Supplementary Material: Concept Drift Detection for Multivariate Data Streams and Temporal Segmentation of Daylong Egocentric Videos

Pravin Nagar
Indraprastha Institute of Information Technology Delhi

Mansi Khemka*
Columbia University

Chetan Arora
Indian Institute of Technology Delhi

CCS CONCEPTS

• **Computing methodologies** → **Scene understanding**; **Video segmentation**.

KEYWORDS

Temporal segmentation; Concept drift detection; Egocentric video; Multivariate data; Long videos

ACM Reference Format:
Pravin Nagar, Mansi Khemka, and Chetan Arora. 2020. Supplementary Material: Concept Drift Detection for Multivariate Data Streams and Temporal Segmentation of Daylong Egocentric Videos. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3394171.3413713

1 DATASETS

As we discussed in main text, the detail description of video datasets, viz HUJI [5, 6], Disney [2], and UTEgo [3, 4], and the standard photo-stream dataset, viz EDUB-Seg20 [1, 7] is as follows.

HUJI dataset consists of video sequences captured by GoPro camera by three users at a temporal resolution of $30fps$. The dataset comprises several small video clips of less than 30 minutes. For each user, we merged their corresponding small clips into one big video in the specified order. We have evaluated on the videos (of length 4 hours and 2 hours) recorded by only two users using the ground truth boundaries made available by [1]. This is due to the unavailability of the ground truth for the third one. The number of frames in the longest video sequence is 72217.

Disney dataset consists of videos captured at Disney world by 6 individual for three days. Similar to the HUJI dataset, for each user, we have merged several small video clips in the order of the numbering provided by the user. After merging we have a total of 8 video sequences of 4-8 hours for each individual user. We have generated our own ground truth by three different annotators. The number of frames in the longest video sequence is 151695.

*The work was done during the internship with Prof. Chetan Arora.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413713

HUJI dataset (video)					
Videos	F-score	δ	ρ_c	Pred.	GT
Yair	78.94	10^{-2}	0.95	37	38
Chetan	28	10^{-2}	0.95	4	5
Weighted Fscore	**73.01**				

Table 1: F-Measure performance of our method on HUJI video dataset

UTEgo dataset comprises of 4 videos captured by Looxcie wearable camera at a temporal resolution of $15fps$. These videos are 3-5 hours long and captured in an unconstrained setting. We have manually labeled the ground truth for this dataset as well. We will make our annotations public, post acceptance. The number of frames in the longest video sequence is 92287.

EDUB-Seg20 We also demonstrate results on a photo-stream dataset namely EDUB-Seg20. The dataset comprises 18735 images captured through Narrative Clip which captures 2 pictures per minute. The pictures are taken by 7 different users over 20 days. The dataset comprises a variety of scene contexts, viz, attending a conference, traveling, working in the office, etc. EDUB-Seg dataset is released in two versions EDUB-Seg12 comprises 12 videos and EDUB-Seg20 which is the extension of EDUB-Seg12 with 8 new videos. Though our focus is on long videos and not short photo-streams, the evaluation of this dataset allows us to compare our technique against existing temporal segmentation methodologies for egocentric photo-streams.

2 ALGORITHM

As discussed in the main text the detailed pseudo-code of the proposed framework is shown in Algorithm 1.

3 DETAILED F-MEASURE

Table 1 to 7 shows the detailed F-Measure for photostream as well as the video sequence datasets. Tables show the δ, correlation coefficient threshold (ρ_c), and predicted segment for each video sample. As discussed in the main text we have used $\rho_c = 0.95$ for video datasets and $\rho_c = 0.999$ for the photo-stream datasets. Similarly, We set the δ for the corresponding granularity as 10^{-2}, 10^{-4}, and 10^{-6} respectively for video datasets and 10^{-1}, 10^{-3}, and 10^{-7} for photostream datasets.

UTEgo dataset (video)					
Videos	F-score	δ	ρ_c	Pred.	GT
P01	59.79	10^{-4}	0.95	42	55
P02	59.01	10^{-6}	0.95	35	25
P03	56.52	10^{-4}	0.95	25	21
P04	58.33	10^{-4}	0.95	39	32
Avg. Fscore	**58.41**				

Table 2: F-Measure performance of our method on UTEgo video dataset

Disney dataset (video)					
Videos	F-score	δ	ρ_c	Pred.	GT
Alin Day 1	64.36	10^{-6}	0.95	54	32
Alireza Day 1	72.83	10^{-2}	0.95	86	77
Alireza Day 2	66	10^{-2}	0.95	131	72
Alireza Day 3	72.72	10^{-4}	0.95	32	33
Denis Day 1	68.42	10^{-6}	0.95	41	34
Hussein Day 1	65.67	10^{-2}	0.95	67	66
Michael Day 2	71.32	10^{-6}	0.95	77	65
Munehike Day 1	59.67	10^{-6}	0.95	68	57
Avg. Fscore	**67.63**				

