Therbligs in Action: Video Understanding through Motion Primitives

Anonymous authors
Paper under double-blind review

Abstract

In this paper we introduce a rule-based, compositional, and hierarchical modelling of action using Therbligs as our atoms - a consistent, expressive, contact-centered representation of action. Over these atoms we introduce a differentiable method of rule-based reasoning to regularize for logical consistency. Our approach is complementary to other approaches in that the Therblig-based representations produced by our architecture augment rather than replace existing architectures' representations. We release the first Therblig-centered annotations over two popular video datasets - EPIC Kitchens 100 and 50-Salads. We evaluate our system for the task of action segmentation, demonstrating a substantial improvement using a base GRU architecture over baseline of 5.6% and 4.1% (14.4% and 6.5% relative) increase in accuracy (and increases with respect to all other metrics as well) over EPIC Kitchens and 50-Salads, respectively. We also demonstrate benefits to adopting Therblig representations for two state-of-the-art approaches - MSTCN++ and ASFormer - observing a 10.3%/10.7% relative improvement, respectively, over EPIC Kitchens and 9.3%/6.1% relative improvement, respectively, over 50 Salads. All code and data is to be released upon paper acceptance.

1 Introduction

The question of how to structure action is non-trivial, and one which action datasets by-and-large do not resolve. The typical solution is to intuit a set of actions belonging to a given domain, and label each datapoint of the dataset as belonging to one of these action categories. If attention is not given to relations among actions, difficulties can arise in knowing the correct label to assign an action segment.

Existing action understanding approaches typically rely on visual representations of input (e.g., I3D features (Carreira and Zisserman, 2017)), high level categorizations (e.g., sequences of action labels), or both. This gives rise to **limitation #1**: a large representational gap between visual features and symbolic action labels.

Additional challenges come from the existence of segments which could reasonably be classified into multiple action categories, or into no action category. However, data annotations (as well as vision models) typically assume frames are associated with one and only one label. This assumption is broken by **limitation #2**: a difficulty in defining action boundaries [1], **limitation #3**: action categories which are not mutually exclusive (e.g., *take* and *move* as defined in EPIC Kitchens (Damen et al., 2020)). These inconsistencies force annotators to rely on subjective assessment to assign action labels and their boundaries.

We introduce a framework confronting the above described limitations. This framework involves: compositionality; hierarchy; rule based modeling; contact. In the realization of this framework, we propose the use of *Therbligs* - a low-level mutually exclusive contact demarcated set of sub-actions. These Therbligs are consistent in that a given action segment has only a single Therblig representation, and Therbligs are expressive in that they capture the meaningful physical aspects of action relevant to action modeling. Therbligs were introduced in the early 20th century as a set of 18 elemental motions used to analyze complex movement - see the Appendix for a brief historical

[1] See (Alwassel et al., 2018) for a case study on how annotators have difficulties coming to a consensus on when actions begin and end.

background. We adopt 8 Therbligs pertaining to those involving the manipulation of objects. See Figure 1 for our Therblig set.

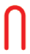
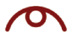
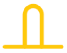
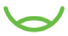
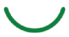
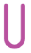
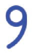
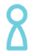
Symbol	Name	Description
	Grasp (**G**)	When the worker's hand grabs the object
	Release (**R**)	The releasing of the object when it reaches its destination
	Hold (**H**)	The retention of an object such that it undergoes no movement while in operation
	Move (**M**)	Moving a loaded hand to the point of 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
	Position (**P**)	Positioning an object for the next Use operation
	Preposition (**Pp**)	Positioning an object for the next Use operation, relative to an approximate location

Figure 1: Listed above are the Therbligs we select, their symbolic illustrations, and brief descriptions of their usage.

The benefits of a Therblig-centered framework include compositionality & hierarchy; rule based reasoning; resolution of semantic ambiguity; contact-centered precision of temporal boundaries of action; propogation and projection of action dynamics from the past and into the future.

Contact transitions demarcate Therblig boundaries, giving Therbligs a consistency which methods relying on annotators' intuited demarcations lack. Between points of contact exist contact states represented by a binary class (contact, no contact) for each object present, which are wholly captured by Therbligs. As objects in contact are the primary objects of interaction and define the space of possible actions, they provide meaningful information for the modelling of action.

Therblig atoms are then composable into higher entities, including full actions. These actions are in turn composable into sequence constituting activities. We then have the hierarchy of representation illustrated in Figure 2. At the lowest, and instantaneous, level are points of contact, between which exist Therbligs with temporal extension, on top of which exist action, permutations of which constitute longer activities.

Architectures built upon Therbligs for the modelling of action gain temporal precision through points of contact as well as meaningful information captured by contact states. Therbligs also exhibit semantic mutual exclusivity in that there is one and only one Therblig interpretation of a sequence, as opposed to the many interpretations when action labels are intuited [2], leading to semantic ambiguity. As a consequence of Therbligs, semantic ambiguity at the action-level is constrained by the deeper grounding of action labels in explicit action dynamics (see Figure 2). And unlike higher level actions, Therbligs enable the imposing of a contact-based logic defining their preconditions and postconditions in the form of the state of contact before and after them. For example, an object being *moved* must be preceded by *grasp* and proceeded by a *release*. These rules interface at the Therblig level of the hierarchy.

