Selective Keyframe Summarisation for Egocentric Videos Based on Semantic Concept Search*

Paria Yousefi
School of Computer Science and Electronic Engineering
Bangor University
Bangor, UK
paria.yousefi@bangor.ac.uk

Ludmila I Kuncheva
School of Computer Science and Electronic Engineering
Bangor University
Bangor, UK
l.i.kuncheva@bangor.ac.uk

***Abstract*—Large volumes of egocentric video data are being continually collected every day. While the standard video summarisation approach offers all-purpose summaries, here we propose a method for selective video summarisation. The user can query the video with an unlimited vocabulary of terms. The result is a time-tagged summary of keyframes related to the query concept. Our method uses a pre-trained Convolutional Neural Network (CNN) for the semantic search, and visualises the generated summary as a compass. Two commonly used datasets were chosen for the evaluation: UTEgo egocentric video and EDUB lifelog.**

***Index Terms*—egocentric video, video summarisation, keyframe selection, first person vision, semantic search**

I. Introduction

It is a foregone conclusion that in the near future every aspect of life will be captured on camera. Wearable cameras such as Narrative Clip and GoPro allow consumers to record any single moment of their lives. As a result, vast amounts of unconstrained data are produced. However, those recorded visual memories may never be revisited by the device wearer. The important images can be organised into keyframe summaries, whilst the repetitive or meaningless images are discarded. Here we propose a method for generating a keyframe summary of a video to answer a user's query, for example: when and what did I eat today?

Many state-of-the-art summarisation methods were built for optimising a predefined criterion related to story coherence: diversity [1], [2]; representativeness [3], [4]; importance [5], [6]; visual aesthetics [7], [8] and first-person engagement [9]. These methods generate a single summary for all users which may not suit everyone, given the unconstrained scenarios in most egocentric videos and lifelogging data streams. A single summary can be suitable in some controlled domains such as video surveillance of a specific area with constant background and predefined salient events. Available annotated data also show considerable discrepancies between summaries made by different users [10]. Users may prefer to obtain a summary of related to a specific concept or event. For instance, a user who follows a diet would be interested in a summary of their eating routine during the day. An elderly user may want to extract summary of faces of the people they have met during their day. Several query-based summarisation methods have been proposed recently either in the form of a sequence of shots (video skimming) [11]–[13] or as an interactively constructed collection of keyframes [14]. These methods, however are often supervised or require user interaction to guide the shot selection.

Here we propose a new summarisation approach where, unlike any of the studies before, we preserve the frame-time relationship in order to answer the question 'when?' (Figure 1). First, the video is mined for frames related to a given concept. The user's query is given as a word (e.g., food, phone, laptop, book). The identified frames are grouped along the timeline to form events. Each event is subsequently represented with one frame in the final summary, which we visualise as a compass diagram. Including the time tags in constructing the final keyframe summary set, our approach apart from the other query-based summarisation approaches.

We refer to the final set of keyframes as a "selective summary". The proposed method can be useful in retrieving memories of daily experiences, behaviours of interest or concern, of in spotting rare occasions when a certain object becomes a part of the view.

The rest of the paper is organised as follows: Related work has been reviewed in the Introduction. Our new summarisation approach is described in Section II, followed by its quantitative evaluation and summarisation examples in Section III. Finally, Section IV offers the conclusions and outlines our future work.

II. Methodology

A. Description of the proposed process

Figure 1 illustrates the proposed approach[1], and Figure 2 depicts the steps of the implementation algorithm. First, after obtaining the user's query, we identify all frames in the video related to it through semantic concept search. Next, we apply an algorithm which we call "occurrence-led clustering" to find time intervals which will be the events to summarise. At the next step, we extract keyframes from the events. Finally, we visualise the summary using a new approach, which we term a "compass summary".

This study was supported by Project RPG-2015-188 sponsored by the Leverhulme Trust, UK.

[1]Matlab code is available at: https://github.com/pariay/Selective-Summary.

