}$. The Therblig Encoder is trained first and frozen during the training of the Context Encoder.

architecture for the modeling of the Therblig Encoder. In adapting models #1 and #2, we preserve temporal information necessary for prediction of the Therblig sequence by removing the depth-wise aspect of the final average-pooling operation. In adapting models #3 and #4, we independently apply the temporal attention layers over 6 separate chunks of the input video (intuitively spanning 6 possible Therblig predictions) rather than over the input video in its entirety. We concatenate the feature outputs of the temporal attention layers over the temporal axis.

We feed all 100 RGB frames of video input to the Therblig Encoder, producing Therblig features f_i. We set the initial hidden state of a 2-layer GRU to f_i with an input vector of $\vec{0}$ and the network is rolled out to iteratively predict a sequence of up to 6 Therbligs $\hat{t}_{j_n}$, where n corresponds to the n-th Therblig predicted for $1 \leq n \leq 6$. The GRU is fed $\vec{0}$ as the initial input, and outputs of previous hidden layers as inputs for subsequent timesteps. The network is trained via categorical cross-entropy loss between $t^n_{i_j}$ and $\hat{t^n}_{i_j}$ for each time step $1 \leq n \leq 6$, where $t^n_{i_j}$ corresponds to the ground truth Therblig annotation sequence.

The Therblig Encoder and GRU are trained in conjunction. Their weights are frozen during the training of the Context Encoder described below.

4.2. Context Encoder

Contextual cues useful for the understanding of action, include but are not limited to: object identity, object state transitions, long-term temporal semantics, etc. While there exist architectures better suited towards the modeling of each of these cues, the incorporation of these architectures makes it difficult to assess the individual contribution of our formulation of complementarity.

We feed the output of the Context Encoder along with the output of the Therblig Encoder to a mutual information estimator (**MI Estimator**), which consists of a single fully connected layer, modeling $p(x_i) = N(\mu_c, \sigma_c)$ where μ_c and σ_c are regressed mean and variance associated with c_i and $q(x_i) = N(\mu_t, \sigma_t)$ where μ_t and σ_t are regressed mean and variance associated with f_i. In addition, we adopt the re-parameterization trick for reasons related to training stability, where z_c and z_t are latent vectors representing $p(x_i)$ and $q(x_i)$, respectively.

Mutual information is defined as the difference between the entropy terms below:

$$M_I(z_c, z_t) = H(z_c) + H(z_t) - H(z_c, z_t) \quad (1)$$

where $H(z_c)$, $H(z_t)$ are the marginal entropies of z_c, z_t and $H(z_c, z_t)$ is the joint entropy of z_c and z_t. In estimating the marginal entropies $H(z_c)$ and $H(z_t)$ we adopt

the Kullback-Leibler divergence formulation with the re-parameterized z_c and z_t of $p(x_i)$ and $q(x_i)$ respectively:

$$KL(P||Q) = H_{z_t}(z_c) - H(z_c) \tag{2}$$

$$KL(Q||P) = H_{z_c}(z_t) - H(z_t) \tag{3}$$

Combining equations (1), (2), and (3), we have

$$\begin{aligned} L_{MI} = & H_{z_c}(z_t) + H_{z_t}(z_c) \\ & - (KL(Q||P) - KL(P||Q)) \end{aligned} \tag{4}$$

where $H_{z_c}(z_t)$ and $H_{z_t}(z_c)$ correspond to cross-entropies, P and Q are random variables associated with context and Therblig features respectively, and KL corresponds to the KL-divergence between the two latent features.

By backpropagating through the combined loss of

$$L = L_{CE}^{a} + \alpha\beta L_{MI} \tag{5}$$

where α is a scaling term (set to 0.1 for all experiments in this paper) and β is a linear annealing term (increasing from 0 at epoch 0 to 1 at the final epoch to avoid posterior collapse), the context head learns representations of action complementary to representations of movement produced by the Therblig Encoder.

5. Experiments

Our experiments over the tasks of action recognition and action anticipation explore the extent to which our novel context formulation helps in the modeling of action. In the action anticipation setting, video leading up until $t_a = t - \tau_a$ (where τ_a is the offset anticipation time and t corresponds to the start time of the action) is fed to the model, which produces likelihoods corresponding to the action most likely to begin at t_a. We adopt $\tau_a = 1$ second for all action anticipation experiments.

