Budget-constrained Online Video Summarisation of Egocentric Video Using Control Charts *

Paria Yousefi[0000−0002−5889−7426], Clare E Matthews[0000−0001−6994−3945], and Ludmila I Kuncheva[0000−0002−0415−6964]

School of Computer Science, Bangor University, Bangor, United Kingdom
Email: paria.yousefi@bangor.ac.uk
http://pages.bangor.ac.uk/~mas00a/activities/Leverhulme/project_RPG_2015_188.html

Abstract. Despite the existence of a large number of approaches for generating summaries from egocentric video, online video summarisation has not been fully explored yet. We present an online video summarisation algorithm to generate keyframe summaries during video capture. Event boundaries are identified using control charts and a keyframe is subsequently selected for each event. The number of keyframes is restricted from above which requires a constant review and possible reduction of the cumulatively built summary. The new method was compared against a baseline and a state-of-the-art online video summarisation methods. The evaluation was done on an egocentric video database (Activity of Daily Living (ADL)). Semantic content of the frames in the video was used to evaluate matches with ground truth. The summaries generated by the proposed method outperform those generated by the two competitors.

Keywords: egocentric · summarisation · control chart.

1 Introduction

Wearable camcorders provide consumers with the ability to record their daily activities all day long. Having a voluminous and at the same time largely redundant stream of frames makes browsing the videos a disagreeable task. A fast-speed, user-friendly system would be required to replace the multitude of video images with a concise set of frames containing valuable information [12]. The system must be capable of generating keyframes from forthcoming data streams. Online video summarisation addresses the issue of generating summary on-the-fly from a video stream, in which the algorithm performs under the constraints of low computational processing time and a limited amount of memory. Such an approach could be useful in applications including monitoring the daily routines of elderly people [14], memory support [10, 23, 9], and health behavior monitoring such as sedentary behavior [8] or dietary analysis [15].

* Supported by project RPG-2015-188 funded by The Leverhulme Trust, UK.

Nine online video summarisation methods were described and experimentally compared on non-egocentric video in our previous study [11]. While these methods work reasonably well for non-egocentric videos, it is reasonable to expect that loosely defined event boundaries in egocentric videos will render their performance inadequate. Therefore, this paper proposes a new online summarisation method suitable for egocentric video (Figure 1).

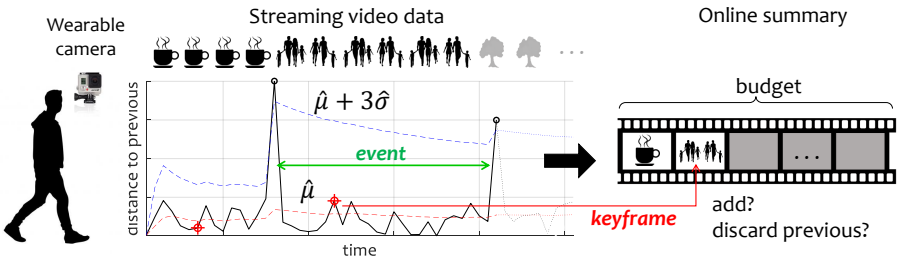

Fig. 1: A sketch of the proposed online video summarisation method for egocentric video. The plot shows the Shewhart chart of the distance between consecutive frames, with the mean μ and the 3σ event-detection boundary, both calculated from the streaming data.

At any moment of the recording video, a valid summary is accessible up to that moment. We required that the new method has low computational complexity and is robust with respect to the feature representation of the video frames. We compare our method against the top-performing online method from our previous study (called 'submodular convex optimisation' [6]) and a baseline method of uniform sampling of events (named 'uniform events').

The rest of the paper is organised as follows: Section 2 reviews related works. Our new summarisation method is introduced in Section 3, followed by its quantitative evaluation and summarisation examples in Section 4. Finally, Section 5 offers the conclusions.

