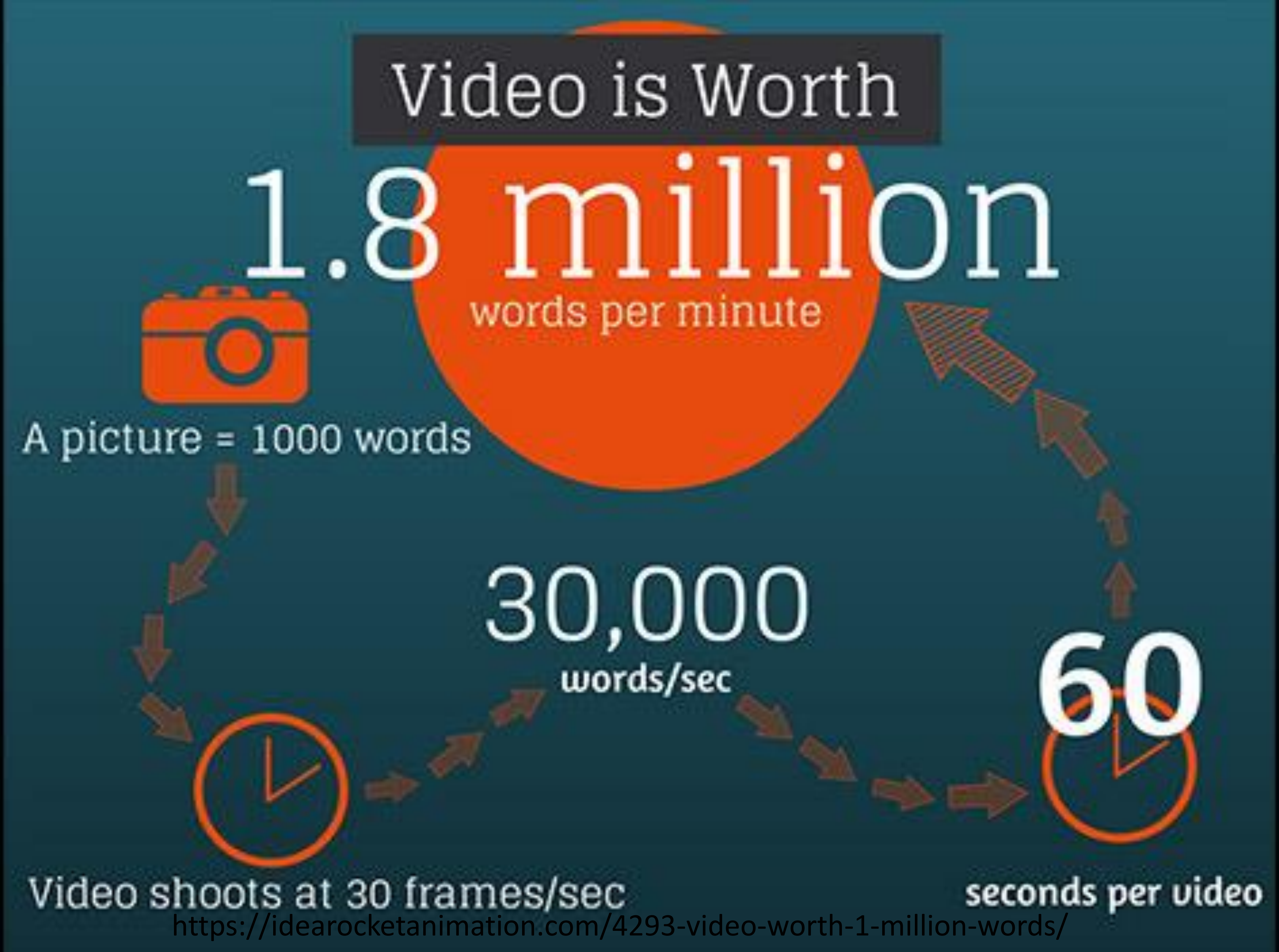
Realistic Video Summarization through VISIOCITY: A New Benchmark and Evaluation Framework

