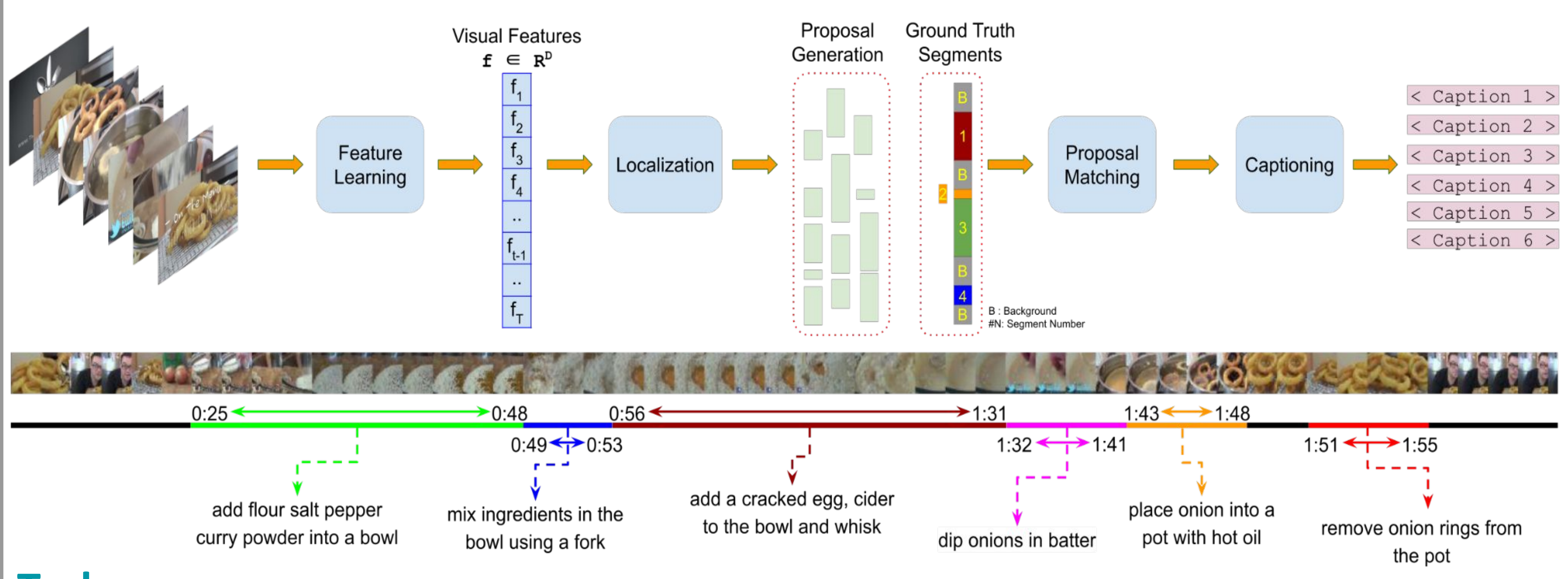
A Closer Look at Temporal Ordering in the Segmentation of Instructional Videos

THE UNIVERSITY of EDINBURGH informatics

Anil Batra Shreyank N Gowda Frank Keller Laura Sevilla-Lara

UKRI CENTRE FOR DOCTORAL TRAINING NLP

Overview



Task

- Learning temporal representation of instructional videos.
- Temporally identify the key steps and generate textual summary.
- Temporal segmentation is critical for generating correct textual summary.
- Closely related to Dense Video Captioning Task[3].
- Datasets: YouCook2 [1] and Tasty [2]

Challenges

Proxy Evaluation Metrics

- ✓ Do not include 1-to-1 mapping between Ground Truth and Predicted segment
- ✓ Recursive search to find the best match with highest overlap
- ✓ Proposal detection metrics are used and overestimate

