Appendix: Weakly-Supervised Action Segmentation and Unseen Error Detection in Anomalous Instructional Videos

Reza Ghoddoosian Isht Dwivedi Nakul Agarwal Behzad Darisuh
Honda Research Institute, USA
{reza_ghoddoosian, idwivedi, nakul_agarwal, bdariush}@honda-ri.com

1. Overview

In this appendix, we provide a glossary of the symbols used in the paper as a reference. Then, we describe details of our 11 defined error classes and clarify the privacy concerns in our subject recruitment for the ATA dataset. We conclude by a discussion on the real-time performance of our proposed unconstrained Viterbi algorithm.

2. Table of Symbols

Table 1 provides definitions of symbols for the readers to refer to.

3. ATA Dataset

3.1. Error Labels

Table 2 and Table 3, respectively, describe our defined error labels and error functions.

3.2. Subject Recruitment

We recruited 32 volunteer participants to collect the ATA dataset. Each participant took part in the data collection process after signing a consent form, and they agreed to having their data released to the public. However, in order to protect the privacy of the participants, faces are blurred and videos are muted.

4. Real-time Performance Details

Table 3 of the main paper shows the average fps given different beam size and threshold values. We report our fps on the ATA dataset based on inference every 30 frames on a single NVIDIA Quadro RTX6000. Note that a precision of 30 frames (1 second) to segment actions is a reasonable choice as no action duration is less than 1 second. Theoretically, with an inference rate of 1 inference per frame and $\tau = 0.7$, the fps values in Table 3 should be divided by 30. Even in this case with a smaller beam size ($B < 150$), real-time performance is achieved with hardly losing any accuracy.

Table 1. Definitions of symbols used in the paper.

Symbol	Definition
$\mathbb{A}$	The set of all actions in the dataset
$\|\mathbb{A}\|$	Number of actions classes
a_n	Action variable at segment n
$\boldsymbol{a}_1^n$	Sequence of action labels from segment 1 to segment n
$\overline{\boldsymbol{a}}_1^n$	Inferred sequence of action labels
$\vec{\boldsymbol{A}}_t$	Anticipated action probability vector of time t
$\vec{\boldsymbol{A}}_i^j$	Anticipation vector from time i to time j
B	Beam size for pruning
$\mathbb{E}_\tau(\boldsymbol{a}_1^{\acute{n}-1}, t)$	Set of possible action labels for segment $\acute{n}$
$\mathcal{E}$	Set of error labels
$\mathcal{F}_e$	Error function for error e (Table 3)
$\mathbf{f}^0$	action frequency vector for $\tau = 0$
$\mathbf{f}^\tau$	action frequency vector for $\tau > 0$
I	Error Instructions (Table 2)
LCS	Longest Common Subsequence
$\mathcal{L}$	Overall loss
$\mathcal{L}_{\text{CDFL}}$	CDFL loss
$\boldsymbol{l}_1^n$	Sequence of action durations from segment 1 to segment n
$\mathcal{N}^+(\boldsymbol{a}_i^j)$	Set of all actions that succeed sequence $\boldsymbol{a}_i^j$
n_t	Segment number at time t
n_e	Number of instances for error e
$P_t[l_n, \boldsymbol{a}_1^n]$	Probability of the most likely alignment of $\boldsymbol{a}_1^n$ until time t where a_n is incomplete and has a duration of l_n
$p_{\text{mode}}()$	Probability in the online or offline **mode**
$Q[]$	Temporary variable in Algorithm 1
$\vec{\boldsymbol{R}}_t$	Current action probability vector of time t
$\mathcal{S}_t^B(a)$	The top B most likely segmentation results ending with action a at time t
$\mathfrak{s}()$	Cosine similarity
S^0	Reference offline segmentation when $\tau = 0$
S^τ	Offline segmentation when $\tau > 0$
$\mathcal{T}_s$	The set of all seen transcripts
$\mathcal{T}_u$	The set of all unseen test transcripts
$\mathcal{T}_v$	The set of all validation transcripts
$\mathcal{T}_{\text{tr}}$	The set of all training transcripts
$\mathcal{T}_{\text{tr}}^{\subset}$	A subset of the training transcripts
$top^B\{A\}$	Function to retrieve the top B values in a set A
$t_{\acute{n}}$	Transition time into segment $\acute{n}$
$\mathbb{V}$	The set of all videos in the dataset
ω	Future anticipation range
$\mathbf{x}_1^t$	Sequence of video features from frame 1 to frame t
θ_a	Parameters for the anticipation network
θ_r	Parameters for the recognition network
λ_a	Coefficient of the anticipation term in the loss function

Table 2. Description of 11 Error classes that are only present in the validation and test sets.

Error	Definition
idle time	Background (bg) segments in the middle of assembly. All training videos include a continuous assembly process without any background segments interrupting the assembly.
dropped item w/o pickup	Not picking up an item that is dropped. In the training set all dropped items are retrieved.
missing part	Not using all three required parts to assemble the airplane in task 1.
missing ring	Not using any rings while assembling the airplane (task 1) or the record player (task 3).
missing leg	Assembling a table (task 2) without the use of all four legs (blocks).
unfastened leg	Not fastening a leg (block) while assembling the table in task 2 irrespective of the number of legs used.
unfastened screw	Not fastening an inserted screw with a nut in the assembly of the record player (task 3) irrespective of the number of screws used.
missing screw	Assembling the record player (task 3) with less than the required two screws.
extra screw	Assembling the record player (task 3) with more than the required two screws.
extra ring	Using more than one ring in the final assembly of the record player in task 3.
no balancing	Not balancing the part on the plate while assembling the record player in task 3.

Table 3. Our defined error functions.

Function	Definition
$\mathcal{F}_{\textbf{idle time}}$	$\mathrm{f}_{\text{in-between background}}$
$\mathcal{F}_{\textbf{dropping item without pickiup}}$	$max(\mathrm{f}_{\text{drop}} - \mathrm{f}_{\text{pickup}}, 0)$
$\mathcal{F}_{\textbf{Missing part}}$	$max(3 - \mathrm{f}_{\text{take part}}, 0)$
$\mathcal{F}_{\textbf{Missing ring}}$	$max(1 - \mathrm{f}_{\text{take ring}}, 0)$
$\mathcal{F}_{\textbf{Missing leg}}$	$max(4 - \mathrm{f}_{\text{take block}}, 0)$
$\mathcal{F}_{\textbf{Unfastened leg}}$	$max(\mathrm{f}_{\text{take block}} - (\mathrm{f}_{\text{fasten screw}} + \mathrm{f}_{\text{spin block}}), 0)$
$\mathcal{F}_{\textbf{unfastened screw}}$	$max(\mathrm{f}_{\text{insert screw}} - \mathrm{f}_{\text{fasten nut}}, 0)$
$\mathcal{F}_{\textbf{Missing screw}}$	$max(2 - \mathrm{f}_{\text{insert screw}}, 0)$
$\mathcal{F}_{\textbf{Extra screw}}$	$max(\mathrm{f}_{\text{insert screw}} - 2, 0)$
$\mathcal{F}_{\textbf{Extra ring}}$	$max(\mathrm{f}_{\text{insert ring}} - 1, 0)$
$\mathcal{F}_{\textbf{no balancing}}$	$max(1 - \mathrm{f}_{\text{balance part}}, 0)$