2 Related Work

Application-specific surveys provide comprehensive comparisons among existing video summarisation methods on egocentric videos [5] and traditional videos (third-person view) [22]. We recently carried out a survey on online video summarisation for traditional videos [11].

For online applications, typically the video is segmented into smaller units of interest (shots, scenes, events) following two strategies: detecting changes of the content information [1, 2, 19, 22]; or grouping frames into clusters using distribution model [21, 16], connectivity model [6, 13] or centroid model [3]. Subsequently,

keyframes are selected based on their temporal positions [2, 19, 22]; central positions in clusters [3, 6, 21] or the relative values of the metric measuring content information [1, 13].

The number of keyframes can be either determined by the algorithm itself [1, 2, 6, 13, 16, 19, 21, 22] or defined by some cardinality constraint [3, 18, 5].

3 Online Video Summarisation

Consider a scenario where the user's daily activities are recorded using a wearable camera. To create an online summary, the video frames are represented as feature vectors in some feature space. A 'budget' is set as the maximal allowed number of frames in the summary. Next, the system saves the extracted keyframes generated by the online video summarisation algorithm if the budget allows for this. Should the limit be reached, one or more of the frames already stored in the summary is removed. Below we explain the steps of our algorithm.

3.1 Budget-constrained online video summarisation

In statistics, control charts have been used to monitor and control ongoing processes over time. Previously [12], we introduced the use of control charts to identify event boundaries from a streaming video. The closest frame to the center of each event, represented as a cluster in the feature space, is selected as a keyframe. Here, we additionally, impose a constraint on the number of keyframes, hence the term 'budget-constrained' video summarisation. We also introduce a dynamic, similarity threshold into the algorithm that varies the probability of selecting new keyframes according to the number of existing keyframes and total budget. The pseudo-code of the algorithm is given in Algorithm 1 [1].

Given an integer constant β, the purpose is to select a set of no more than β keyframes which describe the video as fully and accurately as possible. Unlike the classical summarisation approaches, we derive the summary on-the-go by processing each frame as it comes and selecting keyframes before the full video content is available. The algorithm requires only a limited memory to keep the frames selected thus far, and the frames belonging to the current event.

A control chart is used to detect the event boundaries [20]. The quantity being monitored is the difference between consecutive frames, defined by the distance between the frames in some chosen feature space $\mathbb{R}^L$. Assuming that the frames are represented as points $\mathbb{R}^L$, the hypothesis is that different events in the video are represented by relatively distant clusters. Then transition from one event to the next will be associated with large distance between consecutive frames. As both outlier and transition frames may be detected as an event boundary, we observe a minimum event size, m. If the number of frames in an event is less than m, the algorithm ignores the candidate-event without extracting a

[1] Matlab code is available at: https://github.com/pariay/Budget-constrained-Online-Video-Summarisation-of-Egocentric-Video

Algorithm 1 Budget-constrained online video summarisation

Input: Data stream $F = \{f_1, \ldots, f_N\}, f_i \in \mathbb{R}^L$, initial buffer size b, minimum event length m, threshold parameter for keyframe difference θ, desired number of keyframes β.

Output: Selected set of keyframes $K \subset F, |K| \leq \beta$.

INITIALISATION

$K \leftarrow \emptyset$

$E \leftarrow \{f_1, \ldots, f_b\}$ ▷ initial buffer

Calculate the $b-1$ distances between the consecutive frames in E.

$\mu \leftarrow$ average distance.

$\sigma \leftarrow$ standard deviation.