1. Proposal Detection Metric [3, 5]

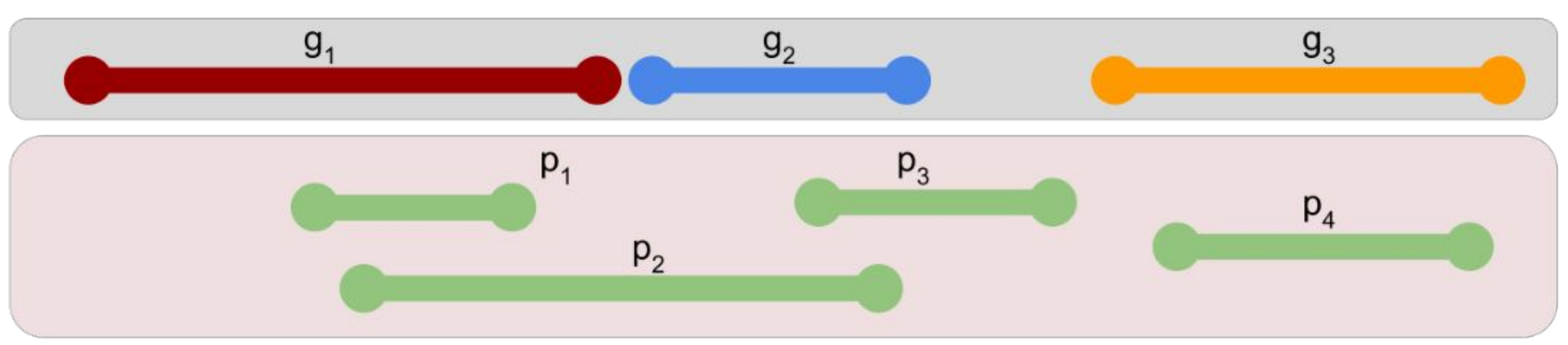
$$P_{g,\tau} = \{p \in \mathcal{P} | IoU(g,p) > \tau\} \qquad G_{p,\tau} = \{g \in \mathcal{G} | IoU(g,p) > \tau\}$$

$$Precision = \frac{|\bigcup_{g\in\mathcal{G}} P_{g,\tau}|}{|\mathcal{P}|} \qquad Recall = \frac{|\bigcup_{p\in\mathcal{P}} G_{p,\tau}|}{|\mathcal{G}|}$$

2. Recursive Overlap Metric [1]

$$mIoU = \frac{\sum_{g\in\mathcal{G}} max(IoU(g,p)|p\in\mathcal{P})}{|\mathcal{G}|}$$

Example



	p_1	p_2	p_3	p_4
g_1	0.26	0.32	0	0
g_2	0.13	0.49	0.21	0
g_3	0	0	0	0.75

Proposal Metric: $Precision_{\tau=0.3} = \frac{|\{p_2, p_4\}|}{4} = 0.5 \qquad Recall_{\tau=0.3} = \frac{|\{g_1, g_2, g_3\}|}{3} = 1.0$

Overlap Metric: $mIoU = \frac{0.32+0.49+0.75}{3} = 0.52$

SODA-D (F1) Score

1. Optimal Matching using Dynamic Programming

- ✓ Utilize matching as Combinatorial Optimization to maximize the average IoU
- ✓ Incorporate temporal order while matching

Dynamic Programming Matching

$$C_{i,j} = IoU(g_i, p_j) \qquad S[i][j] = max \begin{cases} S[i-1][j] \\ S[i-1][j-1] + C_{i,j} \\ S[i][j-1] \end{cases}$$

SODA-D (F1) Score

$$Precision = \frac{\sum_{g\in\mathcal{G}} IoU(g, a(p))}{|\mathcal{P}|} \qquad Recall = \frac{\sum_{g\in\mathcal{G}} IoU(g, a(p))}{|\mathcal{G}|}$$

Example

0.26	0.32	0.32	0.32
0.26	0.75	0.75	0.75
0.26	0.75	0.75	1.5

Dynamic Table

D	D	L	L
T	D	L	L
T	T	T	D

Traceback Table

Sequential Matching Optimzation

1. Hungarian Matching [5]

- ✓ Generate proposal based on Hungarian Matcher - NonDifferentiable
- ✓ Temporal structure of segments is **not** incorporated

2. SODA Matching (Ours)

- ✓ Differentiable matching algorithm.
- ✓ Plug into training pipeline to improve temporal segmentation performance.

