EG UK Computer Graphics & Visual Computing (2018), pp. 1–5
G. Tam and F. Vidal (Editors)

Selecting Feature Representation for Online Summarisation of Egocentric Videos

Paria Yousefi†, Ludmila I. Kuncheva and Clare E. Matthews

School of Computer Science, Bangor University, Bangor, UK

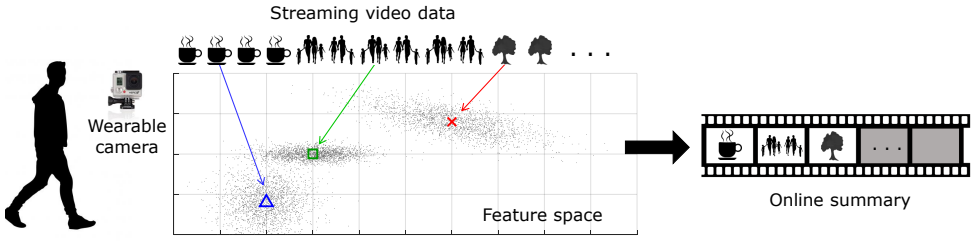

Figure 1: *We are looking to choose the most suitable feature space for creating online keyframe summary from a streaming egocentric video.*

Abstract

Visualising the content of a video through a keyframe summary has been a long-standing quest in computer vision. Using real egocentric videos, this paper explores the suitability of seven feature representations of the video frames for the purpose of online summarisation. Computational speed is an essential requirement in this set-up. We found that simple feature spaces such as HSV histograms and RGB moments are a good compromise between speed and representativeness in comparison with semantically richer but computationally more cumbersome spaces obtained through convolutional neural networks.

CCS Concepts

•***Computing methodologies*** → ***Computer vision; Video summarization; Image representations;***

1. Introduction

Video summarisation is the task of representing a video by a small and informative set of frames (keyframes) from the video. [TV07, MA08,MTLT17]. The lack of clear structure and visual consistency of first-person videos (FPV), called also egocentric videos, make their summarisation substantially more difficult [Bam15,MTLT17, BDR17,BMT*15]. Adding to the challenge, here we are interested in online summarisation of egocentric videos. In online summarisation, the keyframe summary is built on-the-fly so that at any moment during the video capture, there is a valid summary of the video up to that moment (Figure 1). While studies on online video summarisation exist [AA08,ALT13,OLS*15,RS03,MGW*15,AMT16, EK17], none is specifically dedicated to egocentric videos.

One important aspect of the video summarisation pipeline is the extraction of features from the video frames. Ideally, the feature representation will capture both the semantic content and the visual appearance of the frame. Many such representations have been proposed in the literature, ranging from low-level features (e.g., colour spaces) to high-level features (semantic-level description of the image content). In this paper we examine the suitability of seven feature spaces for online summarisation of egocentric video. Our experimental analysis is based on two egocentric videos from an activity recognition database available at: https://www.csee.umbc.edu/~hpirsiav/papers/ADLdataset/.

The rest of the paper is organised as follows. Section 2 describes the feature spaces. Section 3 contains our experiments, and Section 4, our conclusions and future work.

2. Feature representations

Figure 2 shows a possible taxonomy of the multitude of feature spaces (also called descriptors) used in the wider area of video