5.1. Datasets

5.1.1 EPIC Kitchens

The EPIC Kitchens 100 dataset contains unscripted, egocentric activity of roughly 100 hours of activity in kitchen environments. The dataset is annotated with non-overlapping action clips paired with verb and object labels (v, o). There are roughly 125 verbs and 300 objects, making for a total of $2,514$ unique actions. We augment the EPIC Kitchens dataset with our Therblig annotations and refer readers to the Therblig annotation process described in [13]. Results are reported over the official validation set.

5.1.2 50 Salads

The 50 Salads dataset contains 50 long sequences of scripted activity involving the preparation of a salad. Each sequence ranges from 5 to 10 minutes long, and contains 35 unique actions (i.e. *cut tomato*). While the dataset includes accelerometer information and depth, we only rely on the RGB video.

5.2. Models

We aim to demonstrate the ability of our approach towards improving the classification accuracy of video architectures. We perform our experiments over four popular video architectures - **I3D** [9], **TimeSFormer** [6] (TimeSFormer-B) , **ViViT** [1] (ViViT-B/16x2) and **MoViNet** [19] (MoViNet-A3). All models are pre-trained over Kinetics 600. Due to resource constraints, we adopt these popular video architectures over existing state-of-the-art ensembles of architectures. For purposes of reproducibility, we describe all details in the Supplementary Materials.

Models	50 Salads	EPIC Kitchens
TimeSFormer [6]	39.7%/60.1%	41.0%/63.9%
ViViT [1]	36.9%/60.9%	40.3%/60.3%
I3D [9]	41.0%/59.6%	41.7%/64.7%
MoViNet [19]	41.9%/64.0%	42.7%/66.0%

Table 1. **Therblig prediction** accuracies for backbone architectures over EPIC Kitchens and 50 Salads datasets. Results shown as order-aware and order-unaware accuracy, respectively.

5.3. Evaluation

Table 1 evaluates the performance of each Therblig Encoder backbone over the Therblig sequence prediction task. Element-wise accuracy alone is a harsh evaluation metric due to its strict ordering requirement - we include an order-unaware accuracy metric as well, considering a predicted Therblig element correct if it exists in the ground truth, irrespective of its place within the ground truth Therblig sequence. This metric takes the subset of predicted Therbligs for which a 1-1 mapping can be constructed to the ground truth Therblig sequence, and defines accuracy as the cardinality of this set over the length of the predicted Therblig sequence.

In Tables 2 and 3 we showcase the results of our proposed method using classification accuracy as our metric of choice. We evaluate our approach over the tasks of action recognition and action anticipation.

For both tasks, we report our results as follows: **Model** corresponds to the original single-stream, base video architecture, **Model (+T)** corresponds to the complete two-stream architecture without the addition of the L_{MI} loss

	50 Salads	EPIC Kitchens
I3D	64.2%	65.1%/49.7%/39.2%
I3D (+O)	66.9%	67.4%/50.1%/40.9%
I3D (+T) [13]	69.1%	68.2%/51.9%/41.6%
I3D (+TC)	**71.0%**	**69.5%/54.0%/43.5%**
TimeS	73.1%	62.1%/55.4%/41.1%
TimeS (+T) [13]	76.4%	67.6%/57.0%/44.0%
TimeS (+TC)	**77.9%**	**68.0%/57.9%/44.9%**
ViViT	72.1%	62.4%/56.0%/43.3%
ViViT (+T) [13]	74.3%	66.4%/56.9%/45.9%
ViViT (+TC)	**75.2%**	**67.8%/58.1%/47.6%**
MoViNet	73.9%	67.9%/52.9%/41.1%
MoViNet (+O)	76.0%	69.3%/53.1%/42.9%
MoViNet (+T) [13]	76.5%	69.1%/55.0%/43.8%
MoViNet (+TC)	**77.7%**	**69.9%/56.7%/45.0%**

Table 2. **Action recognition** accuracies over EPIC Kitchens and 50 Salads datasets. Results under EPIC Kitchens are provided as: verb/object/action prediction accuracies, respectively.

