Sahar Mohammadi
Ferdowsi university of Mashhad

Review and Analysis of State of the Art Approaches for Action Segmentation

Sep 2023

Introduction to the Problem

Action segmentation plays a vital role in various video analysis applications
Dividing a continuous video stream into meaningful action segments
Labelling each frame of a video with an action

Figure 1. Segmentation output example from Breakfast Dataset [14]: P46 webcam02 P46 milk. Colors indicate different actions in chronological order: ϕ, take cup, spoon powder, pour milk, stir milk, ϕ, where ϕ is background shown in while color.

USAGE

surveillance
human-computer interaction
Content-based Video Retrieval
Video Summarization

CHALLENGES

Variability in Actions
Complex Background Scenes
Varying Action Durations
Complex Motion Patterns
Personalized Styles
Lack of Video-level Contextual Understanding

Thesis objectives

Comprehensive review and analysis of action segmentation methodologies, including TW-FINCH, SSTDA and HASR

Dissect the strengths, weaknesses, and contributions of these approaches

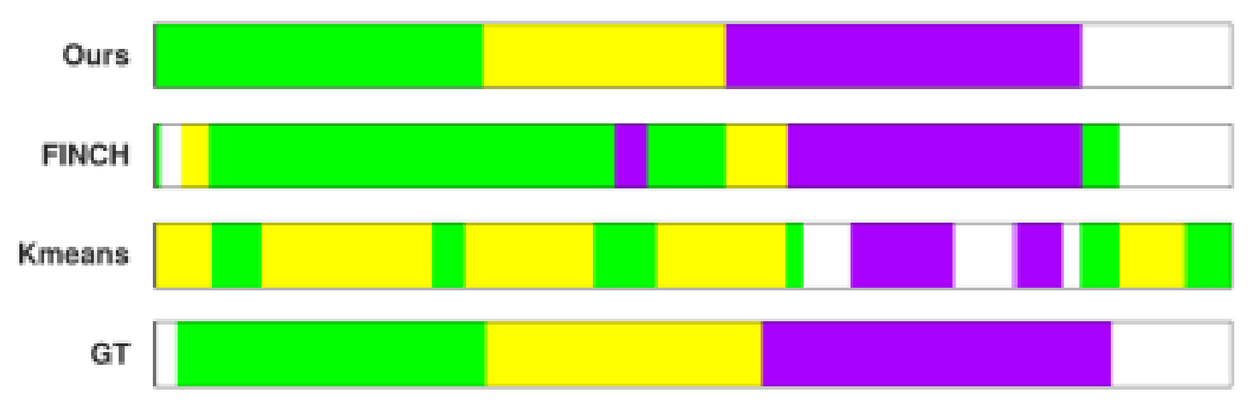
Visualizing the results and conducting error analysis

Overview of TW-FINCH

The Temporally-Weighted Hierarchical Clustering model is to automatically identify action boundaries in unsegmented video sequences without requiring any manually annotated training data

Temporal relationships between frames

More accurate segment lengths



Model architecture

01

Preprocessing

The model begins with preprocessing the input video data to extract relevant features:
spatial and temporal features
optical flow
histograms of oriented gradients (HOG)
deep learning-based representations

02

Temporal Weighting

assigns weights to each frame based on its temporal position within the video sequence

03

Hierarchical Clustering

A hierarchical clustering process based on the weighted frames.
Aims to group frames with similar temporal patterns, forming clusters that correspond to different actions

04

Action Boundary Detection

The points of transition between different clusters are considered as action boundaries, leading to the final action segmentation result