† Corresponding author. E-mail: p.yousefi@bangor.ac.uk

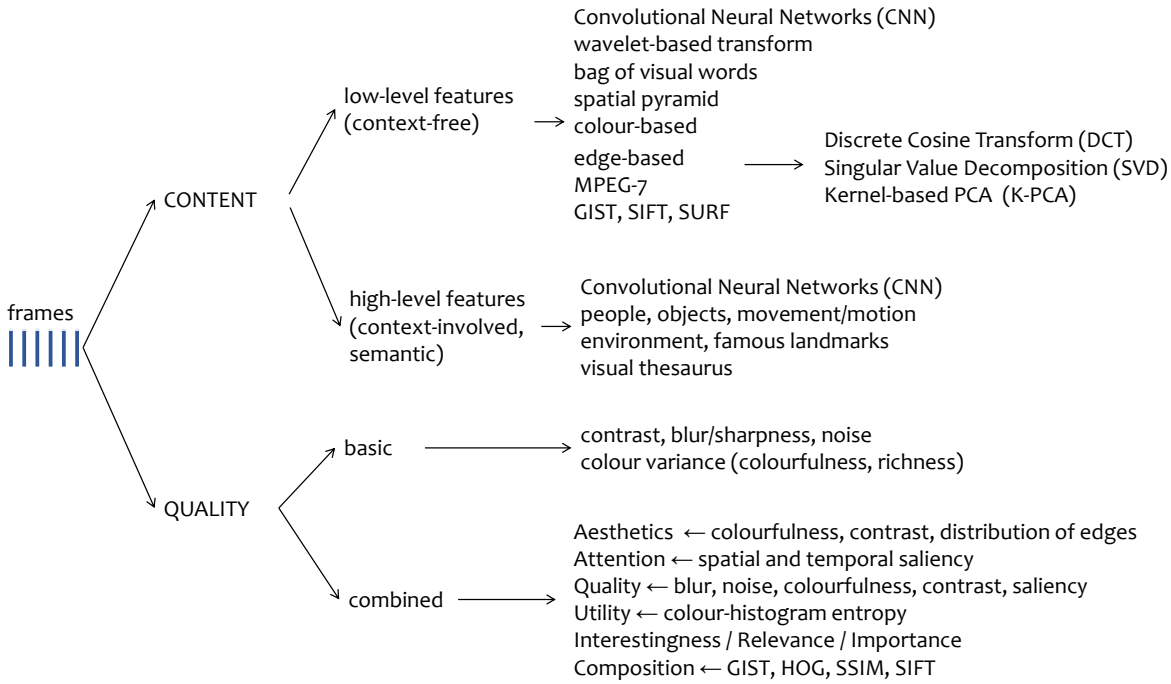

Figure 2: *A taxonomy of feature spaces (descriptors) used in video processing and summarisation.*

summarisaion. One of the most universal and useful group are the colour feature spaces, specifically RGB and HSV [ALT13,ALT12b, DLDD11, RS03, AA08, YX00].

For an online application, two factors must be considered when choosing a descriptor: (1) the ability of the chosen feature space to identify the meaningful attributes of the scene; (2) the computational cost of processing (the extraction process, and algorithm running time associated with the feature dimensionality). In order to select an appropriate feature space for an online algorithm, we collect a number of different features including those employed by non-egocentric online summarisation methods.

For our analyses, we selected the following feature spaces:

1. *RGB moments.* The RGB colour moments 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. Thus, each frame is represented by 54 features.
2. *HSV histograms.* This feature space is extracted by a quantisation of the HSV color space into a 256-dimensional histogram vector: (a) 32 bins for Hue, 4 bins for Saturation and 2 bins for Value ([32 4 2]). To increase speed, before extracting the HSV space the original image is resized to 1/64th of its original size.
3. *CENTRIST descriptor* CENsus TRansform hISTogram (CENTRIST) [WR11]. Census Transform is a nonparametric local transform which compares the intensity value of a pixel with its eight neighboring pixels. The binary results from the 8 comparisons are transformed in a decimal number between 0 and 255 (the order or arrangement does not matter as long as it is consistent across pixels and images). A histogram of these numbers is then generated with 256 bins, one for each Census intensity. The two end bins (corresponding to 0 and 255) are removed, leaving a 254-dimensional feature space. This feature space has been find particularly useful for recognising topological places or scene categories, especially in indoor environments. For extracting the CENTRIST descriptor, we used the MATLAB implementation from: `https://github.com/sometimesfood/spact-matlab`.
4. *GIST.* [OT01] The Gist descriptor is a low dimensional representation of the scene which is computed by convolving an image with 32 Gabor filter (4 scales and 8 orientations), producing 32 feature maps. Each feature map is divided into 4×4 regions and the average feature values of each region is set into the corresponding region. Subsequently, the 16 average values of 32 feature maps are concatenated resulting 512-dimensional descriptor.
5. *Colour Layout MPEG7.* [KY] The Colour Layout descriptor (CLD) represents the spatial distribution of colour in an image. An input RGB image is uniformly divided into 8×8 blocks. The average value of the pixel colours for each block is calculated, producing a 'tiny image'. The tiny image is converted into YCbCr colour space and then quantized into three sets of 64 DCT coefficients (total of 192 features).
6. and 7. *places205-AlexNet CNN and VGG CNN.* We included two high level feature descriptors extracted through deep learning neural networks. The 4096 deep features are extracted right before the classification (soft-max) layer of two pre-trained