	50 Salads	EPIC Kitchens
I3D	39.5%	31.1%/18.3%/10.1%
I3D (+O)	39.3%	29.4%/18.4%/9.9%
I3D (+T) [13]	43.1%	33.0%/18.9%/10.9%
I3D (+TC)	**45.0%**	**32.9%/20.9%/12.2%**
TimeS	48.6%	31.6%/28.2%/13.6%
TimeS (+T) [13]	51.4%	33.9%/28.5%/14.8%
TimeS (+TC)	**53.3%**	**34.5%/29.8%/15.7%**
ViViT	45.6%	31.9%/29.8%/13.9%
ViViT (+T) [13]	49.1%	33.7%/29.4%/14.7%
ViViT (+TC)	**51.6%**	**34.3%/30.8%/15.6%**
MoViNet	46.4%	34.2%/25.6%/13.1%
MoViNet (+O)	46.9%	34.0% / 25.9% /13.3%
MoViNet (+T) [13]	48.2%	36.1% / 26.9% /14.0%
MoViNet (+TC)	**49.7%**	**36.9%/27.5%/14.7%**

Table 3. **Action anticipation** accuracies over EPIC Kitchens and 50 Salads datasets. Results under EPIC Kitchens are provided as: verb/object/action prediction accuracies, respectively.

component, **Model (+O)** corresponds to a two-stream architecture with RGB and optical flow frames processed separately and concatenated for action classification and **Model (+T+C)** corresponds to the complete architecture as proposed in Section 4.

Finally, in Figure 5 we extract representations of action immediately prior to the final action classification layer of the I3D model and display low-dimensional projections using t-SNE for action groupings which are difficult to disambiguate using motion characteristics alone. We do this for the original single-stream base video architecture along with our proposed two-stream framework to demonstrate how incorporation of context allows for better understanding of actions highly similar at the level of movement.

6. Discussion

We observe the findings reported in [17] - that Transformers underperform with respect to architectures built on 3D Convolutions when it comes to the capturing of fine-grained motion - are in alignment with the results we observe in Table 1, where even an I3D model outperforms both Transformer architectures on the Therblig prediction task.

We observe sizable improvements for each of the I3D, MoViNet, TimeSFormer and ViViT models over EPIC Kitchens and 50 Salads in both action recognition and action anticipation (shown in Table 2 and Table 3, respectively). It can be seen based on the difference in performance between $Model$ and $Model(+T)$ that the Therbligs play a particularly large role in driving accuracies upwards. A likely reason for this is the physical nature of both EPIC Kitchens and 50 Salads. In both tasks, the low-level physical activity between the hands and objects of manipulation plays a pronounced role over contextual cues in the understanding of the action being performed. We note the incorporation of optical flow results in marginal gains over baselines, especially in the task of action anticipation where the incorporation of optical flow even decreases performance.

Our context formulation provides sizable benefits across all models and datasets, beyond the benefit provided by Therblig annotations alone. Contextual cues are important in the action recognition setting, but are even more important in the action anticipation setting where the physical aspects of action are unobserved. In accordance with this intuition, we observe larger relative improvements due to our modeling of context in action anticipation over action recognition.

Furthermore, we find it motivating that our context formulation provides improvements across both datasets, and especially so in 50 Salads, where the action ontology is centered around low-level movement and contextual information is primarily limited to objects, surrounding activity, and timings. Context plays a significantly larger role in EPIC Kitchens where it becomes useful in disambiguating between actions highly similar in motion characteristics.

To demonstrate that the Therblig-Model primarily captures representations of movement, we point the reader to spatio-temporal GradCAM visualizations (see here). The spatial activation maps of the Therblig-Model are centered over the hands and are tied to the movement of the hands (activation magnitudes are high during periods of move-

Figure 5. t-SNE visualizations over (LEFT: subfigures **a** and **c**) I3D video representations without our context formulation, and (RIGHT: subfigures **b** and **d**) I3D video representations with our context formulation. Features c_i are extracted immediately following the output of the Context Encoder. The groupings are selected based on the degree of similarity shared by the physical movements they comprise. The verbs in the groupings are as follows: (TOP: subfigures **a** and **b**) wash, pour and mix, (BOTTOM: subfigures **c** and **d**) open and close.

ment, and low when hands are out of view or stationary).

See Figure 5 for t-SNE visualizations over features belonging to action sets which are difficult to distinguish based on motion characteristics alone. Our approach improves separability between actions sharing not only similar motion, but whose associated contexts overlap, such as *wash*, *pour* and *mix*. Inspired by these findings, we plan to explore future domains involving more complex notions of context.