Vishal Kaushal[1], Suraj Kothawade[2], Rishabh Iyer[2], Ganesh Ramakrishnan[1]

[1]Indian Institute of Technology Bombay, [2]University of Texas at Dallas



https://idearocketanimation.com/4293-video-worth-1-million-words/

Flip Side: Lot of Redundancy

Capture Now, Process Later Mentality

Need for Automatic Video Summarization

Motivation 1: Dataset

Name	# Videos	Duration of Videos	Total Duration	# summ	# cat
SumMe [8]	25	Avg: 2 min, 9	1 Hour, 10 min	15-18	-
TVSum [29]	50	Avg: 4min, 11 sec	3.5 Hours	20	10
MED Summaries [26]	260	Dur: 1-5 mins, Avg: 2.5min	9 Hours	2-4	15
UT Egocentric [16]	4	Avg: 254 mins	16 Hours	-	1
Youtube 1 [2]	50	Dur: 1-10 min, Avg: 1 min, 39 sec	1.38 Hours	5	-
Youtube 2 [2]	50	Dur: 1-4 min, Avg: 2min, 54sec	2.41 Hours	5	-
Tour20 [24]	140	Avg: 3 min	7 Hours	-	-
TV Episodes [35]	4	Avg: 45 min	3 Hours	-	1
LOL [5]	218	Dur: 30 to 50 min	-	-	1
VISIOCITY (OURS)	**67**	**Dur: 14-121 mins, Avg: 55 mins**	**71 hours**	**160**	**5**

- Either very short videos or very few long videos or long videos of only a particular type
- We release VISIOCITY (VIdeo SummarizatIOn based on Continuity, Intent and diversiTY)
- Rich annotations to support different flavors of summarization and other tasks as well

Motivation 2: No Single Right Answer

Context Dependent

- Depends on purpose for which summary is required

Subjective

- Preferences of two persons may not match

Depends on higher-level semantics of the video

- Visually same, semantically different vs semantically same, visually different

Challenge: No Single Right Answer

- Difficult to get many reference summaries of different lengths of long videos
- A recipe for creating reference summaries of desired length having different characteristics using annotations (concepts present in each shot)
- Learning from a "combined" reference summary vs learning from separate reference summaries
- Evaluating video summaries
    - With respect to reference summaries
        - Avg vs Max
        - F1 could be deceptive!
    - Using the annotations (indirect ground truth)
        - We propose a suite of measures that score a summary on different characteristics
- Learning from a combination of loss functions instead of a single loss function