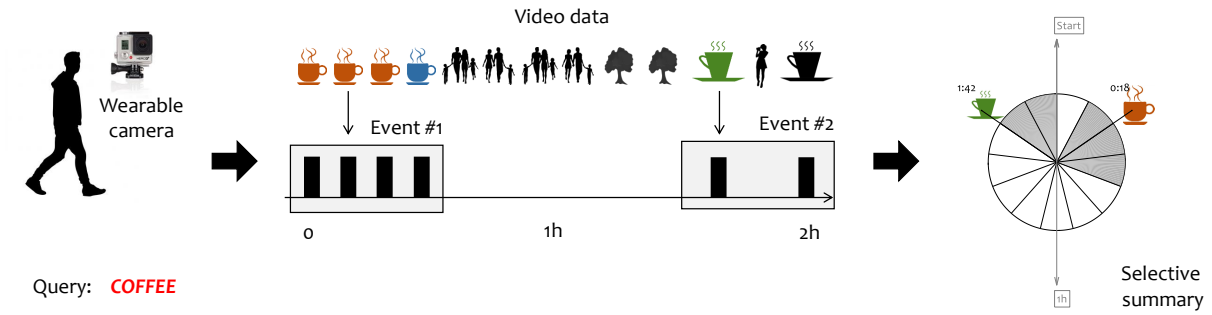


Fig. 1: Diagram of the proposed method for selective egocentric video summarisation.

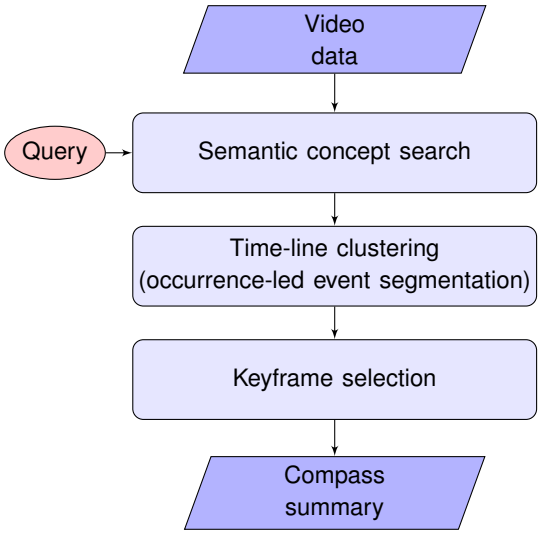


Fig. 2: Flowchart of the proposed method for selective video summarisation.

B. Semantic Concept Search

In order to compute the object representation, we propose to use the winner of the ImageNet Large Scale Visual Recognition Competition 2015 (ILSVRC), Residual Network (ResNet) [15]. As a result, for each frame, the network returns a set of lexical concepts of the detected dominant object along with a prediction score. For example, a dog could be presented as the dominant object in the frame with score 0.2, measuring the certainty that the identified object corresponds to the image content. Inspired by Dimiccoli et al. [16], we used WordNet [17], [18] to post-process the results and calculate the similarity score between a detected object and the user's query. WordNet is a lexical database which groups English words into a set of synonyms, provides a short definition of the words and shows usage examples. The value for a given frame is calculated as follows. The word representing the dominant object detected by ResNet and the query are entered in WordNet, which then outputs a degree of similarity. This degree varies from 0 for dissimilarity to 1 for identity. We considered the frame to be relevant to the user query if the similarity was equal to 1.

The semantic search algorithm returns a vector representing the presence (label 1) or absence (label 0) of the user's query for each frame in the video.

The CNN (ResNet) used here has been pre-trained on images with a canonical view and correct level of illumination without any motion blur. These conditions are rarely met in egocentric images. Therefore, we set a threshold of 0.3 on the probability prediction score of the CNN. Frames with dominant objects whose score is less than the threshold are considered to be empty.

Some popular queries have bespoke solutions. An example is 'food'. For a user with an eating disorder problem (overeating or under-eating), it is important to regularly check their dietary routine (by themselves or by a doctor). Being of a great public interest, the problem of detecting food has been addressed in the past as a binary classification problem where the algorithm has to distinguish whether the given image contains food or not [19]–[22]. Our approach can make use of such solutions at the semantic search step, bypassing the need to use ResNet and WordNet.

An example of the semantic search step is shown below. Figure 3 shows a frame from video P02. ResNet returned description: car mirror. The search query was "automobile". The similarity score between the tokenised frame description and the query was assessed at value 1 by WordNet. According to our threshold, the frame was given label 1 indicating that it matches the query.