These rules provide benefits in that they 1) allow for bias towards consistency between contact states and Therblig predictions within a loss term, as described in Section 3.2, 2) allow the enforcement of logical consistency over Therblig sequences and in turn, resolution of action sequences. 3) allow for modelling of relations over greater time spans, and 4) aid in resolution of inconsistent interpretations of action.

In producing sub-action level symbolic representations, our proposed hierarchical architecture is comprised of two main components; the **Therblig-Model**, which maps I3D features to Therbligs; and, the **Action-Model**, which maps Therbligs and I3D features to actions. The Therblig-Model is optimized over a loss including structure-aware terms for contact consistency and Therblig consistency by incorporating differentiable reasoning. Figure 3 illustrates our architecture. This architecture is complementary to, rather than in competition with, existing architectures for action modeling - Therblig representations can be easily integrated through concatenation with existing feature representations. We demonstrate this with two state-of-the-art approaches to action segmentation - MSTCN++ (Li et al., 2020) and ASFormer (Yi et al., 2021).

We evaluate our approach over the task of action segmentation - the task of assigning each frame of a video sequence to action labels. The challenges facing action modeling generally apply to the

[2] Some additional structure is needed for complete mutual exclusivity - see the Appendix for discussion on this structure.

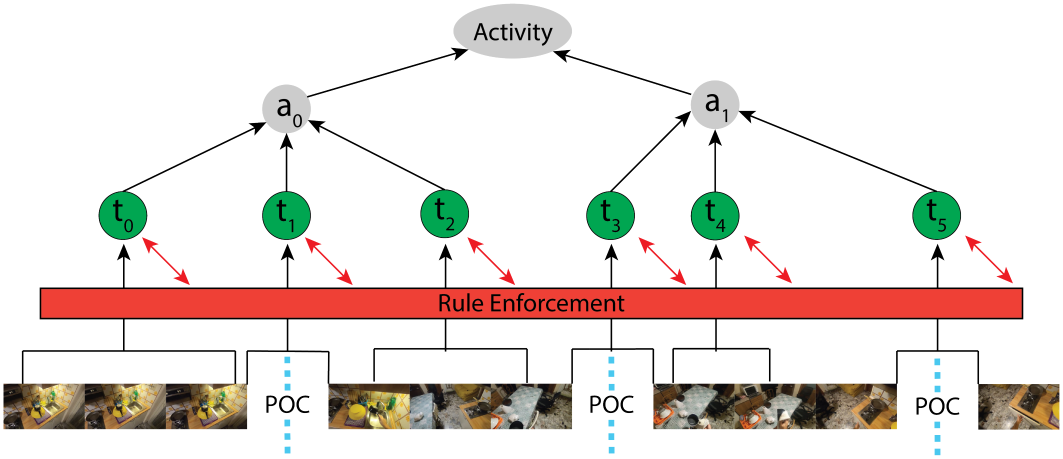


Figure 2: We introduce the use of Therbligs (t_i) in video understanding as a consistent, expressive, symbolic representation of sub-action. Points of Contact (indicated by the divider dashes) are necessarily associated with Therbligs and/or their boundaries. Because of the unambiguity of Points of Contact, Therblig boundaries gain precision and are non-overlapping. On top of Therblig atoms we construct a framework for Rule Enforcement, enforcing greater logical consistency through commonsense rules. This rule-based framework allows for the easy introduction of long-term constraints. Therblig atoms are then composable into actions (a_i), which are in turn composable into activities.

task of action segmentation: imprecise action boundaries; issues of concurrent action categories; inconsistencies in annotations; the tractability of modelling videos, particularly those longer than a few seconds; the over-reliance on appearance cues over long-term temporal semantics. Therbligs present a solution to these challenges. We evaluate over the EPIC Kitchens 100 and 50-Salads datasets.

The primary contributions of our work are as follows:

- Therbligs, a consistent, expressive symbolic representation of sub-action centered on contact.
- Rules: Flexible and differentiable constraining of intuitive constraints on arrangement of atomic actions, informed by therblig ontology.
- Novel hierarchical architecture composed of a Therblig-Model and Action-Model. Representations produced by the Therblig-Model can be easily integrated into other approaches, which we demonstrate with MSTCN++ and ASFormer.
- Dataset: We release the first Therblig-centered annotations over two popular video datasets.

The rest of this paper is structured as follow: Section 2 discusses related works, Section 3 introduces our proposed method, Section 4 describes the experiments, and in Section 5 we conclude.