VISIOCITY

67 long and diverse videos

6 Categories: Friends, Soccer, Surveillance, TechTalk, Birthday, Wedding

Concept annotations for every shot

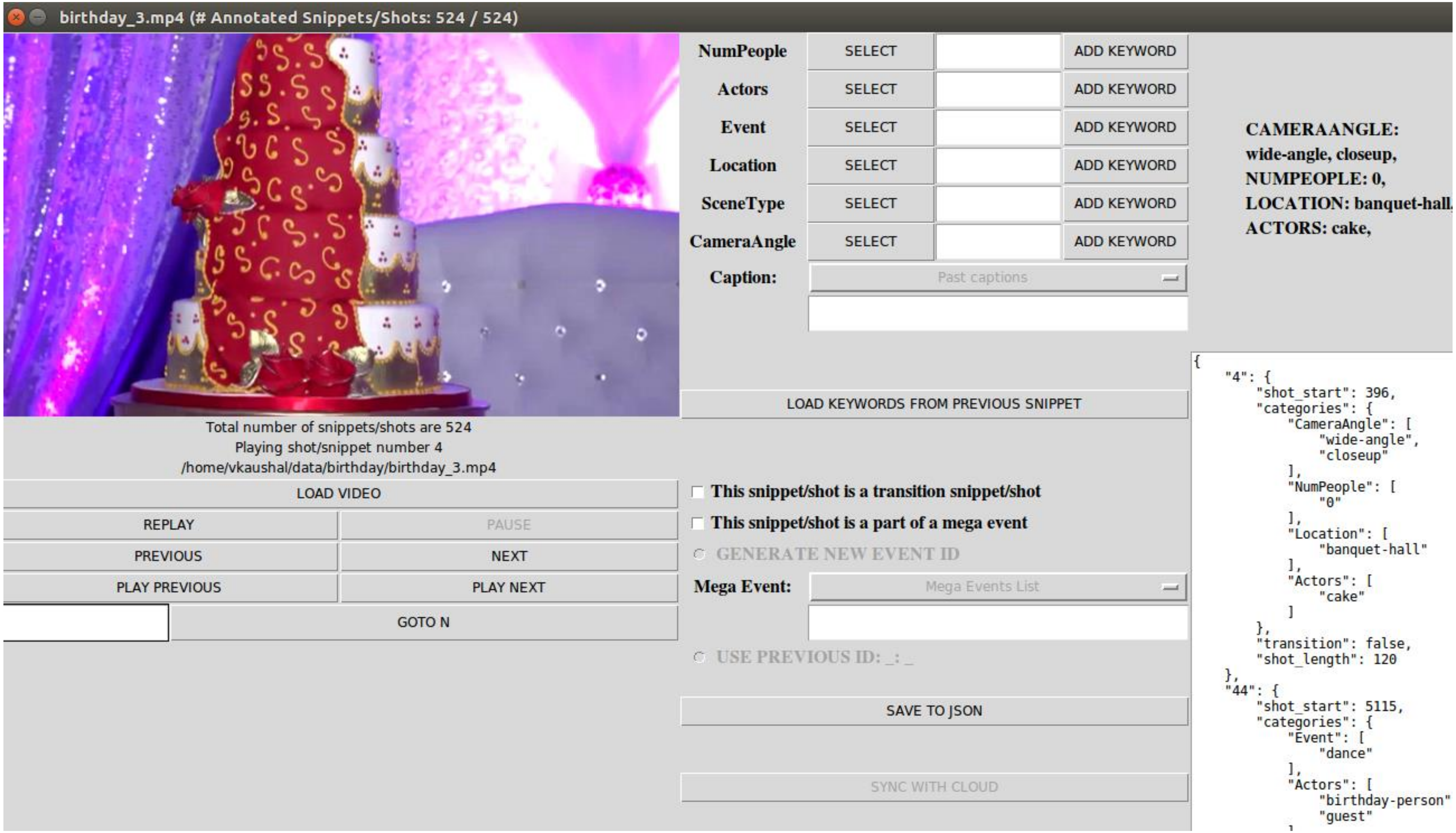
Domain	# Videos	Duration	Total Duration
Sports(Soccer)	12	(37,122,**64**)	12.8
TVShows (Friends)	12	(22,26,**24**)	4.8
Surveillance	12	(22,63,**53**)	10.6
Educational	11	(15,122,**67**)	12.28
Birthday	10	(20,46,**30**)	5
Wedding	10	(40,68,**55**)	9.2
All	67	(26,75,**49**)	54.68




Annotations

- Concepts for each shot (indirect ground-truth)
  - Generator of ground truth summaries
  - More informative
  - Easier and more objective
- 'Mega events'
- Concept annotations are better than scores or ratings
  - Easier and more accurate
  - No chronological bias
  - Semantic content better captured through text both from importance and diveristy perspective
  - Lends well to a wide variety of problems

Annotation Tool



Evaluation

$$Div_{sim}(X) = \max \min_{i,j \in X} d_{ij}$$

Discourages two similar looking snippets in a summary

$$Div(X) = \sum_{i=1}^{|C|} \max_{j \in X \cap C_i} r_j$$

Discourages similar consecutive snippets but doesn't exclude similar snippets separated in time

$$MegaCont(X) = \sum_{i=1}^{E} r^{mega}(M_i)|X \cap M_i|^2$$

(Semantic) Continuity

$$\mathrm{Imp}(X) = \sum_{s \in X \cap A \setminus M} r(s)$$

Importance

Automatic Generation of Reference Summaries

$$score(X, \Lambda) = \lambda_1 \mathrm{MegaCont}(X) + \lambda_2 \mathrm{Imp}(X) + \lambda_3 \mathrm{Div}(X)$$