The poor quality (e.g., motion blur, composition, illumination) of the images in egocentric videos often leads to false positive and false negative detections. Two such examples are shown in Figure 4. The image in subplot (a) is a false positive detection for query 'television', and the image in (b) is a false negative for query 'food'. The true dominant objects in these images were respectively 'car window' or 'street' in (a) and 'food' in (b).

C. Occurrence-led Event Segmentation

An Occurrence-led Event Segmentation (OLES) is proposed here as the next step. The term "occurrence-led" is coined by us to denote the process of finding temporal clusters on the time line based on presence-absence (occurrence). After the frames relevant to the query have been identified, we cluster only their time occurrences (not the frame content

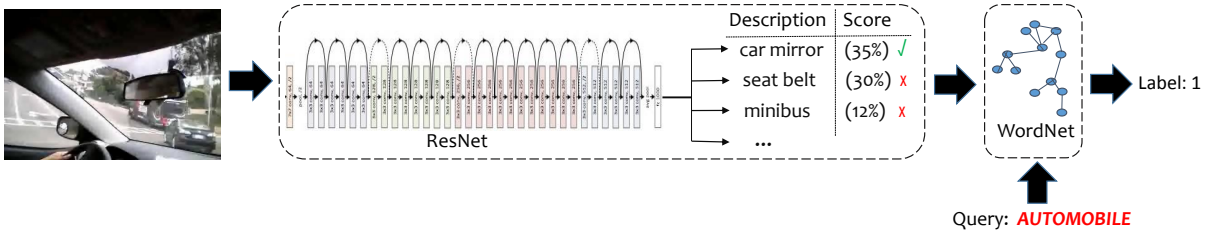


Fig. 3: Illustration of the semantic search process using a frame from video P02 (UTEgo dataset) and query 'automobile'.

(a) False positive (b) False negative

Fig. 4: Frames from egocentric videos P02 and P01 (UTEgo dataset) mislabelled by the semantic labelling algorithm. (a): false positive for 'television', and (b): false negative for food.

or feature representation). For a given concept, we prepare a binary vector with consecutive elements corresponding to the frames in the video. Value 1 indicates that the respective frame contains the concept of interest, and value 0, that it does not. Hierarchical Agglomerative Clustering (AC) was applied to cluster *time-adjacent* frames together based on their geometric centroid.

Consider the toy example in Figure 1. The query "coffee" returns the following vector relating the 13 frames with the searched concept:

$$\underbrace{1\ 1\ 1\ 1}_{Event\ \#1}\ 0\ 0\ 0\ 0\ 0\ 0\ \underbrace{1\ 0\ 1}_{Event\ \#2}$$

The data which we cluster here is the sequence of *occurrences* of the query concept on the time line. We apply the single linkage procedure using the centroid method.

One drawback of nearly any clustering method, including hierarchical clustering, is that the number of clusters is not known in advance. When we cluster a single-dimensional time variable, we have the advantage of being able to interpret the clusters and pose time constraints as deemed necessary. For the video summarisation purposes, we can argue that an event should not be shorter than a given time interval, and that the time gap between events should be no less than a given amount. If two candidate-events are closer to one another than this gap, they are likely parts of the same event. In the toy example, imposing the restriction that the centroids of two clusters must not be closer than 3 frames, the method returns two clusters marked as Events above.

As to the minimum length of an event, we decided not to pose any restrictions. The reason for this are twofold. First, even a glimpse of a certain object may be of high interest. For example, a casual glance at a shelf with wines in the supermarket may need to be flagged in the summary. Second, the camera wearer may not be focusing their gaze on a particular object for a long time even though they may be interacting with this object. An example of this is a chat on the phone. The user may look at the screen for a moment to verify the caller's identity, and then the phone will be pressed to the user's ear, and out of the camera view. For the gap between events, though, we chose a 20-minute threshold. Given the typical length of the egocentric videos (few hours), and lifelog records, we found that this threshold leads to summaries of reasonable length.