2 Related Works

2.1 Sub-Action Video Datasets

There exist several datasets that provide sub-action level annotations as a means of resolving semantic and temporal ambiguity in annotation, and enabling the hierarchical modelling of action (Ji et al., 2020; Shao et al., 2020a;b). FineGym (Shao et al., 2020a) introduces fine-grained action annotations for actions in gymnastics, but suffers from a difficult and expensive data collection process. TAPOS (Shao et al., 2020b) manually breaks actions into sub-actions for Olympics videos via temporal action parsing. Other datasets producing sub-action level annotations are of an instructional nature (Kuehne et al., 2014; Rohrbach et al., 2015; Stein and McKenna, 2013), providing annotations for

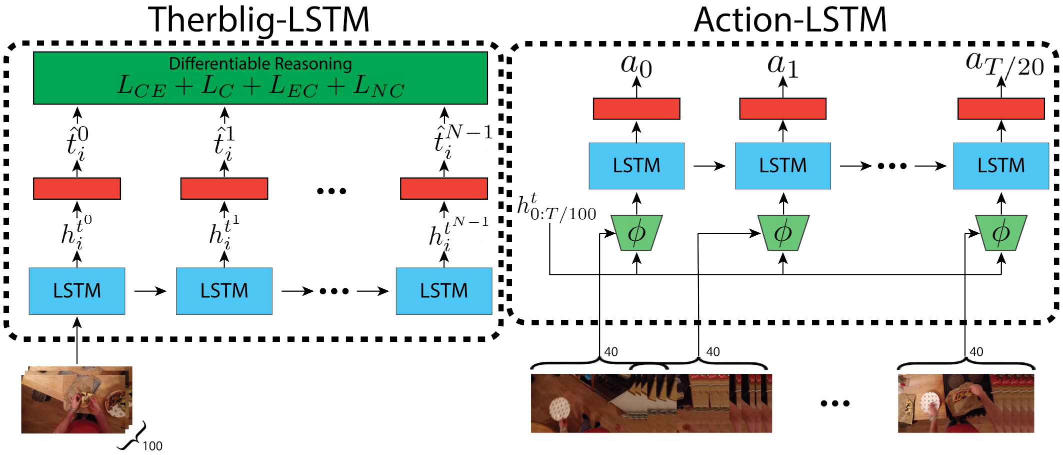


Figure 3: Architectural diagram of our framework. **Therblig-Model** takes a stack of $K = 100$ frames as input, then feeds their I3D representations to a 2-layer GRU (LSTM), which in turn produces hidden states h_i^t. Hidden states h_i^t are fed to fully connected layers, followed by a Gumbel-Softmax operation, producing Therblig predictions amenable to differentiable reasoning. **Action-Model** takes a sliding window with window size $W = 40$ over the original video sequence with stride $s = 20$. These windows' I3D representations are fed to ϕ, an attention mechanism consisting of a 2 layer MLP - this MLP attends over the hidden states produced by **Therblig-Model**. The blended features produced by ϕ are fed to a 2-layer GRU (LSTM) followed by a fully connected layer predicting action class likelihoods $a_{0:T/20}$. See Sections 3.1.1 and 3.1.2 for details.

steps of various cooking activities. Our Therblig annotations 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.

2.2 Contact in Action Understanding

Contact has proven to be a useful feature in several tasks of interest in computer vision, such as hand-object pose estimation (Cao et al., 2021; Karunratanakul et al., 2020; Shan et al., 2020), character animation, kinematic pose estimation (Rempe et al., 2020), etc. However, the vast majority of contact-centered approaches apply in the single-image setting, and few works (Ji et al., 2020) consider modelling of contact for use in action understanding, despite contact being a defining characteristic of all physical human interaction. Ego-OMG (Dessalene et al., 2020) approaches the task of action anticipation, representing long sequences of manipulation activity through sequences of discrete states, each state delineated by the making and breaking of contact. Rather than directly extract contact from each video frame as in (Dessalene et al., 2021; 2020), we instead propose the adoption of Therbligs, a contact-centered representation governed by contact-based commonsense rules.

2.3 Action Segmentation

Action segmentation is the task of assigning each frame in a video to a particular class of action. The bulk of approaches to action segmentation involve better methods of temporal aggregation and/or better methods of action boundary estimation (Ahn and Lee, 2021; Ishikawa et al., 2021; Huang et al., 2020; Wang et al., 2020; Chen et al., 2020). We stress that our Therblig framework is complementary to such approaches and models. Therbligs can be temporally aggregated over long timespans thanks to their low-level symbolic nature and the existence of commonsense rules governing their usage. Due to their grounding in contact events in video, Therbligs also open up a variety of exciting opportunities for action boundary estimation via the modelling of contact transitions.

3 METHODS

For video segment s_i with T frames, the Therblig representation t_i is a sequence of N Therblig tuples, and the contact representation c_i represents the objects in contact at the end of that video segment. 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 $V = \{Re, M, G, R, P, Pp, U, H\}$. In other words, v_j indicates the Therblig verb and each o_j indicates the object of interaction. We set the maximum number of possible Therblig annotations per sequence N to 6. Each contact annotation c_i is a tuple of the form (c_i^r, c_i^l), where c_i^r corresponds to the class of the object held by the right hand, and c_i^l corresponds to the class of the object held by the left hand.

Given video segment s_i, our goal is to infer the action class likelihoods of each frame. We do this by means of a novel hierarchical architecture as described in subsection 3.1. This architecture consists of two levels; a Therblig-Model 3.1.1 and an Action-Model 3.1.2. We then describe our rule-based reasoning formulation in subsection 3.2, and detail the Therblig annotation collection process in subsection 3.3.