Table 1: *Comparison of the average time of feature extraction for the ADL video (toy video obtained from video#8), and the main characteristics of the selected features.*

Descriptor	Size		Visual Information			Time(sec)	Dimensions	Used in
	resized	original	colour	scene	deep learning			
RGB moments		✓	✓			50	54	[MYK18]
HSV histogram [32 4 2]	✓		✓			30	256	[ALT12a]
CENTRIST		✓		✓		160	254	[MGW*15]
Gist	✓			✓		232	512	—
Color Layout MPEG7		✓	✓			519	192	[OLS*15]
places205-AlexNet CNN		✓		✓	✓	494	4096	—
VGG CNN		✓			✓	2377	4096	[AMT16]

Convolutional Neural Networks (CNNs), known as: VGGNet architecture [SZ14] available through the MatConvNet toolbox [VL15]; and Places205-AlexNet model [ZLX*14] using Caffe deep learning toolbox [JSD*14].

The feature spaces are detailed in Table 1

3. Experiment

The purpose of the experiment is to evaluate the feature spaces in regard to their suitability for online keyframe summarisation from egocentric video. Thus, we consider two aspects: ease of calculation of the feature space and the quality of the produced summary.

3.1. Data

For our experiment, we picked two egocentric videos from the Activity of Daily Living (ADL) dataset [PR12]. The ADL dataset was recorded using a chest-mounted GoPro camera which captures video at 30 frames per second at 1280×960 resolution. It consists of 20 videos of subjects performing their daily activities in the house. The first video is the video #1 from the dataset, consisting of 1,794 frames. The second video, which is called the 'toy video', is a selection of the initial 495 frames from video #8 of the same dataset. For this experiment, the selected videos are subsampled at rate one frame per second.

3.2. Extraction time

All experiments were carried out on a laptop, 2.20 GHz Intel Core *i*5 CPU, with 8GB RAM.

The first part of our analyses compares the processing time to extract the different features for the toy video. For each descriptor, we calculated the average time of extraction by repeating the process 20 times. The results are shown in Table 1. The extraction time for the simple colour spaces (RGB moments and HSV histograms) is shorter than the time for the other descriptors, whereas the popular VGG CNN has the longest extraction time.

3.3. Quality of the keyframe summary

The second part of our analyses compares the qualities of the summaries based on the different feature spaces.

3.3.1. The online summarisation algorithm

The online summarisation algorithm is sketched in Algorithm 1. The idea is to monitor the distances between consecutive frames (points in the chosen feature space $\mathbb{R}^L$) and detect transition frames between events in the video when this distance exceeds a threshold. Following the detection of an event boundary, the frame closest to the centroid of the current event (collection of points in $\mathbb{R}^L$) is selected and added to the summary. We start the algorithm with initial buffer B of a chosen cardinality $|B| = b$. Next we calculate the pairwise distances between consecutive frames in B. We opted for the control chart approach for monitoring the distance. Hence, we use the initial b frames in B to calculate the mean μ and the standard deviation σ of the distances in order to have a first reference value. Next, we start accumulating the incoming frames in the 'current' buffer B. Each subsequent frame which is deemed close enough to its predecessor, we add it to B and recalculate the reference value. Should we come across a frame that triggers the detector, we check whether the current size of the buffer is too small for the buffer to be perceived as an event. If so, we empty the buffer and start collecting frames anew. If, however, B is large enough, we have identified an event. The frame representing the event, k, is the one whose point in $\mathbb{R}^L$ is closest to the centroid of the points in B. Before adding the k to the summary, we make sure that the last stored keyframe, if it exists, is different enough from k. Otherwise, if the two consecutive keyframes are close, we might have misidentified an outlier as an event boundary. Therefore, we calculate the similarity between k and the latest stored keyframe in S. If the two frames are similar above a certain threshold (similarity does not have to be defined in the space of interest $\mathbb{R}^L$), we pool B with the the previous buffer, B_{last} (the previous, as well as current buffer is maintained in memory), and select the centroid keyframe k^* to add to the summary S. If k is not similar to its predecessor, we add k to S instead.