7. Conclusion

We have argued that action decomposes into base movement and *context*, further arguing that separate representations for each provides benefits. We employ Therbligs as a consistent, expressive, contact centered representation through which to model body motion. We present an approach for modeling context based on complementarity: removing base motion from action modeling, through use of a contrastive mutual information loss set against Therblig representations, produces context representations. We demonstrate the utility of our model across experiments in action recognition and action anticipation, showing sizable improvements across a variety of models, and across EPIC Kitchens 100 and 50 Salads datasets. All code will be made publicly available upon paper acceptance.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 6

[2] Sadjad Asghari-Esfeden, Mario Sznaier, and Octavia Camps. Dynamic motion representation for human action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 557–566, 2020. 3

[3] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019. 3

[4] Ralph M Barnes. Motion and time study. 1949. 2, 3

[5] Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context r-cnn: Long term temporal context for per-camera object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13075–13085, 2020. 3

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 6

[7] Malik Boudiaf, Jérôme Rony, Imtiaz Masud Ziko, Eric Granger, Marco Pedersoli, Pablo Piantanida, and Ismail Ben Ayed. A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. In *European conference on computer vision*, pages 548–564. Springer, 2020. 3

[8] Diane M Browder, Levan Lim, Chien-Hui Lin, and Phillip J Belfiore. Applying therbligs to task analytic instruction: A technology to pursue? *Education and Training in Mental Retardation*, pages 242–251, 1993. 2, 3

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[10] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1130–1139, 2018. 3

[11] Anthony Cioppa, Adrien Deliege, Silvio Giancola, Bernard Ghanem, Marc Van Droogenbroeck, Rikke Gade, and Thomas B Moeslund. A context-aware loss function for action spotting in soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13126–13136, 2020. 3

[12] Dima Damen, Hazel Doughty, Giovanni Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1, 2020. 2, 3

[13] Eadom Dessalene, Michael Maynord, Cornelia Fermüller, and Yiannis Aloimonos. Therbligs in action: Video understanding through motion primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10618–10626, June 2023. 2, 3, 4, 6, 7

[14] Alvin I Goldman. *Theory of human action*. Princeton University Press, 2015. 1

[15] Pranay Gupta, Anirudh Thatipelli, Aditya Aggarwal, Shubh Maheshwari, Neel Trivedi, Sourav Das, and Ravi Kiran Sarvadevabhatla. Quo vadis, skeleton action recognition? *International Journal of Computer Vision*, 129(7):2097–2112, 2021. 3

[16] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3148–3159, 2022. 3

[17] Ziyuan Huang, Zhiwu Qing, Xiang Wang, Yutong Feng, Shiwei Zhang, Jianwen Jiang, Zhurong Xia, Mingqian Tang, Nong Sang, and Marcelo H Ang Jr. Towards training stronger video vision transformers for epic-kitchens-100 action recognition. *arXiv preprint arXiv:2106.05058*, 2021. 7

[18] Seung-kook Jun, Pankaj Singhal, Madusudanan Sathianarayanan, Sudha Garimella, Abeer Eddib, and Venkat Krovi. Evaluation of robotic minimally invasive surgical skills using motion studies. In *Proceedings of the Workshop on Performance Metrics for Intelligent Systems*, pages 198–205, 2012. 2, 3

[19] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16020–16030, 2021. 6

[20] Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Information maximizing visual question generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2008–2018, 2019. 3

[21] EJ MS and Mccormick Ej. Human factors engineering design. *National Defense Industry Press*, 1992. 2, 3

[22] Xiaojiang Peng and Cordelia Schmid. Multi-region two-stream r-cnn for action detection. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 744–759. Springer, 2016. 4

[23] Suman Saha, Gurkirt Singh, and Fabio Cuzzolin. Two-stream amtnet for action detection. *arXiv preprint arXiv:2004.01494*, 2020. 4

[24] Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Ortner. Learning disentangled representations via mutual information estimation. In *European Conference on Computer Vision*, pages 205–221. Springer, 2020. 3

[25] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2020. 3

[26] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 4

[27] Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738, 2013. 3

[28] An Tran and Loong-Fah Cheong. Two-stream flow-guided convolutional attention networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3110–3119, 2017. 4

[29] Xuanhan Wang, Lianli Gao, Peng Wang, Xiaoshuai Sun, and Xianglong Liu. Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length. *IEEE Transactions on Multimedia*, 20(3):634–644, 2017. 4

[30] Mike Wu, Chengxu Zhuang, Milan Mosse, Daniel Yamins, and Noah Goodman. On mutual information in contrastive learning for visual representations. *arXiv preprint arXiv:2005.13149*, 2020. 3

[31] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. Rgb-d saliency detection via cascaded mutual information minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4338–4347, 2021. 3

[32] Shuai Zhao, Yang Wang, Zheng Yang, and Deng Cai. Region mutual information loss for semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 3