3.1 ARCHITECTURE

The architecture of our proposal is illustrated in Figure 3. We apply an I3D network (pre-trained over Kinetics-400) for each segment $s_i \in S$ where S is composed of video segments $S = \{s_0, ..., s_{T/100}\}$, where T is the number of frames in S. This results in I3D features $F = \{f_1, ..., f_{T/100}\}$. Our Therblig-Model predicts a sequence of Therbligs $\hat{t}_i$ for each $f_i \in F$. Our Action-Model takes the representations produced by Therblig-Model along with S, and produces per-frame action class likelihoods.

As the Therblig annotations t_i and action annotations a_i do not exhibit one-to-one overlap between their respective video sequences, the Therblig-Model and Action-Model are trained separately. It is otherwise possible for the two models to be trained in end-to-end fashion.

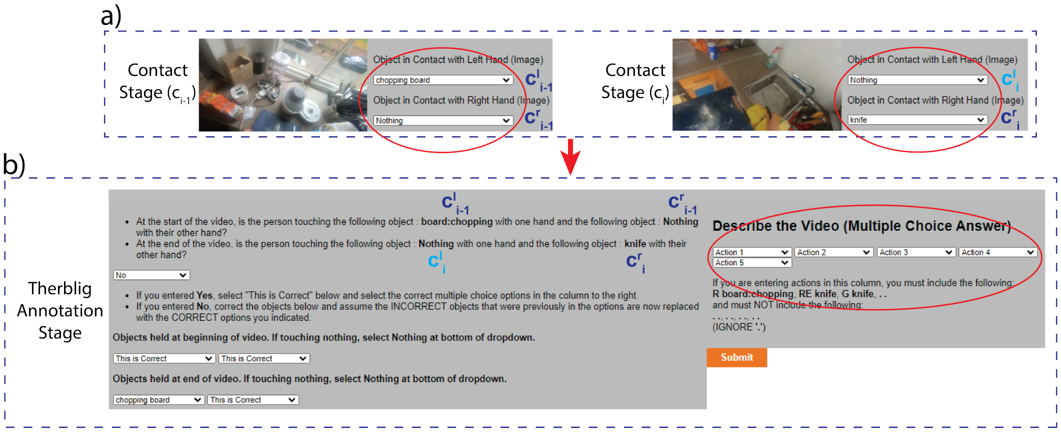
3.1.1 THERBLIG-MODEL

We adopt a 2-layer GRU as our primary base architecture for Therblig-Model. Due to the lack of precise temporal alignment between the input video segments $s_i \in S$ and the Therblig annotations t_i, we adopt an encoder-decoder schema as follows: The hidden state of the Therblig-Model is set to f_i, and the network is rolled out to iteratively predict a sequence of Therbligs $\hat{t}_i = \{\hat{t}_i^0, ..., \hat{t}_i^5\}$, feeding $\vec{0}$ as the initial input, and outputs of previous hidden layers as the inputs to the decoder for subsequent timesteps. We adopt the practice of teacher forcing, where the outputs of previous hidden layers are occasionally replaced with the ground truth, with probability $p = 0.5$. After training the Therblig-Model for 50 epochs and selecting the model instance with the top validation accuracy, we freeze the model for the training of the Action-Model. We additionally incorporate two state-of-the-art base architectures (MSTCN++ (Li et al., 2020) and ASFormer (Yi et al., 2021)), as a replacement to the architecture described above to, with training procedures outlined in the respective papers.

3.1.2 ACTION-MODEL

We adopt a 2-layer GRU as our primary base architecture for Action-Model. The Action-Model computes I3D representations over sliding windows of size $W = 40$, taken with a stride of $s = 20$ from input video S of T frames. This produces features $f_j \in \{f_0, ..., f_{T/20}\}$. We pair indices i from h_i^t and j from f_j by cross-referencing the closest times in video S associated with i and j. I3D features f_j and the Therblig-Model hidden state output sequence h_i^t are then fed to the Action-Model network, producing action segmentation predictions $\hat{a}_j \in \{a_0, ..., a_{T/s}\}$. As for the ASFormer base architecture, for each Therblig prediction $\hat{t}_i^t$ we extract the Therblig representations at the level immediately prior to the fully connected layers of the decoder, instead of using h_t^t. For the MSTCN++ base architecture, for each Therblig prediction $\hat{t}_i^t$ we extract the Therblig representations immediately prior to the refinement step instead of using h_t^t.

For the GRU base architecture of Therblig-Model, we adopt a temporal attention mechanism ϕ, feeding it I3D features f_j to output learned attention weights $\alpha_i^a = \{\alpha_i^{a^0}, ..., \alpha_i^{a^{N-1}}\}$ over hidden



Object in Contact with Left Hand (Image)

chopping board

Object in Contact with Right Hand (Image)

Nothing

Object in Contact with Left Hand (Image)

Nothing

Object in Contact with Right Hand (Image)

knife

- At the start of the video, is the person touching the following object : **board:chopping** with one hand and the following object : **Nothing** with their other hand?
- At the end of the video, is the person touching the following object : **Nothing** with one hand and the following object : **knife** with their other hand?