Different configuration of λ generates different summaries

We compute the Pareto optimal set of configurations and use them to generate the reference summaries

Imp vs Continuity (Eg. Soccer_18)

Imp	Continuity
• 2 (Kick Off) • 83 (Save) • 147 (Goal) • 383 (Save) • 469 (Goal) • 555 (Save) • 578 (Goal) • 688 (Goal) • 804 (Save) • 843 (Save) • 990 (Save) • 1142 (Save)	• 146-147-148 (Goal) • 468-469-470 (Goal) • 576-577-578 (Goal) • 686-687-688 (Goal)

Simple Recipe for a Better Model

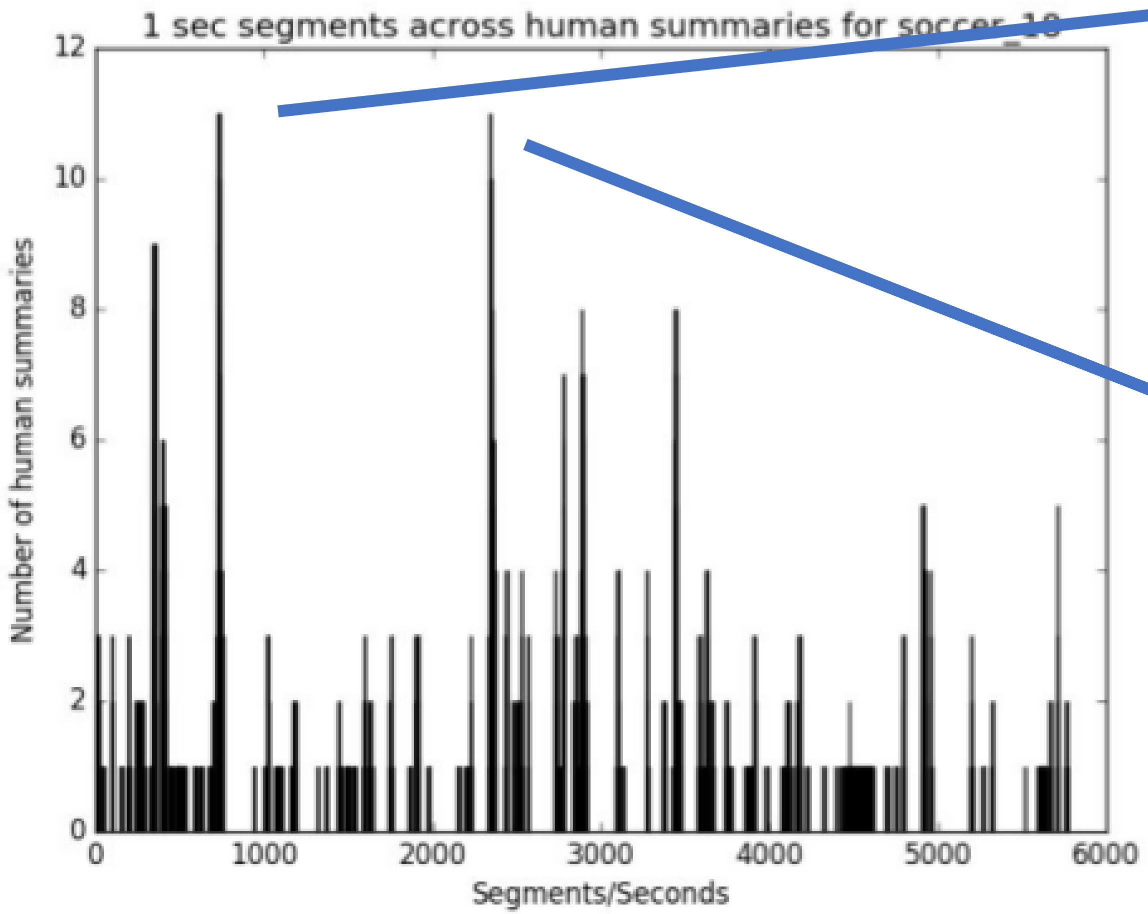
$$y^* = \operatorname*{argmax}_{y \subseteq Y_v, |y| \leq k} o(x_v, y)$$