Algorithm 1: Online Video Summarisation

Data: Streaming video frames $f_1, \ldots, f_N$, $f_i \in \mathbb{R}^L$ (the chosen feature space); distance measure $d(.,.)$ in $\mathbb{R}^L$; minimum event size m; initial buffer size b; frame similarity measure $\tau(.,.)$ comparing HSV histograms; threshold value for keyframe similarity δ^*

Result: Set of keyframes S.

begin
- $B \leftarrow$ first b frames (initial buffer)
- $S \leftarrow \emptyset$
- Calculate the $b-1$ distances between all consecutive frames in B. Find the mean μ and the standard deviation σ of these distances.
- **for** $i = b+1 \ldots N$ **do**
  - **if** $d(f_{i-1}, f_i) < \mu + 3\sigma$ **then**
    - update μ and σ with f_i (same event)
    - $B \leftarrow B \cup \{f_i\}$ (store in the current buffer)
  - **else if** $|B| > m$ **then**
    - $k \leftarrow$ the frame closest to the centroid of B
    - **if** $|S| > 0$ **then**
      - $\delta \leftarrow \tau(k,$last stored keyframe$)$
      - **if** $\delta > \delta^*$ **then**
        - $k \leftarrow$ merge events B and B_{last} and select single representative
        - Remove the last keyframe from S.
    - $S \leftarrow S \cup k$
    - $B_{last} \leftarrow B$
    - $B \leftarrow f_i$ (re-initialise the buffer)
  - **else**
    - $B \leftarrow f_i$ (scrap the non-event)

return S

3.3.2. Performance measure

We chose the F-measure between the selected summary S and a given ground truth as a performance indicator. The ground truth for the data sets was created as follows: Each event in the video is distinguished by a number of terms. The frames in an event are labelled as informative/not informative based on whether they contain semantic information that is included in the relevant terms for this event. Consequently, any informative frame from the event can be considered ground truth for that event.

The F-value is calculated as the number of matched frames between the two summaries, divided by the average cardinality of the two sets. The higher the value, the better the quality of the summary.

3.3.3. Results

The F-values using the chosen feature spaces for video #1 from the ADL database are shown in Table 2.

Comparing the values in this table, it can be seen that the HSV histogram descriptor has the highest F-value, and the CNN descriptor, the lowest value. The difference between F-value for the RGB moments and the HSV histogram descriptor is not large. However, having fewer dimensions, the RGB moments space may have an advantage in the further processing compared to the HSV histogram space.

Table 2: *Comparison of the F-measure values among different chosen features for video* #1.

Descriptor	F-measure
HSV histogram [32 4 2]	0.78
RGB moments	0.76
CENTRIST	0.76
Color Layout MPEG7	0.36
Gist	0.21
places205-AlexNet CNN	0.21
VGG CNN	0.2

As can be seen from Tables 1 and 2, using complex descriptors such as places205-AlexNet [ZLX*14] would not always improve the performance thereby justifying its high computational cost. CENTRIST and Gist feature spaces gave better performance than CNN, and were also faster to extract.

4. Conclusions

Our experiments show that for egocentric videos, simple, colour-based descriptors offer a substantially more efficient and higher quality summary than the complex CNN features tested. Extending the study to additional videos is necessary to assess the robustness of descriptors across different content.

For the colour-based descriptors, the use of resized images does not appear to adversely affect the summary quality. Image compression is therefore an interesting area to explore for online video summarisation, with a potential for further gains in efficiency.

References

[AA08] ABD-ALMAGEED W.: 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)* (Oct. 2008), pp. 3200–3203. 1, 2

[ALT12a] ALMEIDA J., LEITE N. J., TORRES R. D. S.: Vison: Video summarization for online applications. *Pattern Recognition Letters 33*, 4 (Mar. 2012), 397–409. URL: https://doi.org/10.1016/j.patrec.2011.08.007. 3

[ALT12b] ALMEIDA J., LEITE N. J., TORRES R. S.: VISON: VIdeo Summarization for ONline applications. *Pattern Recognition Letters 33*, 4 (2012), 397–409. 2

[ALT13] ALMEIDA J., LEITE N. J., TORRES R. D. S.: Online video summarization on compressed domain. *Journal of Visual Communication and Image Representation 24*, 6 (Aug. 2013), 729–738. URL: http://dx.doi.org/10.1016/j.jvcir.2012.01.009. 1, 2

[AMT16] ANIRUDH R., MASROOR A., TURAGA P.: Diversity promoting online sampling for streaming video summarization. In *IEEE International Conference on Image Processing (ICIP2016)* (Sept. 2016), pp. 3329–3333. 1, 3

[Bam15] BAMBACH S.: A survey on recent advances of comp. vision algorithms for egocentric video. *arXiv:1501.02825* (2015). 1