PROCESSING OF THE VIDEO

for frame number $i = b+1, \ldots, N$ **do**

  $d_i \leftarrow d(f_i, f_{i-1})$ ▷ calculate distance to previous frame

  if $d_i <= \mu + 3\sigma$ **then** ▷ - **same event**

    $[\mu, \sigma] \leftarrow$ update μ & σ with d_i

    $E \leftarrow E \cup f_i$ ▷ store the frame

  else if $|E| < m$ **then** ▷ - - - - - - - - - - - - - - - - - - **event too short**

    $E \leftarrow f_i$ ▷ remove frames in E and start a new event

  else ▷ - - - - - - - - - - - - - - **event sufficiently long**

    $k \leftarrow$ SELECT-KEYFRAME(E)

    if K empty **then** ▷ - - - - - - - - - - - - - - - - - - **first keyframe**

      $K \leftarrow k$

    else ▷ - - k **included if sufficiently different to K**

      $k_{last} \leftarrow$ last keyframe in K

      $\delta \leftarrow$ KEYFRAME-DIFF(k, k_{last})

      $\delta_{min} \leftarrow$ smallest distance among consecutive keyframes in K

      if $|K| < \beta$ & $\delta >$ DIFF-THRESHOLD($|K|, i, \theta, \beta, N$) **then** ▷ - **in budget**

        $K \leftarrow K \cup k$

      else if $\delta >= \delta_{min}$ **then** ▷ **over budget**

        Remove from K one of the keyframes in the closest pair.

        $K \leftarrow K \cup k$

    $E \leftarrow f_i$ ▷ new event

FUNCTIONS

Function $f =$ SELECT-KEYFRAME($data$)

  $f \leftarrow \arg\min_{x \in data} d(x, \text{mean(data)})$

Function $\delta =$ KEYFRAME-DIFF(f_1, f_2)

  $h_i \leftarrow$hist16(hue(f_i)) ▷ Normalised 16-bin Hue histogram

  $\delta = \frac{1}{16} \sum_{j=1}^{16} |h_1(j) - h_2(j)|$

Function $\theta_{new} =$ DIFF-THRESHOLD(n_k, t, θ, β, T)

  $n_t \leftarrow \beta \times t/T$ ▷ Expected number of keyframes, assuming linear distribution

  if $n_t == \beta$ **then**

    $\theta_{new} = 0$

  else

    $\theta_{new} \leftarrow \frac{\theta \times (\beta - n_k) + (n_k - n_t)}{\beta - n_t}$

keyframe. This approach is suitable for clearly distinguishable shots (events) [12]. For application to egocentric videos, in this paper we adapt the approach to allow for less well-defined shots. In addition, the budget constraint provides a means of defining an expected or desired number of events to be captured. Egocentric videos are not easily split into coherent events. To improve the event detection, we compare a selected keyframe with its immediate predecessor. If the keyframes of the adjacent events are deemed similar, the new event is ignored, without extracting a keyframe. The tolerance for accepting similarity between frames varies in relation to how close to the overall budget the existing set of keyframes is, and how many more events may be expected in the video. Note that this assumes prior knowledge of roughly how long the video will be. If the budget for keyframes is reached while frames are still being captured, keyframes from any additional events are only saved if the keyframe set is made more diverse by the substitution of the new keyframe for an existing keyframe.

Assume a video stream is presented as a sequence of frames, $F = \{f_1, \ldots, f_N\}$, $f_i \in \mathbb{R}^L$, where L indicates the dimensions of the frame descriptor. For any upcoming frame, the similarity of consecutive frames f_i and f_{i-1} is calculated using Euclidean distance $d(.,.)$ in $\mathbb{R}^L$. Denote $d_i = d(f_i, f_{i-1})$. In the process of monitoring quality control, the probability p of an object being defective is known from the product specifications or trading standards. This probability is the quantity being monitored. For the event boundary detection in videos, we need to monitor the distance d_i. The initial values can be calculated by taking average values of the first b distances: $\mu = \frac{1}{b-1}\sum_{i=2}^{b} d_i$, and computing the standard deviation value of the first b distances as: $\sigma = \sqrt{1/(b-1)\sum_{i=2}^{b}(d_i - \mu)^2}$. At time point $i+1$, the distance value d_{i+1} is calculated and compared with the μ and σ at time point i. A change is detected if $d_{i+1} > \mu + \alpha\sigma$. The value of α typically is set to 3, but other alternatives are also possible.