No

- If you entered **Yes**, select "This is Correct" below and select the correct multiple choice options in the column to the right.
- If you entered **No**, correct the objects below and assume the INCORRECT objects that were previously in the options are now replaced with the CORRECT options you indicated.

Objects held at beginning of video. If touching nothing, select Nothing at bottom of dropdown.

This is Correct | This is Correct

Objects held at end of video. If touching nothing, select Nothing at bottom of dropdown.

chopping board | This is Correct

Describe the Video (Multiple Choice Answer)

Action 1 | Action 2 | Action 3 | Action 4 | Action 5

If you are entering actions in this column, you must include the following:
R board:chopping, **RE knife**, **G knife**, **. .**
and must NOT include the following:
. ., . ., . ., . .
(IGNORE '.')

Figure 4: An overview of our two-stage croudsourcing pipeline. The **Contact Annotation Stage** provides an image (shown) paired with a video (not shown) and asks the user to indicate the objects held by the actor in the image/video via multiple choice (circled). In the **Therblig Annotation Stage**, an annotator first validates the correctness of c_i and c_{i-1} and then produces Therblig annotations t_i. As can be seen above, in the Contact Annotation Stage a worker mistakenly indicated that nothing was held by the left hand for c_i. In the Therblig Annotation Stage, after correcting these erroneous contact annotations, a new worker annotates the sequence of Therbligs t_i (circled) - the multiple choice options shown to this worker have been filtered so as to achieve consistency with rules discussed in 3.2.1.

layer outputs $\hat{h}_i = \{\hat{h}_i^{t^0}, ..., \hat{h}_i^{t^{N-1}}\}$ (adopting the extracted Therblig representations instead of $\hat{h}_i$, for ϕ in MSTCNN++/ASFormer), producing blended hidden state features. For the GRU base architecture of Action-Model, those blended features from Therblig-Model are fed stepwise through the GRU, whose outputs are fed to fully connected layers to predict action class $\hat{a}_j$. For the prediction of action class $\hat{a}_j$ in the MSTCN++/ASFormer base architectures of Action-Model, those blended features from Therblig-Model are concatenated with existing I3D-derived representations prior to the application of the first fully connected layer.

3.2 Therblig Rules

Therbligs enable the introduction of contact-centered rules that 1) provide significant structure to the annotation process and 2) provide structure during training in the form of separate differentiable loss components. See Sections 3.2.1 and 3.2.2 for more.

3.2.1 Explicit Rules

Below we enumerate the explicit rules we introduce over the Therblig ontology.

Rule 1 The additions and subtractions of objects in contact produced by the Therblig sequence linking c_i and c_{i+N-1} must produce object contact set c_{i+N-1} from c_i.

Rule 2 Objects in contact state c_i cannot be *grasped* or *reached* without first being *released*.

Rule 3 Objects not in contact state c_i cannot be *moved*, *(pre)positioned*, *used*, or *released* without first being *grasped*.

These rules structure the second stage of croudsourcing annotations in the form of a filter over all possible annotations - see Section 3.3 for details.

3.2.2 Differentiable Rules

We wish to incorporate the rules discussed in Section 3.2.1 into the training of the Therblig-Model. However, the rules are non-differentiable, and as such we approximate each rule with differentiable closed-form expressions. However, the logic requires discrete representations of Therblig predictions. As the softmax activation outputs are continuous, they must be converted into their discrete equivalents. The *argmax* operation is non-differentiable, and so we adopt the use of Gumbel Softmax as a differentiable alternative, applying it over the pre-softmax features of the Therblig-Model to arrive at discrete one-hot-encodings of Therbligs ($\hat{g_i}$) while maintaining differentiability.

The rules are represented as follows:

Rule 1 Cross Contact-Therblig Consistency Loss

$$L_C = \sum_{k=1}^{5} \|c_i + \beta \hat{g}_i^k - c_{i+1}\| \tag{1}$$

Rule 2 Contact Enforcement Loss

$$L_{EC} = \sum_{k=1}^{5} \|a_{i,k} - \gamma \hat{g}_i^k\| \tag{2}$$

$$\text{where } a_{i,k} = a_{i,k-1} + \beta \hat{g}_i^{k-1} \text{ and } a_{i,0} = c_i$$

Rule 3 Non-Contact Enforcement Loss

$$L_{NC} = \sum_{k=1}^{5} \|a_{i,k} - \delta \hat{g}_i^k\| \tag{3}$$

$$\text{where } a_{i,k} = a_{i,k-1} \beta \hat{g}_i^{k-1} \text{ and } a_{i,0} = c_i$$

Each of β, γ, and δ, correspond to a vector of length 8, each index of which corresponds to a single Therblig verb. For β, values take on 1 for *grasp*, -1 for *release*, and 0 otherwise. For γ, values take on 1 for *reach* and *grasp*, 0 otherwise. For δ, values take on 1 for *move*, *(pre)position*, *use*, or *release*, and 0 otherwise. Loss component L_C of Rule 1 measures the offset of c_i and c_{i+1} against βg_i^k. Loss component L_{EC} of Rule 2 iteratively compares the Therblig-derived contact states $a_{i,k}$ against $\gamma \hat{g}_i^k$. Loss component L_{NC} iteratively compares $a_{i,k}$ against $\delta \hat{g}_i^k$.