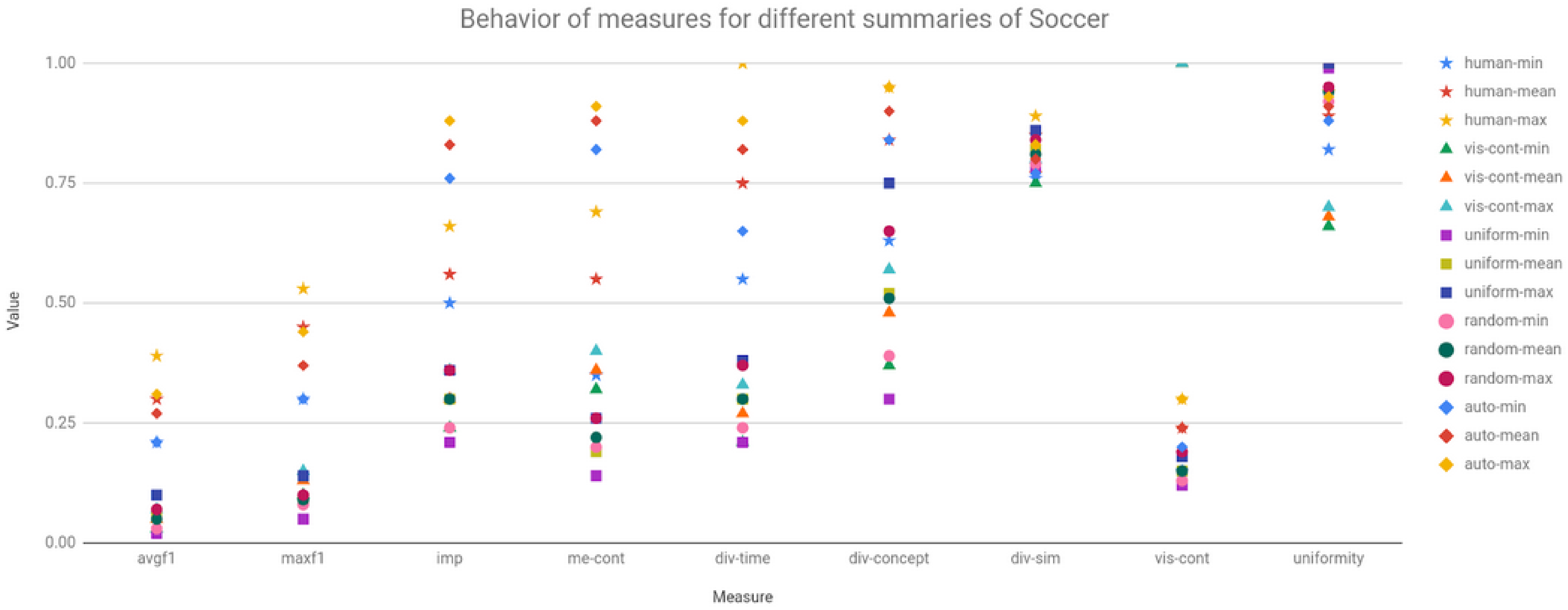
$$o(x_v, y) = w^T f(x_v, y)$$

$$\min_{w \geq 0} \frac{1}{N} \sum_{n=1}^{N} L_n(w)$$

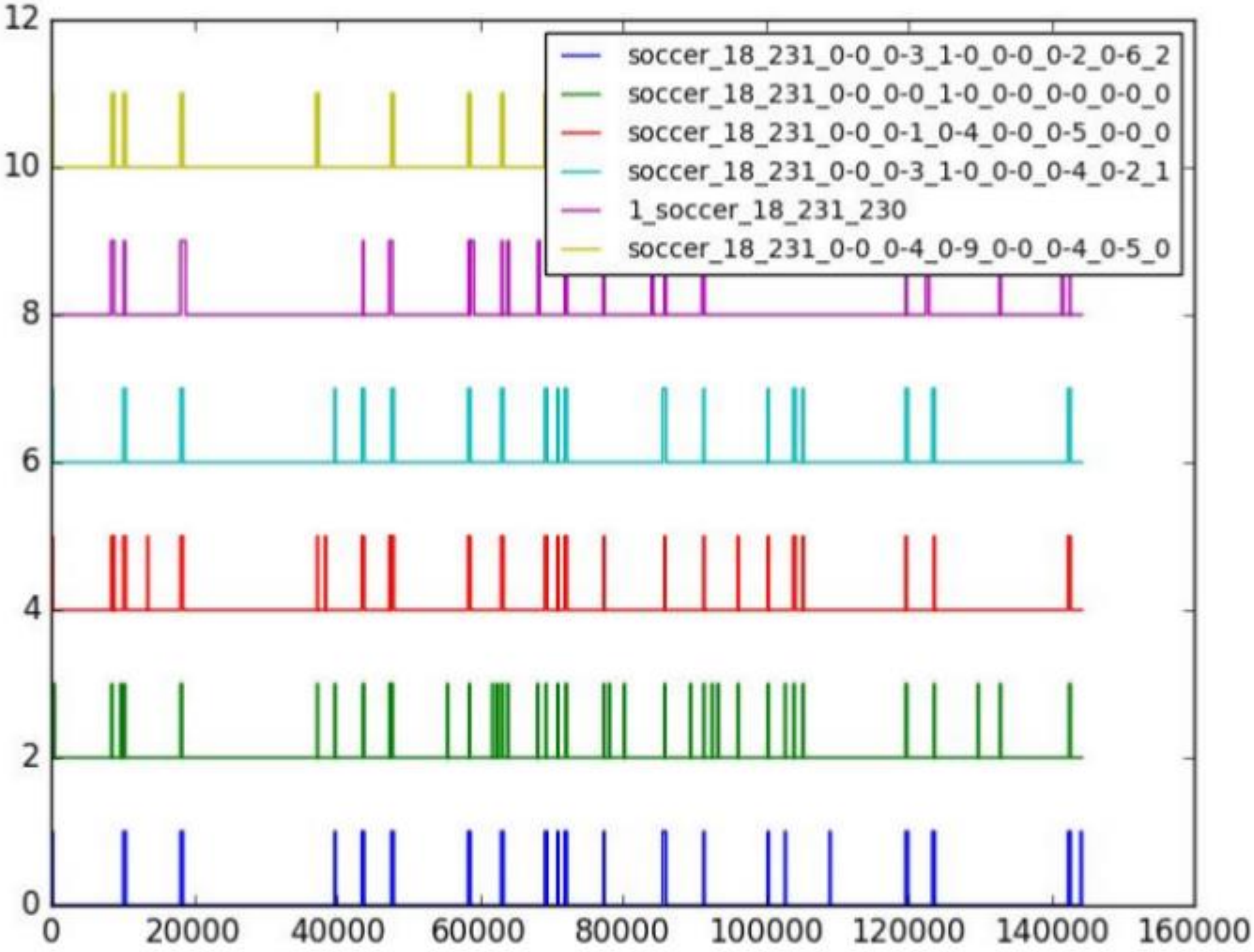
$$L_n(w) = \max_{y \subseteq Y_v^n} (w^T f(x_v^n, y) + l_n(y)) - w^T f(x_v^n, y_{gt}^n)$$

Consistency and InConsistency in Human Summaries





- Automatic summaries are at par with human summaries and are much better than uniform, random or vis-cont baselines
- Vanilla diversity doesn't seem to be a good evaluation measure for Soccer videos



soccer_18_231_0-0_0-3_1-0_0-0_0-2_0-6_2
soccer_18_231_0-0_0-0_1-0_0-0_0-0_0-0_0
soccer_18_231_0-0_0-1_0-4_0-0_0-5_0-0_0
soccer_18_231_0-0_0-3_1-0_0-0_0-4_0-2_1
1_soccer_18_231_230
soccer_18_231_0-0_0-4_0-9_0-0_0-4_0-5_0
0
2
4
6
8
10
12
0
20000
40000
60000
80000
100000
120000
140000
160000