Evaluation

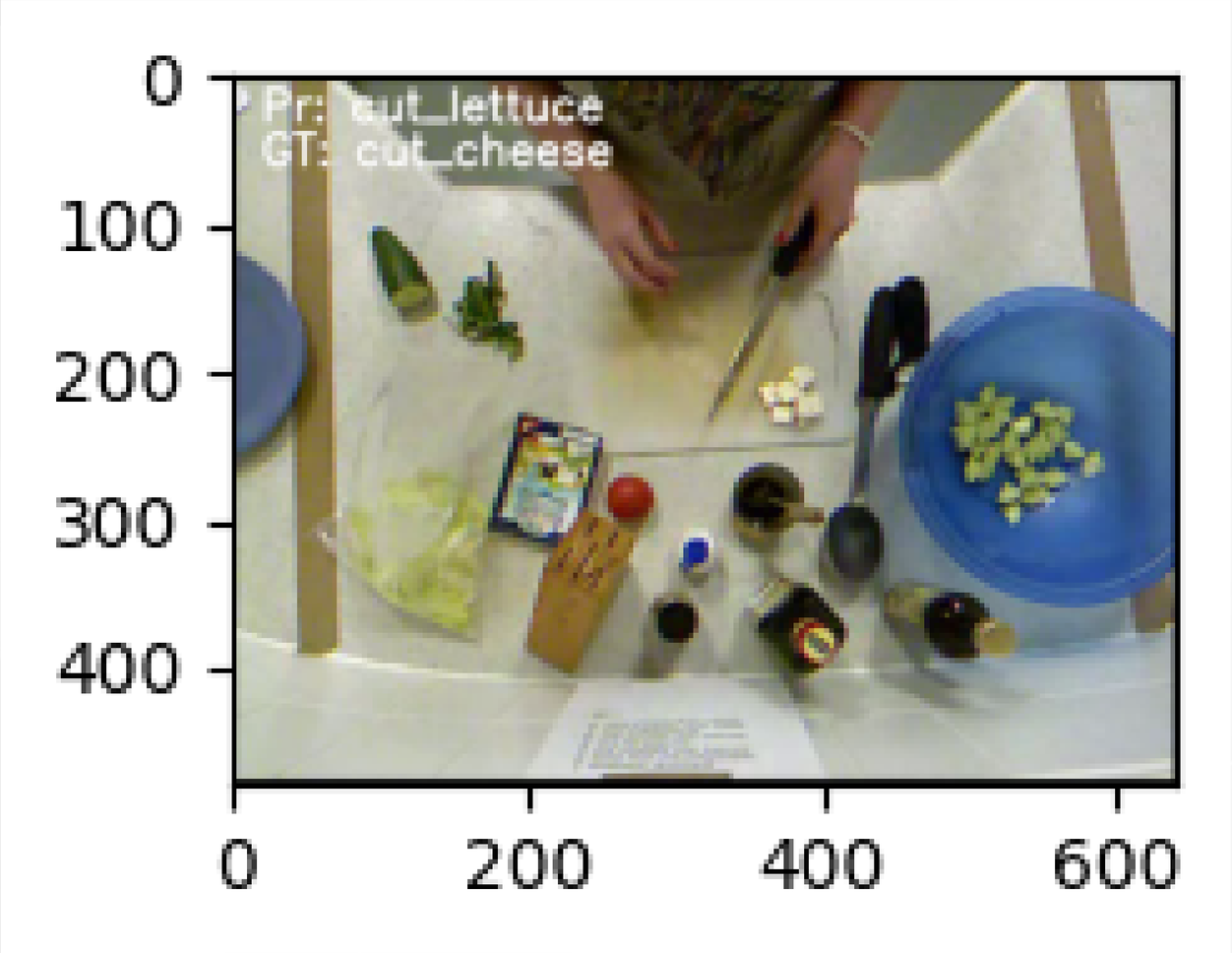
a dataset all three conduct their experiments on is 50salads
an average of 19 actions per video

	50Salads			
Supervision	Method	eval	mid	T
Fully Sup.	ST-CNN [21]	68.0	58.1	✓
	ED-TCN [20]	72.0	64.7	✓
	TricorNet [8]	73.4	67.5	✓
	MS-TCN [10]	80.7	—	✓
	SSTDA [7]	83.8	—	✓
Weakly Sup.	ECTC [13]	—	11.9	✓
	HTK+DTF [15]	—	24.7	✓
	RNN-FC [26]	—	45.5	✓
	NN-Vit. [28]	—	49.4	✓
	CDFL [22]	—	54.7	✓
Unsup. Baselines	Equal Split	47.4	33.1	✗
	Kmeans	34.4	29.4	✗
	FINCH	39.6	33.7	✗
Unsup.	LSTM+AL [1]	60.6	—	✓
	TW-FINCH	71.1	66.5	✗

Weaknesses

as well as the high MoF(66.8%), some weaknesses still exits. finding them would help us to improve the model in those aspects.

test on FS video 21.1, accuracy: 61%



test on FS video 11.2, accuracy: 62%

70%

LONG/SHORT ACTIONS

when the action is too long it clusters the beginning half or the ending quarter of it and assigns another action to it which is so randomly. It might be a really short action which happened some 10 actions before or later. And when an anction is too short it almost make the previous action longer.