We adopt each of these loss terms in addition to Categorical Cross Entropy loss L_{CE} to arrive at combined loss $L = L_{CE} + L_{NC} + L_{EC} + L_C$. These loss terms provide several benefits. For one, they relax the strict ordering constraint induced by our L_{CE} term, which helps due to the concurrently-overlapping execution of the Therbligs by the actor in the video. These loss terms also provide meaningful constraints in guiding the learning process, and we demonstrate this finding empirically in Section 4.2.

3.3 CROUDSOURCING

We croudsource our annotations in two stages; see Figure 4. The first stage involves the croudsourcing of objects in contact with the hands. The annotator is shown an image, along with a slightly slowed-down video roughly 5 seconds long. The addition of the video adds temporal context for images too difficult to annotate alone. The annotator is asked to indicate via multiple choice the objects in contact with the actor's hands at the end of the interaction. We pool 5 independently collected responses per image, taking the mode of the answers for consistency. The average time per assignment was 3.4 minutes.

In the second stage, the possible Therbligs for each item in the sequence t_i^j for $0 \leq j \leq N$ are determined by taking the cross product of the contact annotations c_i and c_{i+1} with $V = \{Re, M, G, R, P, Pp, U, H\}$ and filter the results such that *consistency* is observed with respect to the three rules defined in subsection 3.2.1. **We note there are** $1,067$ **possible Therblig** $(verb, object)$ **tuples; through our rules, we are able to significantly reduce the average number of multiple choice annotation possibilities to just** 19 $(verb, object)$ **tuples!** We set N, the number of Therblig tuples per video clip, to 6.

4 Experiments

Our experiments on action segmentation explore the extent to which we are able to predict Therbligs and the extent to which their incorporation benefits action segmentation[3]. We report our numbers by early stopping over validation accuracy for 3 independent runs, reporting only the mean. All code, data and results will be released upon acceptance.

4.1 Datasets

EPIC Kitchens We choose the EPIC Kitchens 100 dataset because of the benefits the egocentric perspective provides in allowing for full view of the hands and objects in contact. We augment portions of the EPIC Kitchens dataset with densely labelled Therbligs, for a total of $14,600$ croudsourced annotations. See the Appendix for details on the dataset. We report our results over the validation set provided by the original paper, and create an additional held-out validation set used for early stopping from 10% of the training set.

50 Salads We choose the 50 Salads dataset because each activity performed is strongly structured by the making and breaking of contact. We augment the entirety of the video in this dataset, for a total of $6,500$ croudsourced annotations. See the Appendix for details on the dataset. We bin videos into training/validation/testing according to a 80%:10%:10% split.

Full Data	50-Salads	EPIC Kitchens
L_{CE}	22.1%/3.19/2.25	9.5%/5.68/2.491
$L_{CE} + L_C$	23.0%/3.15/1.96	9.9%/5.209/2.124
$L_{CE} + L_{EC}$	23.2%/2.99/1.92	11.7%/5.238/1.90
$L_{CE} + L_{NC}$	23.1%/2.83/1.98	12.7%/**5.16**/1.91
All L	**25.1%/2.6/1.70**	**13.7%**/5.19/**1.83**
Low Data		
L_{CE}	12.1%/5.91/2.53	7.4%/5.37/2.36
All L	**16.9%/5.25/1.96**	**10.1%/5.21/1.86**

Table 1: Evaluation of Therblig-Model when trained over all Therblig annotations (Full Data) and when trained over a subset (Low Data). Results reported in order of : Accuracy ↑/Levenshtein Distance ↓/ Logical Consistency ↓.

4.2 Therblig Prediction

In this set of experiments, we answer the extent to which our Therblig-Model is capable of mapping video chunks s_i to the sequence of Therbligs $\hat{t}_i = \{\hat{t}_i^0, ..., \hat{t}_i^5\}$.

Metrics We evaluate our results, comparing $\hat{t}_i^j$ and t_i^j, for $0 \leq j \leq 5, \forall i$) over the following metrics of evaluation: Element-wise accuracy and Levenshtein distance (**L**). Element-wise accuracy metrics suffer from the strict ordering requirement; and so are not reflective of sequence-level similarity. Therefore we also evaluate over Levenshtein distance, the number of edits (insertions, deletions, and swaps) to transform $\hat{t}_i^j$ into t_i^j. In addition, we evaluate the logical consistency of our predictions as measured by the normalized number of violations per sequence of the rules described in 3.2.1.

Comparisons Table 1 illustrates the results of various forms of Therblig-Model over the EPIC Kitchens and 50 Salads datasets. L_{CE} refers to a simple, 2-layer bidirectional GRU trained solely over Categorical Cross Entropy. We train our GRU with and without each of the loss components discussed in Section 3.2.2. In addition, we include results when training over $L_{CE} + L_C + L_{EC} + L_{NC}$ and L_{CE}, but in the low-data setting where roughly 10% of our annotations are trained over for both datasets, highlighting the value of the structure defined in Section 3.2.2. See Table 1 for results.

4.3 Action Segmentation

In this experiment we evaluate the extent to which the incorporation of Therblig-Model benefits performance in action segmentation.