Domain	Technique	AF1	MF1	IMP	MC	DT	DC	DSi
Soccer	Auto	59.3	93.3	83.2	84.3	82.6	85.9	76.2
	DR-DSN	2.8	8.9	23.7	20.3	23.2	30.4	**83.4**
	VASNET	28.4	43.4	**63**	49.3	**62.1**	**67.4**	75.2
	vsLSTM	**31.9**	**48.2**	**62.2**	**60.1**	**62**	**69.5**	76.5
	Ours	**32.6**	**50.3**	**64.2**	**62.6**	**63.4**	**72.2**	78.7
	Random	3.4	9.3	25.7	18.5	25.5	39.2	80.5
Friends	AUTO	66.3	96.9	87.8	84.6	80.3	89.8	83.1
	DR-DSN	4.3	9.4	19.1	6.9	**65.7**	51.5	**98.5**
	VASNET	**17**	**29.6**	**41**	**39.3**	49	**60.6**	86.7
	vsLSTM	15.5	27.2	**40.4**	**39.2**	**64.7**	59	91.1
	Ours	**17.4**	**31.2**	**42.5**	**40.5**	50.2	**64**	90.3
	Random	7.7	17.9	31.5	19.8	34.8	45.2	85.9
Surveillance	Auto	62.4	96.8	81.8	83.2	78.6	98	85.2
	DR-DSN	10	17.7	33.6	20.2	21.8	54.5	**57.2**
	VASNET	**19.4**	**31.4**	**39.5**	**42.6**	**28.4**	**65.4**	37.6
	vsLSTM	10.3	23.6	34.4	18.4	22.8	55.2	**58.4**
	Ours	**20.5**	**32.6**	**41.7**	**44.3**	**29.6**	**68.2**	38.5
	Random	3.9	8	16.6	12	15.3	49.4	69.4
TechTalk	Auto	64.7	91.5	79.8	-	80.5	88.4	94
	DR-DSN	13.5	22.5	49.3	-	24.8	29.9	35.2
	VASNET	**18.2**	**35.7**	52.1	-	**47.3**	**43.3**	**43.2**
	vsLSTM	15.1	32.2	**60.3**	-	38.8	35.3	41.7
	Ours	**18.7**	**37.5**	53.2	-	**50**	**45.8**	**45.5**
	Random	4.5	9.7	38.5	-	28	44	40.6
Birthday	Auto	67.3	97.2	89.7	88.6	68.1	90.8	81.3
	DR-DSN	8.1	14.2	54.7	14.1	**79.4**	63.6	**74.9**
	VASNET	21.6	37.6	50.1	30	36.2	47	48.7
	vsLSTM	**27.3**	**42.1**	**72.1**	**57.2**	59.6	**67.1**	73.6
	Ours	**28**	**44.3**	**74.8**	**60.3**	**62**	**69.5**	**77.6**
	Random	6.9	14.2	51.8	16.9	49.2	54.8	70.3
Wedding	Auto	55.4	94.4	83.9	74.7	67	88	85.7
	DR-DSN	4.2	8.9	40.7	14.4	**76.6**	**62**	**88.4**
	VASNET	4.5	14.4	46.5	22	44	52.7	84.9
	vsLSTM	**9**	**17.3**	**50.2**	**29.5**	50.1	56.9	80.7
	Ours	**9.4**	**17.9**	**52.8**	**30.3**	**51.8**	**58.6**	82.8
	Random	3.5	10	41.1	16.3	40.6	51.6	80

Thank You

Realistic Video Summarization through VISIOCITY: A New Benchmark and Evaluation Framework

Vishal Kaushal[1], Suraj Kothawade[2], Rishabh Iyer[2], Ganesh Ramakrishnan[1]

[1]Indian Institute of Technology Bombay, [2]University of Texas at Dallas