The measure of similarity between two selected adjacent keyframes follows the study of De Avila et al. [4]. Those keyframes are represented by 16-bins histograms of the hue value (H). Keyframes are similar if the Minkowski distance between their normalised histograms is less than a threshold θ, and are dissimilar otherwise.

The proposed algorithm requires four parameters: the initial buffer size (b), the minimum event length (m), the pre-defined threshold value for keyframe similarity (θ), and the maximum number of keyframes (β).

3.2 Choosing parameter values

An empirical value for the desired number of the keyframes, β, has been obtained following the study by Le et al. [9]. The authors collected a total of 80 image sets from 16 participants from 9am to 10pm using lifelogging devices. An average of 28 frames per image set were chosen by the participants to represent their day. Therefore, in our experiment we set this parameter to $\beta = 28$. We sample one frame per second for each video. The buffer size b was selected to be equal to one

minute, $b = 60$. The minimum event length was set to thirteen seconds, $m = 13$. The threshold value for keyframe similarity was set to $\theta = 0.7$.

3.3 Selecting a feature representation

The proposed algorithm is not tailor-made for any particular descriptor, therefore any type of feature space may be applied. For an online application, two factors must be considered when choosing a descriptor: good representation ability and low computational cost. Following a preliminary study involving 7 descriptors, including two convolutional neural networks, we chose the RGB feature space as the best compromise between the two criteria. This work is presented at Computer Graphics and Visual Computing 2018, ("Selecting Feature Representation for Online Summarisation of Egocentric Videos"). The RGB colour moments (mean and standard deviation) are obtained by dividing an image uniformly into 3×3 blocks. The mean and the standard deviation for each block and colour channel are computed, giving a feature space of dimensionality $L = 9 \times 6 = 54$.

4 Experimental Results

4.1 Dataset

The algorithm performance was evaluated on the Activity of Daily Living (ADL) dataset[2] [17]. This dataset was recorded using a chest-mounted GoPro camera and consists of 20 videos (each lasting about 30 minutes to one hour) of subjects performing their daily activities in the house.

4.2 Evaluation

Evaluation of keyframe video summarisation for egocentric videos is still a challenging task [5, 7]. Yeung et al. [24] suggested to evaluate summaries through text using the VideoSET method[3]. In their experiments, the author provided text annotations per frame for the video to be summarised. The VideoSet method converts the summary into text representation. Then the content similarity between this representation and a ground truth text summary was measured through Natural Language Processing (NLP).

Motivated by [24], we annotated the ADL dataset rather using numbers than text. The numbers are organised to describe sequences of events. We made a list of events in each video, using an action list from [17]. The frames are labelled with their relevant event, or as not informative if the event cannot be recognised from the frame (semantic information). Consequently, any informative frame from the event can be considered ground truth for that event. Given a video summary, the number of matches and then the F-measure can be subsequently calculated.

[2] https://www.csee.umbc.edu/~hpirsiav/papers/ADLdataset/
[3] http://ai.stanford.edu/~syyeung/videoset.html

4.3 Online Summarisation Methods

We compared the following methods:

(a) BCC. The proposed Budget-constrained Control Chart algorithm.
(b) SCX. Submodular convex optimisation [6].
(c) UE. Uniform Events (baseline method). To implement the UE algorithm, the video is uniformly divided into ϵ number of events (segments). The ϵ value follows the number of keyframes extracted by our online algorithm. The closest frame to the center of each segment (in $\mathbb{R}^L$) is taken to represent the event.

To have a fair comparison we tuned the SCX and the UE for each video to their best performance. Doing that, the value for ϵ was adjusted with the number of keyframes extracted by our online algorithm. The same adjustment applied for the SCX.