D. Keyframe Selection

Once the events have been determined through OLES, the next step is to select a good subset of keyframes (one keyframe per event). This step needs a feature representation of all frames. For this representation we chose the 4096 deep features extracted from the last fully-connected layer (FC7) of the Convolutional Neural Networks (ConvNets or CNN). The runner-up in ILSVRC 2014, known as VGGNet architecture [23]. Treating the temporal events as "clusters" in the respective 4096-dimensional space, the frame closest to the centroid of the cluster was chosen to represent that cluster.

E. The Compass Summary Visualisation

We demonstrate the result of our summarisation method using a "compass view" as shown in Figure 5. Consider query "phone" in video P01 from the UTEgo dataset [5]. The semantic concept search identified 90 frames containing a mobile phone as the dominant object. (The actual number of frames related to the "phone" query is 153.)

The duration of the video, rounded up to the closest hour, is represented by a circle, and the hours are denoted with annotated long spikes. The individual frames where the query concept is found, are plotted with short black spikes (90 in this case). Shaded sectors of the circle are the events detected through the OLES algorithm. Finally, the spikes with the offset images are the proposed summary. The summary should be read clockwise, starting from the box 'Start' at the top.

The compass view allows the user to see the whole video at a glance and indicates the time positions of the summary frames.

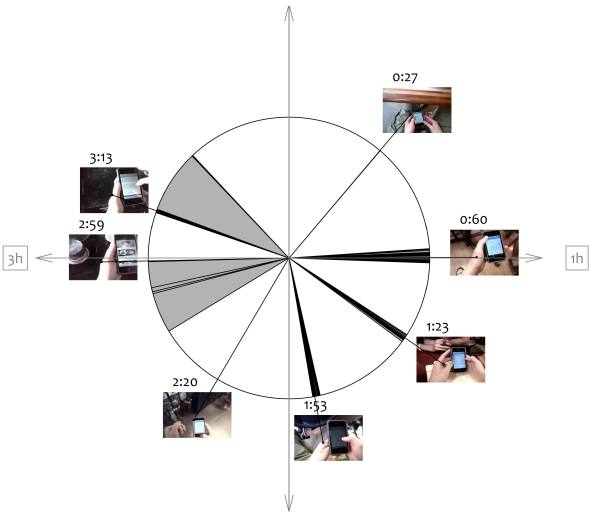


Fig. 5: An example of a compass summary of the system's output for video P01 from the UTE database for query 'phone'.

III. Experimental Results

This section presents quantitative experimental results on two egocentric datasets. The aim of the experiment is to demonstrate the effectiveness of the presented selective keyframe summarisation process. In the first leg of this experiment, we assess quantitatively the semantic concept search. This part of the pipeline pre-determines the success of the subsequent clustering and keyframe selection parts (Figure 2), dictating to a large extent the quality of the final summary. Next, we estimate the effectiveness of the whole selective summary.

A. Datasets

To demonstrate the performance of the approach, two datasets were selected: the University of Texas Egocentric video (UTEgo) [5]; and the Egocentric Dataset of the University of Barcelona-objects (EDUB-obj) [24]. The given results illustrate that our selective summary approach works on both type of data (egocentric video and lifelog series of images).

The UTEgo contains 4 long videos (each lasting about 3-4 hours) of subjects performing their different daily activities: shopping, eating, cooking, attending lectures and driving. The videos were recorded at low-quality frame rate (15 frames/ seconds) with 350x480 resolution per frame. Each video was sub-sampled taking one frame per four seconds, thereby reducing the number of frames as follows: P01: 3464 frames; P02: 4566 frames; P03: 2696 frames; and P04: 4446 frames.

The EDUB-Obj comprises of 4916 images of daily activities: eating, working, attending meetings and shopping. Images were recorded by 4 different subjects in 8 different days (each of them having captured 2 days). This dataset is acquired by the wearable Narrative camera which captures images in a passive way every 30-60 seconds. Number of images per subject are as follows: Subject 1-1: 588 images; Subject 1-2: 721 images; Subject 2-1: 589 images; Subject 2-2: 557 images; Subject 3-1: 726 images; Subject 3-2: 437 images; Subject 4-1: 610 images; and Subject 4-2: 684 images.

We prepared a ground truth by identifying the dominant object for each individual frame for all videos. The most common objects found in both datasets were: car, food, phone, laptop/computer. In addition there were other objects such as: glass, beer, coffee, book, desk, light, sign, refrigerator and television. We are interested in one dominant object per frame, and ignore any other object in that particular frame.