83%

LOGICAL ORDER

there is some logical orders between actions whish is not cosidered in the model.
e.g. 'place' should be only after 'cut'
but not immediately but it should exist.
or when there is 'mix dressing', 'add dressing' is expected.
'action start' and 'action end' must be the first and last.

90%

LATENCY/RUSH

In some parts the order of actions that are recognized are correct but there is rush or latency. So recognition is kind of good but the clustering and finding the exact duration of each action and clarifying boundaries is weak.

Overview of SSTDA

Joint Self-Supervised Temporal Domain Adaptation approach combines self-supervised learning and temporal domain adaptation techniques.

Aims to improve action segmentation performance in the presence of domain shifts.

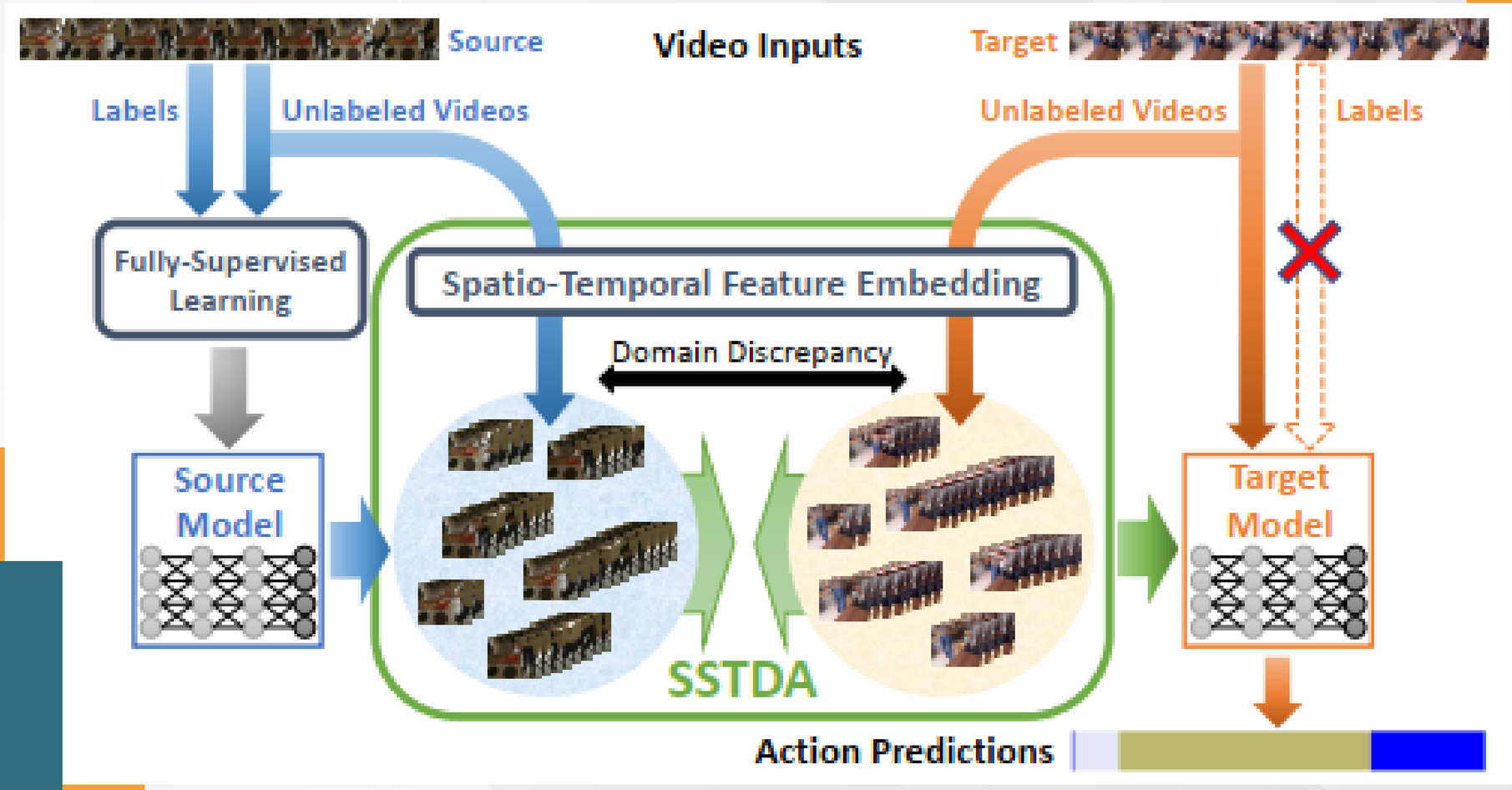
Leverage both labeled source domain data and unlabeled target domain data to enhance the segmentation results