[3] See here for an exhaustive of videos paired with their corresponding Therblig annotations, contact state annotations, Therblig predictions and action annotations.

	50 Salads	EPIC Kitchens
Base GRU	62.8%/8.45/60.9	39.1%/19.3/23.2
GRU w. Therbligs	**66.9%/8.21/65.1**	**44.7%/13.5/25.6**
Base MST	78.19%/7.1/74.16	53.44%/10.4/54.38
MST w. Therbligs	**85.25%/6.7/81.01**	**58.75%/9.9/60.97**
Base ASF	82.97%/6.9/77.23	58.14%/10.3/56.41
ASF w. Therbligs	**88.2%/6.3/84.33**	**64.03%/9.5/61.46**

Table 2: Action Segmentation Results over EPIC Kitchens and 50 Salads datasets for frame-wise accuracy ↑/ segmental edit distance ↓/ and segmental F1-score@25 ↑.

Metrics As action segmentation is a classic task in computer vision, we rely on the works of (Huang et al., 2020; Ishikawa et al., 2021) in adopting the following evaluation metrics: frame-wise accuracy, segmental edit-score, and segmental F1-score. Frame-wise accuracy is used in all action segmentation works, whereas segmental edit-score and F1-score are most commonly used to penalize over-segmentation in particular.

Comparisons Table 2 illustrates the results of ablations of our proposed architecture over the EPIC Kitchens and 50 Salads datasets for the following base architectures of Therblig-Model and Action-Model: **GRU**, **MST** (MSTCN++), and **ASF** (ASFormer). The **Base** models correspond to the mapping of raw video to framewise action labels and the **w. Therbligs** models correspond to the entirety of our proposed framework.

5 Discussion

We stress that accuracy is a poor metric for the evaluation of Therblig-Model, due to its strict ordering requirement (e.g. misalignment between the predicted Therblig sequence and ground truth leads to an accuracy of 0% regardless of how many Therbligs were predicted correctly). As such, it under-reports the performance of Therblig-Model. To give better intuition we show predicted Therblig sequences alongside video of manipulation activity here.

We point the reader's attention towards the ”Low Data” results reported in Table 1, where we observe a large increase in accuracy through the incorporation of all the rule-based loss components. We believe this validates our hypothesis that the constraints imposed by the rules play a particularly outsized role in the low-data setting, where the model would otherwise have to infer the same commonsense structure, motivating future possible directions in the few-shot domain.

Finally, as demonstrated in Table 2, the incorporation of Therbligs results in superior performance over both the 50 Salads and EPIC Kitchens 100 datasets for all base architectures. While the accuracies of the base models we trained come close to matching the reported numbers in the original papers, the accuracy of the models trained with Therbligs reported in Table 2 outperform baseline models trained by us as well as those reported in original papers.

6 Conclusion

In this paper we have presented a method for mitigation of common limitations in action understanding approaches using a novel framework structured around Therbligs - a consistent, expressive, contact-centered representation of action. We demonstrate through ablation studies the utility of our proposed system components and demonstrate benefits to adopting Therblig representations for two state-of-the-art approaches - MSTCN++ and ASFormer - observing a 10.3%/10.7% relative improvement, respectively, over EPIC Kitchens and 9.3%/6.1% relative improvement, respectively, over 50 Salads. We hope the release of these annotations inspires future work towards the hierarchical modelling of action, and will release all code and data upon acceptance.

REFERENCES

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.

Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem. Diagnosing error in temporal action detectors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 256–272, 2018.

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.

Shi-Jie Li, Yazan AbuFarha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. *arXiv preprint arXiv:2110.08568*, 2021.

Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020.

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, 2020a.

Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Intra-and inter-action understanding via temporal action parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 730–739, 2020b.

Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014.

Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, pages 1–28, 2015. ISSN 0920-5691. doi: 10.1007/s11263-015-0851-8. URL http://dx.doi.org/10.1007/s11263-015-0851-8.

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.

Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021.

Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020.

Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9869–9878, 2020.

Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *European Conference on Computer Vision*, pages 71–87. Springer, 2020.

Eadom Dessalene, Michael Maynord, Chinmaya Devaraj, Cornelia Fermuller, and Yiannis Aloimonos. Egocentric object manipulation graphs. *arXiv preprint arXiv:2006.03201*, 2020.

Eadom Dessalene, Chinmaya Devaraj, Michael Maynord, Cornelia Fermuller, and Yiannis Aloimonos. Forecasting action through contact representations from first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Hyemin Ahn and Dongheui Lee. Refining action segmentation with hierarchical video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16302–16310, 2021.

Yuchi Ishikawa, Seito Kasai, Yoshimitsu Aoki, and Hirokatsu Kataoka. Alleviating over-segmentation errors by detecting action boundaries. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2322–2331, 2021.

Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14024–14034, 2020.

Zhenzhi Wang, Ziteng Gao, Limin Wang, Zhifeng Li, and Gangshan Wu. Boundary-aware cascade networks for temporal action segmentation. In *European Conference on Computer Vision*, pages 34–51. Springer, 2020.

Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9454–9463, 2020.

Ralph M Barnes. Motion and time study. 1949.