B. Effectiveness of the Semantic Search Algorithm

For each video, we identified the most represented objects. Then we applied the semantic search, separately for each identified object. To do this, the frames were labelled with 0 and 1, as described in Section II-B. The result from the semantic search was represented in the same format, which allowed us to calculate Precision, Recall, and the F-measure (2*Precision*Recall/(Precision+Recall)). For each video we averaged the Precision, Recall, and the F-measure across the query terms. The results are shown in Table I.

The table shows that our detection algorithm performs well in finding frames related to the user search (high Precision values), however it also misses a considerable number of frames which are related to the concept (low Recall values). Considering that we are using poor quality images (egocentric video), we regard our semantic search as reasonably successful.

C. Effectiveness of the Selective Summarisation Method

The aim of this part are: (1) to determine the success of the Occurrence-led Event Segmentation algorithm followed by

TABLE I: Result of the concept search algorithm for different user queries per video (in %)

Dataset	*Name*	*Precision*	*Recall*	*F-measure*	*Concepts*
	P01	92.2	49.2	60.4	food, car, phone, computer, shoe
UTEgo	P02	80.4	26.2	36.6	food, car, glass, book, television
	P03	88.7	37.5	49.5	food, car, phone, grocery, refrigerator, washbasin
	P04	100	20	31.7	food, laptop, book
	Subject 1-1	88.5	34.5	39.5	food, car, phone, building
	Subject 1-2	80.4	54.1	61.6	food, car, mobile, beer, coffee, glass, cup, sign
	Subject 2-1	100	55.5	67.8	phone, computer, light, grocery
EDUB	Subject 2-2	83.2	37	47.8	food, phone, glass, laptop, light
	Subject 3-1	87.75	40.5	46.5	phone, laptop, book, train
	Subject 3-2	99.5	46	58	food, phone, computer, desk
	Subject 4-1	100	33.3	48.7	computer, desk, building
	Subject 4-2	94.7	24.7	44	car, computer, train

TABLE II: Results of the Selective Summary process for different user queries per video (in %).

		Selective Summary without Concept Search algorithm			*Selective Summary method*		
Dataset	*Name*	*Precision*	*Recall*	*F-measure*	*Precision*	*Recall*	*F-measure*
	P01	96.6	87.6	91.4	70	88.4	75.4
UTEgo	P02	78.2	100	87.4	72.6	78.4	70.8
	P03	95.8	100	97.7	86.2	90.3	86.8
	P04	83.3	100	89	85.7	100	91
	Subject 1-1	85	91.75	87.5	80	79.25	70
	Subject 1-2	81.8	93.8	84.63	54.5	72.9	58.4
	Subject 2-1	75	100	83.5	68.3	81.3	61.8
EDUB	Subject 2-2	93.4	100	96	70	90	72.8
	Subject 3-1	92.5	93.8	92	80	80	70.8
	Subject 3-2	74.3	100	82.5	71.8	87.5	71.8
	Subject 4-1	100	90.3	94.3	100	73.7	83.3
	Subject 4-2	75	100	80	48.7	100	57.7

the keyframe selection; and (2) subsequently to determine the effectiveness of the entire selective summary method.

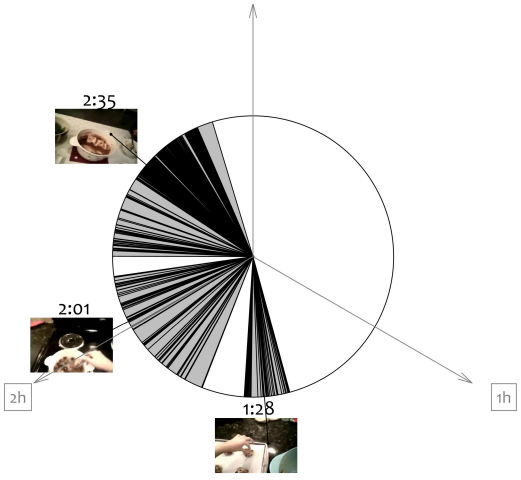
To this end, we made a user summary U for each video and each concept: 'phone', 'food', and 'car'. The selected frames account for the events when the camera wearer is interacting with the object of interest (one frame per event). An ideal output from our method would match reasonably the number, timing and content of U. We must note, however, that many frames of different visual content and at different time moments may represent the same event equally well. Thus, a summary returned by our method may not be an ideal match for U and still be of high quality.