4.4 Keyframe Selection Results

Table 1 shows the F-value for the match between the summaries generated through BCC, SCX and UE, and the semantic-category ground truth for the 20 videos. As seen from these results, the proposed online method performs consistently better than the two competitors.

Figure 2 displays the summaries obtained by the BCC, SCX and UE methods, highlighting matched frames with the ground truth. Our BCC method misses one event in the ground truth (Figure 2a) resulting in the F-value of 0.89.

5 Conclusion

The purpose of the current study was to introduce a fast and effective method (BCC) to extract a keyframe summary from a streaming video. The proposed method applies control charts to detect event boundaries online, and observes a maximum limit on the number of selected keyframes (budged-constrained). Our experiments with 20 egocentric videos from the ADL video database demonstrate that BCC performs well in comparison with two existing methods, state-of-the-art SCX and baseline UE.

The requirement to store all frames for an event before the keyframe is selected could present memory issues in the event of excessively long, sedentary events, e.g. sleeping. One way to deal with this issue is the introduction of a dynamic frame-rate, with far fewer frames recorded during such events.

References

1. Wael Abd-Almageed. Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing. In *IEEE 15th International Conference on Image Processing (ICIP 2008)*, pages 3200–3203, Oct. 2008.

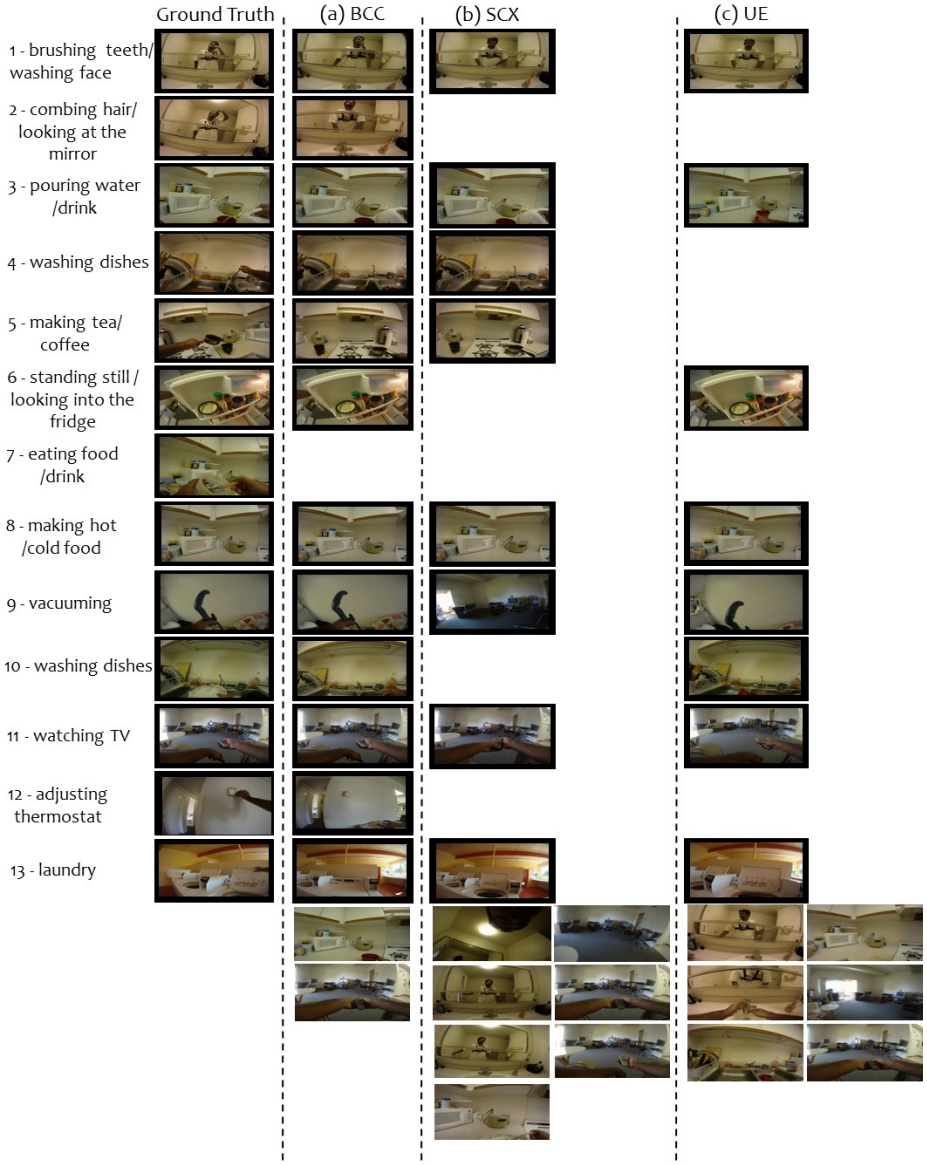

Fig. 2: Example of keyframe summaries obtained by the (a) BCC, (b) SCX and (c) UE methods and their matched frames with the ground truth, for ADL dataset video #16. The total number of events in ground truth for this video is 13, and the BCC just missed one event on eating food/drink.

Table 1: F-values for the comparison of the proposed method (BCC), and the two rival methods (SXC and UE) on the 20 videos in ADL video database.

		F-measure			Parameters	
Video	Number of Frames	BCC	SCX	UE	SCX(λ)	UE(ϵ)