Differentiable SODA Matching

$$C_{i,j} = -IoU(g_i, p_j) \qquad S[i][j] = min^{\gamma} \begin{cases} S[i-1][j] \\ S[i-1][j-1] + C_{i,j} \\ S[i][j-1] \end{cases}$$

Example

	p_1 = [1,9]	p_2 = [1,4]	p_3 = [4,8]
g_1 = [2,5]	0.38	0.5	0.17
g_2 = [7,9]	0.25	0	0.2

Hungarian: [(g_1, p_2), (g_2, p_1)]

SODA: [(g_1, p_2), (g_2, p_3)]

Score: 27.99

Example - 1

	p_1 = [1,4]	p_2 = [1,9]	p_3 = [4,8]
g_1 = [2,5]	0.5	0.38	0.17
g_2 = [7,9]	0	0.25	0.2

Hungarian: [(g_1, p_1), (g_2, p_2)]

SODA: [(g_1, p_1), (g_2, p_2)]

Score: 30.0

Example - 2

Quantitative Results

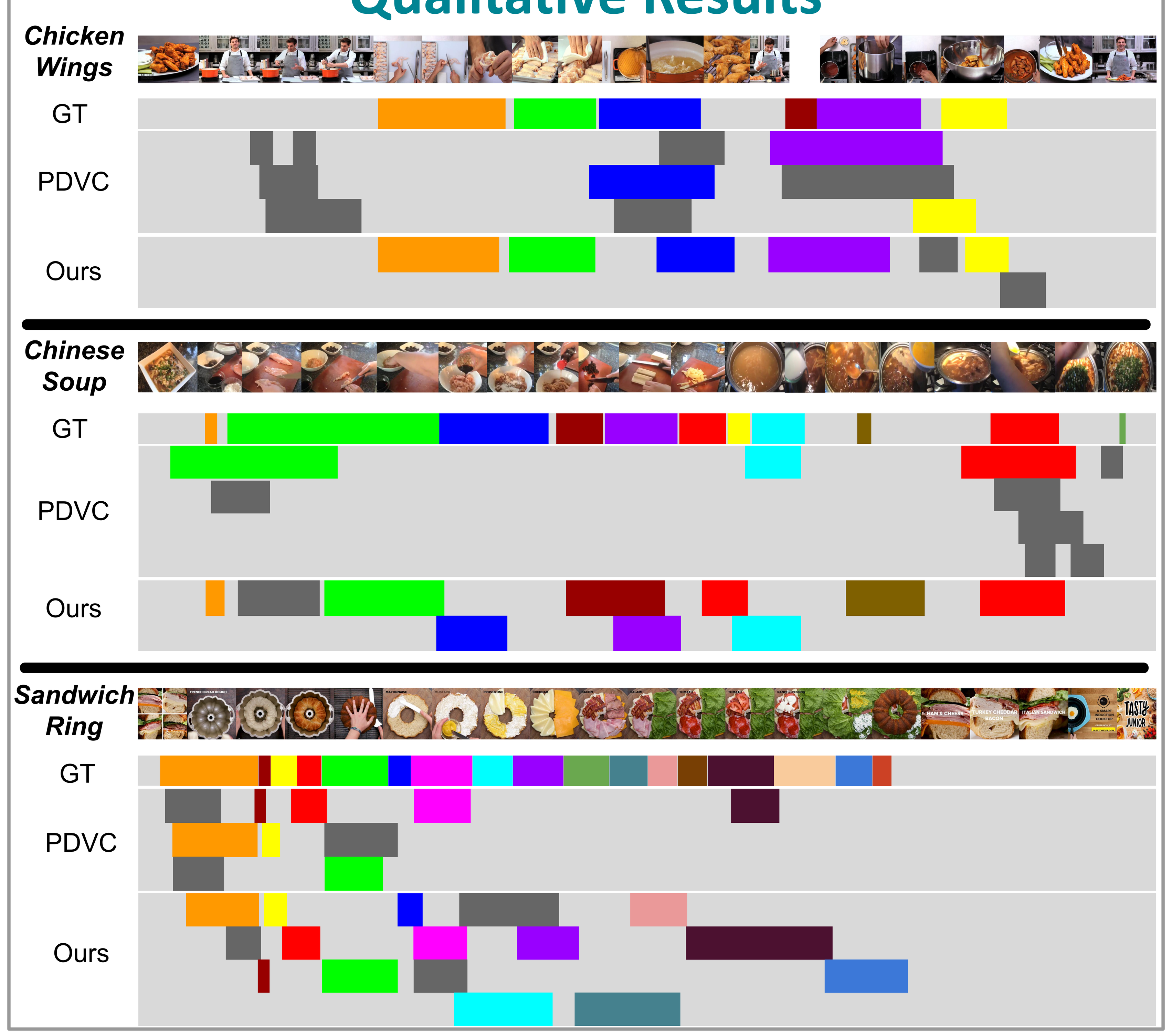
Existing and Proposed Evaluation Metrics Comparison