For the first part of the evaluation, for every concept w, we applied OLES and the keyframe selection algorithm to the frames *manually labelled as* w. Thus we bypass the semantic search part and assume an ideal input for the OLES and keyframe selection. The resultant keyframe summaries were compared with those for U. The left part of Table II provides the experimental results for this part. This time, the matches were calculated as follows: a keyframe containing the object of interest is considered true positive (TP), if the event it represents is also represented by a keyframe in U. Frames in U which were not associated with an event returned by OLES were considered false negative (FN). Finally, a frame representing event which was not included in U is considered false positive (FP). The values are averaged across the queries.

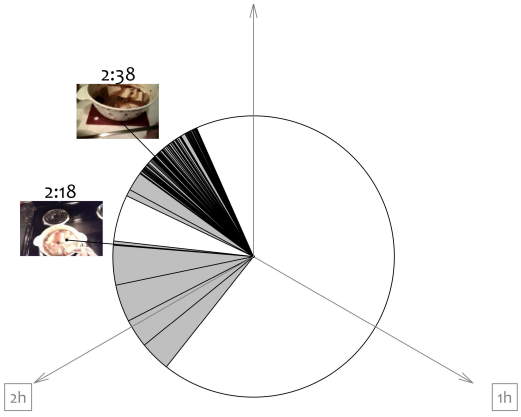
For the second part, we applied OLES and the keyframe selection algorithm to the frames returned by the semantic concept search. The results are presented in the right part of Table II. As expected, the values are lower due to the imperfection of the semantic search part of the pipeline.

Figure 6 displays an example from the UTEgo video (P03)

(a) User keyframe selection (ground truth)



(b) Without Semantic Search



(c) With Semantic Search.

Fig. 6: An example keyframes of the ground truth summary U and the proposed summary for video P03 of the UTEgo dataset. The user's query is 'food'.

answering a user's query on "food". Our selective summarisation method misses an event 1.5 hours into the video (Figure 6-c). We note that the frames returned by the closest-to-centroid keyframe selection method in Figure 6-b are very close to the user selection, both semantically and visually. This indicates that, should we have a better semantic search algorithm, the selective summarisation method may be expected to be accurate and useful.

The match counts for this example are as follows: for Figure 6-(b): $TP = 3$, $FP = 0$, $FN = 0$, ($F_{=}100\%$); and for Figure 6-(c): $TP = 2$, $FP = 0$, $FN = 1$, ($F = 80\%$).

IV. Conclusion

We propose a method to extract a selective, time-aware keyframe summary of an egocentric video. The problem was solved by applying a pipeline of a semantic concept search, occurrence-led event segmentation, and finally a cluster centroid keyframe selection. A compass-type diagram was proposed to visualise the selective summary. We demonstrate the effectiveness of our system through experiments with user-defined ground truth and two egocentric video databases.

We found that the major bottleneck of our approach is the semantic search part. Identifying objects and their related concepts is a challenge when the images are blurred, the illumination is poor, and the scene is cluttered. This is the predominant type of images in egocentric video. Thus, the main possibility to improve the accuracy of our selective summarisation system would come from honing the object detection and recognition in egocentric video.

Comparisons with alternative video summarisation methods would not be useful here because we are solving a different problem whereby the summary preserves the time position of the selected frames. We are not aware of other works proposing summarisation methods for this problem.

Future research direction include incorporating user searches on faces (known persons or general encounter of groups and crowds). This will involve face detection, people detection and face recognition. We were not able to explore this aspect with the publicly available databases because any faces in the frames were purposely blurred for identity protection. Experiments with own egocentric videos will give us the opportunity to expand the system in this direction.

Combining feature spaces is also an interesting area to explore for a potential improvement on keyframe selection.

A commercially built selective summarisation system may be used for monitoring addictive behaviours, e.g., those related to alcohol, smoking, and overeating.

References

[1] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2013*, June 2013, pp. 2714–2721.