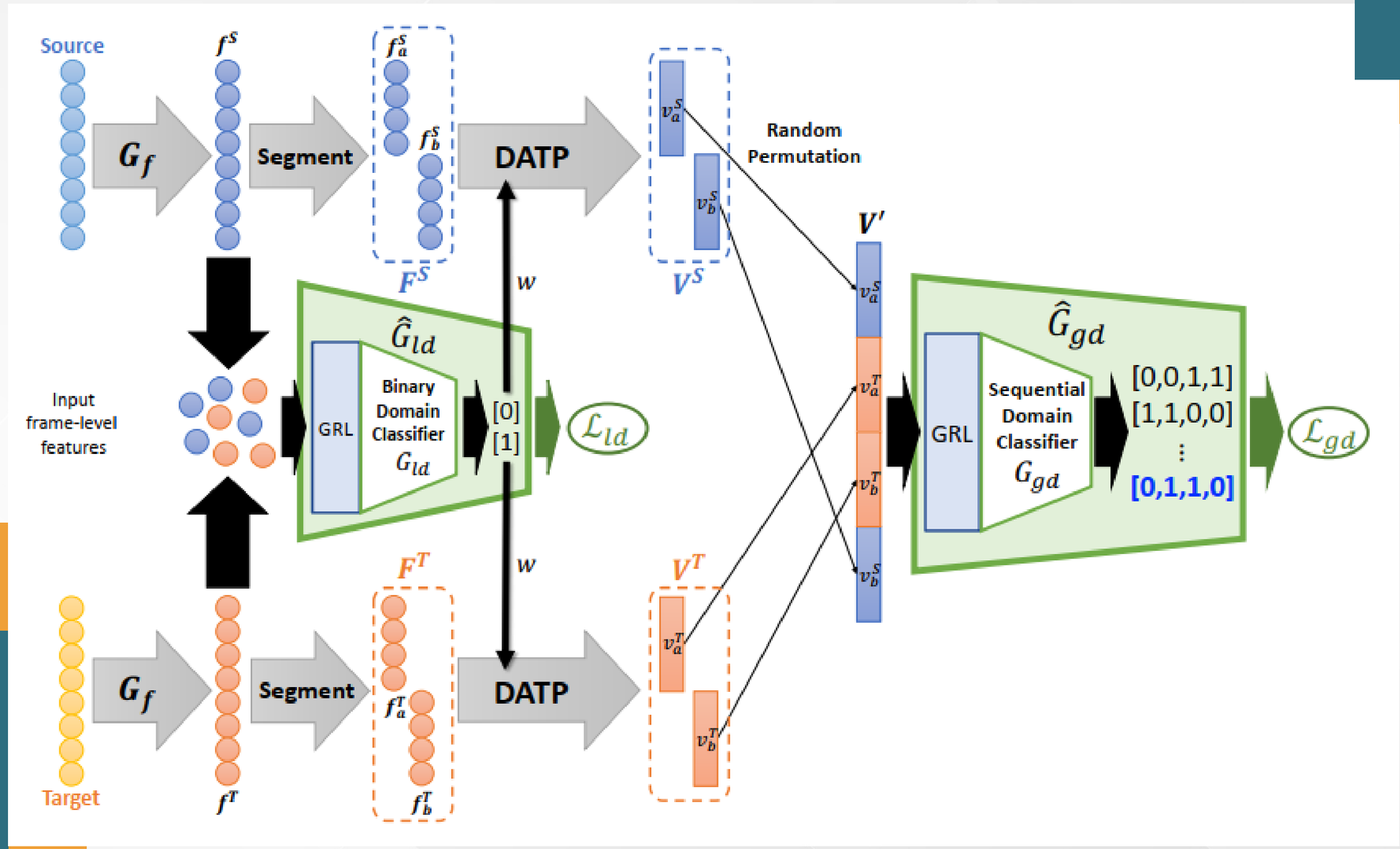
Two sub-tasks are defined to allow the system to recognize this mismatch