Table 3: F-Measure performance of our method on Disney video dataset

HUJI dataset (Phtostream)					
Videos	F-score	δ	ρ_c	Pred.	GT
Yair	59.37	10^{-1}	0.999	27	38
Chetan	60	10^{-1}	0.999	7	5
Weighted Fscore	**59.44**				

Table 4: F-Measure performance of our method on HUJI photostream dataset

UTEgo dataset (Photostream)					
Videos	F-score	δ	ρ_c	Pred.	GT
P01	64.44	10^{-1}	0.999	45	55
P02	60	10^{-3}	0.999	29	25
P03	57.69	10^{-3}	0.999	32	21
P04	60.31	10^{-3}	0.999	35	32
Avg. Fscore	**60.61**				

Table 5: F-Measure performance of our method on UTEgo photostream dataset

Algorithm 1 Proposed Algorithm

Input $F_{i=1}^{N}$**:** Feature vector of video frames
Output $B_{i=1}^{M}$**:** Predicted Boundaries

Initialize the window W
for each frame x_t **do**
  $W \longleftarrow W \cup \{x_t\}$
  Compute average skip factor k in current window by a user defined correlation coefficient ρ_c
  Flag=False
  Possible Boundaries B
  for each n split of W into W_1 . W_2 **do**
    Compute threshold, $\epsilon_{\text{cut}} \geq \sqrt{\frac{4}{m} \log\left(\frac{2n(d+1)}{\delta}\right)}$
    if $\|\mu_1 - \mu_2\|_2 \geq \epsilon_{\text{cut}}$ **then**
      splits = $\|\mu_1 - \mu_2\|_2 - \epsilon_{\text{cut}}$
      $B \longleftarrow B \cup best(splits)$
      Flag = True
    end if
  end for
  if Flag==True **then**
    Drop window W_1 from W along best boundary B
  end if
end for

Disney dataset (Photostream)					
Videos	F-score	δ	ρ_c	Pred.	GT
Alin Day 1	69.56	10^{-3}	0.999	38	32
Alireza Day 1	71.64	10^{-1}	0.999	68	77
Alireza Day 2	62.85	10^{-1}	0.999	72	72
Alireza Day 3	64.28	10^{-3}	0.999	24	33
Denis Day 1	69.84	10^{-3}	0.999	31	34
Hussein Day 1	73.33	10^{-1}	0.999	40	66
Michael Day 2	76.11	10^{-1}	0.999	65	65
Munehike Day 1	63.04	10^{-3}	0.999	44	57
Avg. Fscore	**68.83**				

Table 6: F-Measure performance of our method on Disney photostream dataset

REFERENCES

[1] Mariella Dimiccoli, Marc Bolaños, Estefania Talavera, Maedeh Aghaei, Stavri G Nikolov, and Petia Radeva. 2017. SR-clustering: Semantic regularized clustering for egocentric photo streams segmentation. *Computer Vision and Image Understanding* 155 (2017), 55–69.

[2] Alircza Fathi, Jessica K Hodgins, and James M Rehg. 2012. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition.* 1226–1233.

[3] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *IEEE Conference on Computer Vision and Pattern Recognition.* 1346–1353.

[4] Zheng Lu and Kristen Grauman. 2013. Story-driven summarization for egocentric video. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[5] Yair Poleg, Chetan Arora, and Shmuel Peleg. 2014. Temporal segmentation of egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2537–2544.

[6] Yair Poleg, Ariel Ephrat, Shmuel Peleg, and Chetan Arora. 2016. Compact cnn for indexing egocentric videos. In *2016 IEEE winter conference on applications of computer vision (WACV).* IEEE, 1–9.

EDUB-Seg20 dataset (Photostream)					
Subject-Set	F-score	δ	ρ_c	Pred.	GT
1-1	66.66	10^{-7}	0.999	28	16
1-2	45.71	10^{-7}	0.999	22	12
1-3	70.96	10^{-7}	0.999	17	13
1-4	70	10^{-7}	0.999	40	39
1-5	58.53	10^{-7}	0.999	43	38
2-1	64.615	10^{-7}	0.999	42	22
2-2	56.86	10^{-7}	0.999	67	34
2-3	60.46	10^{-7}	0.999	54	31
2-4	72.72	10^{-7}	0.999	49	38
3-1	71.23	10^{-7}	0.999	37	35
4-1	70.58	10^{-7}	0.999	21	12
5-1	64.70	10^{-7}	0.999	22	11
5-2	61.72	10^{-7}	0.999	36	44
5-3	62.22	10^{-7}	0.999	22	24
6-1	70.12	10^{-7}	0.999	42	34
6-2	69.69	10^{-7}	0.999	35	30
6-3	71.11	10^{-7}	0.999	39	50
6-4	56.75	10^{-7}	0.999	45	28
7-1	54.54	10^{-7}	0.999	70	28
7-2	60	10^{-7}	0.999	26	14
Avg. Fscore	**63.96**				

Table 7: F-Measure performance of our method on EDUB-Seg20 photostream dataset

[7] Estefania Talavera, Mariella Dimiccoli, Marc Bolanos, Maedeh Aghaei, and Petia Radeva. 2015. R-clustering for egocentric video segmentation. In *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 327–336.