Method	mean-IoU	mean-Jaccard	Precision	Recall	SODA-D (F1)
Uniform (Avg # Segment)	35.18	40.79	29.01	31.67	29.46
Uniform (Avg # Duration)	43.43	61.87	22.5	43.77	28.53
Uniform (GT)	34.18	38.26	30.47	30.47	30.47
ProcNet [1]	32.3	46.32	26.72	30.98	27.89
PDVC [5]	33.54	42.61	27.44	31.39	28.35
Ours	**41.38**	**52.63**	**36.01**	**39.67**	**36.8**

Procedure Segmentation and Summarization Comparison

Method	Video Features	Matcher	YouCook2		Tasty	
			SODA-D	SODA-C [4]	SODA-D	SODA-C [4]
Uniform (Avg # Segment)	-	-	29.46	-	36.82	-
Uniform (Avg # Duration)	-	-	28.53	-	39.88	-
Uniform (GT)	-	-	30.47	-	43.35	-
ProcNet [1]	RGB + Flow	-	27.89 ± 1.25	-	34.12 ± 1.01	-
PDVC [5]	R3D	Hungarian	28.35 ± 0.27	4.11 ± 0.05	42.44 ± 0.66	6.58 ± 0.16
Ours	S3D	Hungarian	33.11 ± 0.28	6.13 ± 0.08	46.57 ± 0.65	9.17 ± 0.19
Ours	S3D	SODA	36.32 ± 2.19	6.39 ± 0.51	**50.84 ± 0.41**	**9.71 ± 0.18**
Ours	S3D	SoftSODA	**36.80 ± 1.90**	**6.54 ± 0.44**	50.37 ± 0.63	9.63 ± 0.21

Qualitative Results



References:

1. Luowei Zhou *et. al.* Towards Automatic Learning of Procedures from Web Instructional Videos, AAAI 2018
2. Fadime Sener *et. al.* Sener, Fadime Zero-Shot Anticipation for Instructional Activities, ICCV 2019
3. Ranjay Krishna *et. al.* Dense-Captioning Events in Videos, ICCV 2017
4. Soichiro Fujita *et. al.* SODA:Story Oriented Dense Video Captioning Evaluation Framework, ECCV 2020
5. Teng Wang *et. al.* PDVC: End-to-End Dense Video Captioning with Parallel Decoding, ICCV 2021

Acknowledgement:

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

UK Research and Innovation

BMVC 2022