Binary Domain classifier

Sequential Domain classifier

Domain Adaptation




Source
f^S
G_f
Segment
f_a^S
f_b^S
DATP
v_a^S
v_b^S
Random Permutation
V'
F^S
V^S
w
$\hat{G}_{ld}$
GRL
Binary Domain Classifier G_{ld}
[0]
[1]
$\mathcal{L}_{ld}$
Input frame-level features
$\hat{G}_{gd}$
GRL
Sequential Domain Classifier G_{gd}
[0,0,1,1]
[1,1,0,0]
⋮
[0,1,1,0]
$\mathcal{L}_{gd}$
F^T
V^T
w
v_a^T
v_b^T
G_f
Segment
DATP
f_a^T
f_b^T
Target
f^T

Evaluation

Evaluation metrics are F1-score, accuracy, precision, and recall.

“SSTDA” refers to the full model while “Local SSTDA” only contains binary domain prediction. Global SSTDA requires outputs from local SSTDA, so it is not evaluated alone

GTEA	F1@{10, 25, 50}			Edit	Acc
Source only (MS-TCN)†	86.5	83.6	71.9	81.3	76.5
Local SSTDA	89.6	87.9	74.4	84.5	**80.1**
SSTDA‡	**90.0**	**89.1**	**78.0**	**86.2**	79.8
50Salads	F1@{10, 25, 50}			Edit	Acc
Source only (MS-TCN)†	75.4	73.4	65.2	68.9	82.1
Local SSTDA	79.2	77.8	70.3	72.0	82.8
SSTDA‡	**83.0**	**81.5**	**73.8**	**75.8**	**83.2**
Breakfast	F1@{10, 25, 50}			Edit	Acc
Source only (MS-TCN)†	65.3	59.6	47.2	65.7	64.7
Local SSTDA	72.8	67.8	55.1	71.7	**70.3**
SSTDA‡	**75.0**	**69.1**	**55.2**	**73.7**	70.2

Weaknesses

as well as the high Acc(83.2%), which is a lot higher than the previous method provided by Sarfraz(66.8%), some weaknesses still exits. finding them would help us to improve the model in those aspects.

Computational Time and Resource Intensive

It took almost 8h to run.
The model's training and prediction processes require a significant amount of computational resources and time, which can be a limitation, especially in time-sensitive or resource-constrained environments.

76%

LATENCY/RUSH

In some parts the order of actions that are recognized are correct but there is rush or latency. So again, clarifying boundaries is the problem.
76% was so visible that in some of them with high discrepancy.
3% was so little maybe only a few frames, that the acc of that vidoe wa above 90%.

23%

LONG/SHORT ACTIONS

there is a diff attitude here, some actions that are too small or too long, both are predicted longer than what they really are and some of the times smaller.
But it is so accurate that most of the time it recognizes the small actions at the correct place.