EJ MS and Mccormick Ej. Human factors engineering design. *National Defense Industry Press*, 1992.

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.

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.

A Appendix

A.1 Therbligs & Mutual Exclusivity

To model base movements, we adopt Therbligs - a set of 18 elemental motion primitives that capture 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 Barnes (1949), but have since been applied to represent and understand a wide variety of domains, i.e. assembly lines MS and Ej (1992), education Browder et al. (1993), surgery Jun et al. (2012), etc.

Therbligs from the definitions listed in Table 1 alone do not exhibit complete mutual exclusivity - they sometimes semantically overlap in instances involving possible *Use* and *Hold* Therbligs. For example, the action of turning on a faucet can be described with both of the following Therbligs in no particular order: *Hold faucet* and *Use faucet*. We address these instances by adding a slight change to the definition of *Use* and *Hold* so as to achieve complementarity between the two - *Use* is selected over *Hold* when an object's primary affordance is invoked in an action, regardless if the action involves holding the object, and *Hold* applies over *Use* when an object is otherwise being held in service of the *Use* of another object. That is, in the few instances these two definitions semantically overlap, the instructions inform the annotator to select *Use*.

As manipulation activity is typically two-handed, the Therblig sequences typically involve the interweaving of Therbligs performed by the right and left hands over different objects, leading to potential confusion on the part of the annotator as to the Therblig sequence's correct ordering. This is easily resolved via a post-processing step over the contact states (c_i^r, c_i^l) corresponding to the object of contact between the right and left hands, resulting in a Therblig sequence for the right hand and a Therblig sequence for the left hand.

A.2 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.

A.3 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 (e.g. *cut tomato*). While the dataset includes accelerometer information and depth, we only rely on the RGB video.

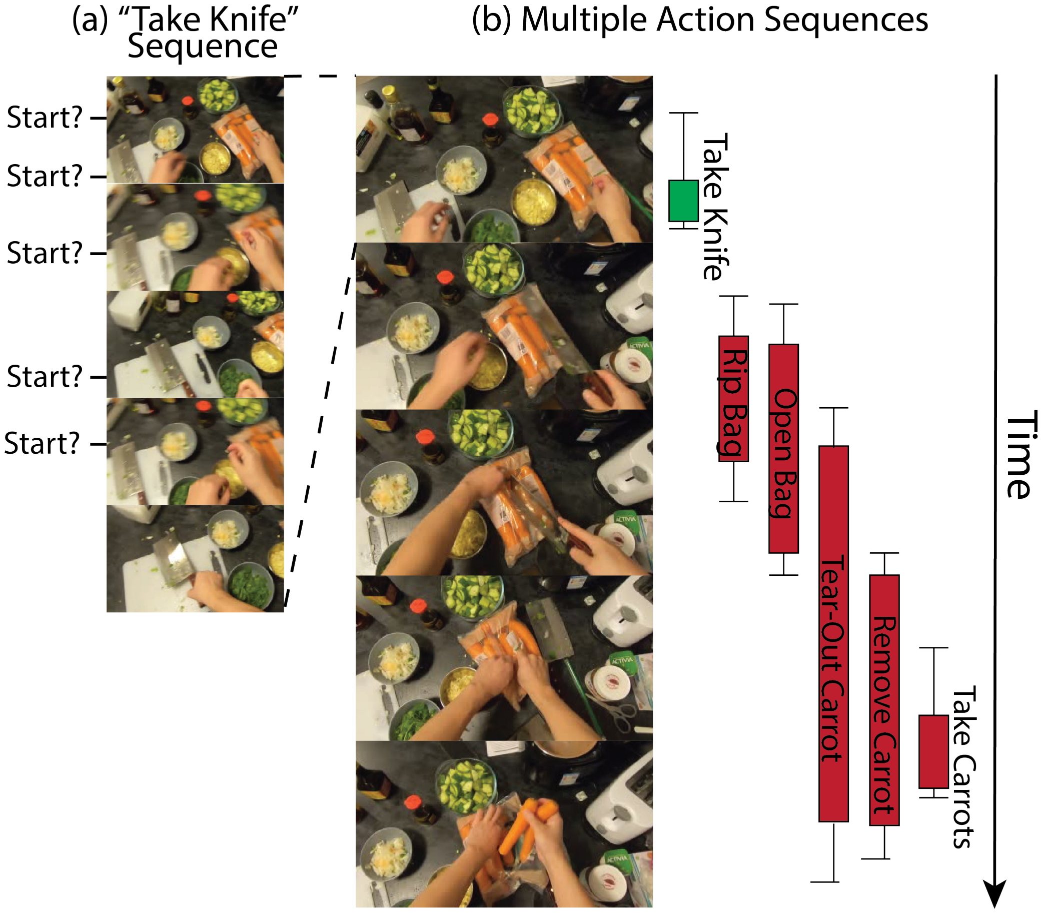


Figure 5: In this figure we illustrate; a) the difficulty of defining the precise starting point of action, as without contact subjective assessments vary; b) a series of boxplots (non-conflicting and conflicting) showing both temporal and semantic overlap of conflicting action labels which might be assigned in datasets without temporal and semantic mutual exclusivity.