P_{01}	1,794	**0.73**	0.45	0.60	0.33	13
P_{02}	2,860	0.63	0.35	**0.67**	0.07	27
P_{03}	2,370	0.50	0.37	**0.56**	0.15	19
P_{04}	1,578	**0.52**	0.31	0.44	0.25	18
P_{05}	1,475	**0.42**	0.30	**0.42**	1	5
P_{06}	1,550	**0.67**	0.53	0.47	0.2	20
P_{07}	2,643	**0.81**	0.43	0.54	0.17	18
P_{08}	1,592	0.56	0.40	**0.60**	0.08	27
P_{09}	1,288	**0.67**	0.61	0.56	0.15	25
P_{10}	956	**0.80**	0.40	**0.80**	0.7	8
P_{11}	493	**0.87**	0.52	0.78	0.6	10
P_{12}	844	**0.69**	0.43	**0.69**	0.3	14
P_{13}	1,768	**0.63**	0.28	0.51	0.11	24
P_{14}	1,531	**0.78**	0.54	0.63	0.09	23
P_{15}	1,585	**0.59**	0.37	**0.59**	0.25	13
P_{16}	840	**0.89**	0.64	0.59	0.19	13
P_{17}	885	**0.44**	**0.44**	0.22	0.28	9
P_{18}	1,150	**0.47**	**0.47**	0.40	0.095	21
P_{19}	3,797	**0.77**	0.33	0.57	0.08	28
P_{20}	1,609	**0.69**	0.31	0.50	0.17	16

2. Jurandy Almeida, Neucimar J Leite, and Ricardo da S Torres. Vison: Video summarization for online applications. *Pattern Recognition Letters*, 33(4):397–409, Mar. 2012.
3. Rushil Anirudh, Ahnaf Masroor, and Pavan Turaga. Diversity promoting online sampling for streaming video summarization. In *IEEE International Conference on Image Processing (ICIP2016)*, pages 3329–3333, Sept. 2016.
4. S. E. F. de Avila, A. P. B. Lopes, A. da Luz, and A. de Albuquerque Araújo. VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68, Jan. 2011.
5. Ana Garcia del Molino, Cheston Tan, Joo-Hwee Lim, and Ah-Hwee Tan. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems*, 47(1):65–76, 2017.
6. Ehsan Elhamifar and M Clara De Paolis Kaluza. Online summarization via submodular and convex optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2017)*, pages 1818–1826, Jul. 2017.
7. Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European conference on computer vision (ECCV14)*, pages 505–520. Springer, 2014.
8. Jacqueline Kerr, Simon J Marshall, Suneeta Godbole, Jacqueline Chen, Amanda Legge, Aiden R Doherty, Paul Kelly, Melody Oliver, Hannah M Badland, and Charlie Foster. Using the sensecam to improve classifications of sedentary behavior