[2] L. I. Kuncheva, P. Yousefi, and J. Almeida, "Edited nearest neighbour for selecting keyframe summaries of egocentric videos," *Journal of Visual Communication and Image Representation (JVCI)*, 2018. [Online]. Available: https://doi.org/10.1016/j.jvcir.2018.02.010

[3] M. Bolanos, R. Mestre, E. Talavera, X. Giró-i Nieto, and P. Radeva, "Visual summary of egocentric photostreams by representative keyframes," in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW2015)*. IEEE, 2015, pp. 1–6.

[4] L. I. Kuncheva, P. Yousefi, and J. Almeida, "Comparing keyframe summaries of egocentric videos: Closest-to-centroid baseline," in *Proceedings of the 7th International Conference on Image Processing Theory, Tools and Applications (IPTA 2017)*, Montreal, Canada, 2017.

[5] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Computer Vision and Pattern Recognition (CVPR12), IEEE Conference on*. IEEE, June 2012, pp. 1346–1353.

[6] Y. J. Lee and K. Grauman, "Predicting important objects for egocentric video summarization," *International Journal of Computer Vision*, vol. 114, no. 1, pp. 38–55, 2015.

[7] B. Xiong and K. Grauman, "Detecting snap points in egocentric video with a web photo prior," in *Proceedings of the European Conference of Computer Vision (ECCV14)*, vol. 8693 LNCS, no. PART 5, September 2014, pp. 282–298.

[8] V. Bettadapura, D. Castro, and I. Essa, "Discovering picturesque highlights from egocentric vacation videos," in *IEEE Winter Conference on Applications of Computer Vision (WACV2016)*. IEEE, 2016, pp. 1–9.

[9] Y.-C. Su and K. Grauman, "Detecting engagement in egocentric video," in *European Conference on Computer Vision (ECCV16)*. Springer, 2016, pp. 454–471.

[10] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *European conference on computer vision (ECCV14)*. Springer, 2014, pp. 505–520.

[11] A. Sharghi, J. S. Laurel, and B. Gong, "Query-focused video summarization: Dataset, evaluation, and a memory network based approach," in *Conference on Computer Vision and Pattern Recognition (CVPR17)*. IEEE, July 2017, pp. 2127–2136.

[12] P. Varini, G. Serra, and R. Cucchiara, "Personalized egocentric video summarization of cultural tour on user preferences input," *IEEE Transactions on Multimedia*, 2017.

[13] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *Proceedings of the European Conference on Computer Vision (ECCV14)*, Zurich, Switzerland, Sep 2014.

[14] A. Garcia del Molino, X. Boix, J.-H. Lim, and A.-H. Tan, "Active video summarization: Customized summaries via on-line interaction with the user," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, Feb 2017, pp. 4046–4052.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR16)*, June 2016, pp. 770–778.

[16] M. Dimiccoli, E. Talavera, S. G. Nikolov, and P. Radeva, "SR-Clustering: Semantic regularized clustering for egocentric photo streams segmentation," *Computer Vision and Image Understanding*, vol. 155, pp. 55–69, 2017.

[17] G. A. Miller, "WordNet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[18] G. Miller and C. Fellbaum, "Wordnet: An electronic lexical database," 1998.

[19] K. Aizawa, Y. Maruyama, H. Li, and C. Morikawa, "Food balance estimation by using personal dietary tendencies in a multimedia food log," *IEEE Transactions on multimedia*, vol. 15, no. 8, pp. 2176–2185, 2013.

[20] H. Kagaya and K. Aizawa, "Highly accurate food/non-food image classification based on a deep convolutional neural network," in *International Conference on Image Analysis and Processing (ICIAP15)*. Springer, 2015, pp. 350–357.

[21] A. Singla, L. Yuan, and T. Ebrahimi, "Food/non-food image classification and food categorization using pre-trained googlenet model," in *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*. ACM, 2016, pp. 3–11.

[22] F. Ragusa, V. Tomaselli, A. Furnari, S. Battiato, and G. M. Farinella, "Food vs non-food classification," in *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*. ACM, 2016, pp. 77–81.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] M. Bolaños and P. Radeva, "Ego-object discovery," *arXiv preprint arXiv:1504.01639*, vol. abs/1504.01639, 2015.