[BDR17] BOLAÑOS M., DIMICCOLI M., RADEVA P.: Towards storytelling from visual lifelogging: An overview. *Journal of Transactions on Human-Machine Systems 47* (2017), 77–90. 1

[BMT*15] BOLAÑOS M., MESTRE R., TALAVERA E., GIRÓ I NIETO X., RADEVA P.: Visual summary of egocentric photostreams by representative keyframes. In *Proc. IEEE Int. Multimedia and Expo Workshops (ICME)* (2015), pp. 1–6. 1

[DLDD11] DE AVILA S. E. F., LOPES A. P. B., DA LUZ A., DE ALBUQUERQUE ARAÚJO A.: VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters 32*, 1 (2011), 56–68. 2

[EK17] ELHAMIFAR E., KALUZA M. C. D. P.: Online summarization via submodular and convex optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2017)* (Jul. 2017), pp. 1818–1826. 1

[JSD*14] JIA Y., SHELHAMER E., DONAHUE J., KARAYEV S., LONG J., GIRSHICK R., GUADARRAMA S., DARRELL T.: Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia* (New York, NY, USA, 2014), MM '14, ACM, pp. 675–678. URL: http://doi.acm.org/10.1145/2647868.2654889, doi:10.1145/2647868.2654889. 2

[KY] KASUTANI E., YAMADA A.: The mpeg-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. In *Proceedings 2001 International Conference on Image Processing (ICIP01), volume=1, pages=674–677, month=October 07-10, year=2001, organization=IEEE, location=Thessaloniki, Greece, Greece, doi=10.1109/ICIP.2001.959135.* 2

[MA08] MONEY A. G., AGIUS H. W.: Video summarization: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation 19*, 2 (2008), 121–143. 1

[MGW*15] MEI S., GUAN G., WANG Z., WAN S., HE M., FENG D. D.: Video summarization via minimum sparse reconstruction. *Pattern Recognition 48*, 2 (Feb. 2015), 522–533. 1, 3

[MTLT17] MOLINO A. G. D., TAN C., LIM J. H., TAN A. H.: Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems 47*, 1 (2017), 65–76. 1

[MYK18] MATTHEWS C. E., YOUSEFI P., KUNCHEVA L. I.: Using control charts for on-line video summarisation. In *SUBMITTED: British Machine Vision Conference (BMVC 2018)* (April 2018). 3

[OLS*15] OU S.-H., LEE C.-H., SOMAYAZULU V. S., CHEN Y.-K., CHIEN S.-Y.: On-line multi-view video summarization for wireless video sensor network. *IEEE Journal of Selected Topics in Signal Processing 9*, 1 (Feb. 2015), 165–179. 1, 3

[OT01] OLIVA. A., TORRALBA A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. Journal of Computer Vision 42*, 3 (2001), 145–175. 2

[PR12] PIRSIAVASH H., RAMANAN D.: Detecting activities of daily living in first-person camera views. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR12)* (June 16-21 2012), IEEE, pp. 2847–2854. doi:10.1109/CVPR.2012.6248010. 3

[RS03] RASHEED Z., SHAH M.: Scene detection in hollywood movies and tv shows. In *Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Jun. 2003), vol. 2, pp. 343–343. 1, 2

[SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014). 2

[TV07] TRUONG B. T., VENKATESH S.: Video abstraction. *ACM Transactions on Multimedia Computing, Communications, and Applications 3*, 1 (2007), 3–es. 1

[VL15] VEDALDI A., LENC K.: Matconvnet – convolutional neural networks for matlab. In *Proceeding of the ACM Int. Conf. on Multimedia* (2015). 2

[WR11] WU J., REHG J. M.: Centrist: A visual descriptor for scene categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 33*, 8 (2011), 1489–1501. doi:10.1109/TPAMI.2010.224. 2

[YX00] YIHONG G., XIN L.: Generating optimal video summaries. In *Proc. IEEE Int. Multimedia and Expo Workshops (ICME)* (2000), vol. 3, pp. 1559–1562. 2

[ZLX*14] ZHOU B., LAPEDRIZA A., XIAO J., TORRALBA A., OLIVA A.: Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems 27* (2014), Ghahramani Z., Welling M., Cortes C., Lawrence N. D., Weinberger K. Q., (Eds.), Curran Associates, Inc., pp. 487–495. 2, 4