13%

LOGICAL ORDER

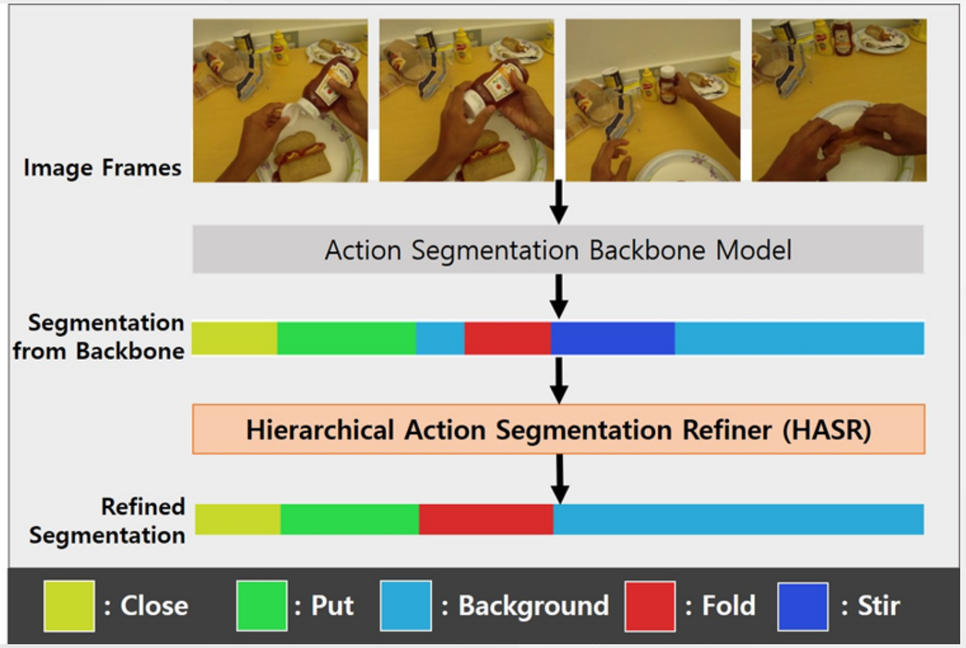
there is some logical orders between actions whish is not cosidered in the model.
e.g. 'place' should be only after 'cut'
but not immediately but it should exist.
or when there is 'mix dressing', 'add dressing' is expected.
'action start' and 'action end' must be the first and last.

test on FS video 6.2, accuracy: 84%

Overview of HASR

Hierarchical Action Segmentation Refiner model extracts the hierarchical video representations to understand the overall context, refine the results from action segmentation backbone models

a supervised way, by referring to the segmentation results from the pretrained backbone models as well as the ground truth to improve the segmentation results from unseen backbone models



Evaluation

Evaluation metrics are F1-score, accuracy

Shows that it has a potential to be extensively used as an effective tool for boosting up the performance of any action segmentation models

50Salads					
Method	F1@{10, 25, 50}			Edit	Acc
GRU	62.4	60.0	52.2	55.6	80.5
GRU + HASR	78.1	76.0	67.7	72.2	80.9
Gain	15.7	16.0	15.5	16.5	0.4
MS-TCN [5]	76.3	74.0	64.5	67.9	80.7
MS-TCN (our impl.)	77.2	74.7	64.8	70.4	80.3
MS-TCN + HASR	83.4	81.8	71.9	77.4	81.7
Gain	6.2	7.1	7.1	7.0	1.4
SSTDA [3]	83.0	81.5	73.8	75.8	83.2
SSTDA (our impl.)	80.6	78.7	70.8	74.9	82.5
SSTDA + HASR	83.5	82.1	74.1	77.3	82.7
Gain	2.9	3.4	3.3	2.4	0.2
ASRF [10]	84.9	83.5	77.3	79.3	84.5
ASRF (our impl.)	85.1	83.3	77.7	79.9	83.7
ASRF + HASR	86.6	85.7	78.5	81.0	83.9
Gain	1.5	2.4	0.9	1.2	0.2

Conclusion

Review and analyze best approaches, with a particular emphasis on the influential articles TW-FINCH, SSTDA and HASR

	TW-FINCH	SSTDA
Long/Short Actions	70%	23%
Logical Order	83%	13%
Latency/Rush	90%	76%

REFERENCES

1.M. Saquib Sarfraz, Naila Murray, Vivek Sharma, Ali Diba, Luc Van Gool, Rainer Stiefelhagen Karlsruhe Institute of Technology, Facebook AI Research, MIT, Harvard Medical School, KU Leuven, ETH Zurich Daimler TSS which is called Temporally-Weighted Hierarchical Clustering for Unsupervised Action Segmentation

2.Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, Zsolt Kira, Georgia Institute of Technology, Baidu USA which is Action Segmentation with Joint Self-Supervised Temporal Domain Adaptation

3.Hyemin Ahn, and Dongheui Lee, German Aerospace Center (DLR), Technical University of Munich which is Refining Action Segmentation with Hierarchical Video Representations

@sahar_mhdd

THANK YOU