in free-living settings. *American journal of preventive medicine*, 44(3):290–296, 2013.
9. Huy Viet Le, Sarah Clinch, Corina Sas, Tilman Dingler, Niels Henze, and Nigel Davies. Impact of video summary viewing on episodic memory recall: Design guidelines for video summarizations. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4793–4805, New York, NY, USA, May 07-12 2016. ACM.
10. Matthew L. Lee and Anind K. Dey. Lifelogging memory appliance for people with episodic memory impairment. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, UbiComp '08, pages 44–53, New York, USA, 2008. ACM.
11. Clare E. Matthews, Ludmila I. Kuncheva, and Paria Yousefi. Classification and comparison of on-line video summarisation methods. *Machine Vision and Applications*, May 2018. Submitted.
12. Clare E. Matthews, Paria Yousefi, and Ludmila I. Kuncheva. Using control charts for on-line video summarisation. In *14th Asian Conference on Computer Vision (ACCV 2018)*, July 2018. Submitted.
13. Shaohui Mei, Genliang Guan, Zhiyong Wang, Shuai Wan, Mingyi He, and David Dagan Feng. Video summarization via minimum sparse reconstruction. *Pattern Recognition*, 48(2):522–533, Feb. 2015.
14. Dorothy N Monekosso and Paolo Remagnino. Behavior analysis for assisted living. *IEEE Transactions on Automation science and Engineering*, 7(4):879–886, 2010.
15. Gillian O'Loughlin, Sarah Jane Cullen, Adrian McGoldrick, Siobhan O'Connor, Richard Blain, Shane O'Malley, and Giles D Warrington. Using a wearable camera to increase the accuracy of dietary analysis. *American journal of preventive medicine*, 44(3):297–301, 2013.
16. Shun-Hsing Ou, Chia-Han Lee, V Srinivasa Somayazulu, Yen-Kuang Chen, and Shao-Yi Chien. On-line multi-view video summarization for wireless video sensor network. *IEEE Journal of Selected Topics in Signal Processing*, 9(1):165–179, Feb. 2015.
17. Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR12)*, pages 2847–2854. IEEE, June 16-21 2012.
18. Bryan A Plummer, Matthew Brown, and Svetlana Lazebnik. Enhancing video summarization via vision-language embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1052–1060, July 2017.
19. Zeeshan Rasheed and Mubarak Shah. Scene detection in hollywood movies and tv shows. In *Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 343–343, Jun. 2003.
20. Walter Andrew Shewhart. *Economic control of quality of manufactured product*. Van Nostrand Company, 1931.
21. Mingzhou Song and Hongbin Wang. Highly efficient incremental estimation of gaussian mixture models for online data stream clustering. In *Intelligent Computing: Theory and Applications III, SPIE*, volume 5803, pages 174–184, Mar. 2005.
22. Ba Tu Truong and Svetha Venkatesh. Video abstraction: A systematic review and classification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 3(1):3, Feb. 2007.
23. Emma Woodberry, Georgina Browne, Steve Hodges, Peter Watson, Narinder Kapur, and Ken Woodberry. The use of a wearable camera improves autobiographical memory in patients with alzheimer's disease. *Memory*, 23(3):340–349, 2015.
24. Serena Yeung, Alireza Fathi, and Li Fei-Fei. VideoSET: Video summary evaluation through text. *CoRR, arXiv preprint arXiv:1406.5824*, abs/1406.5824, 2014.