Comparing Keyframe Summaries of Egocentric Videos: Closest-to-Centroid Baseline

Ludmila I. Kuncheva
School of Computer Science
Bangor University
Bangor, United Kingdom
l.i.kuncheva@bangor.ac.uk

Paria Yousefi
School of Computer Science
Bangor University
Bangor, United Kingdom
paria.yousefi@bangor.ac.uk

Jurandy Almeida
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo – UNIFESP
São Paulo, Brazil
jurandy.almeida@unifesp.br

***Abstract*—Evaluation of keyframe video summaries is a notoriously difficult problem. So far, there is no consensus on guidelines, protocols, benchmarks and baseline models. This study contributes in three ways: (1) We propose a new baseline model for creating a keyframe summary, called Closest-to-Centroid, and show that it is a better contestant compared to the two most popular baselines: uniform sampling and choosing the mid-event frame. (2) We also propose a method for matching the visual appearance of keyframes, suitable for comparing summaries of egocentric videos and lifelogging photostreams. (3) We examine 24 image feature spaces (different descriptors) including colour, texture, shape, motion and a feature space extracted by a pretrained convolutional neural network (CNN). Our results using the four egocentric videos in the UTE database favour low-level shape and colour feature spaces for use with CC.**

***Index Terms*—Video summarisation, Keyframe selection, Egocentric video, Image feature descriptors, Closest-to-Centroid baseline model, Keyframe evaluation protocol.**

I. Introduction

Keyframe summary of a video is a collection of frames which reflects the content of the video in a succinct and expressive way. One common problem faced by researchers is the evaluation of a keyframe summary [1]–[8]. At present, authors often develop a bespoke experimental set-up in which their proposed method for keyframe selection compares favourably to one or two alternative methods.

The methods for obtaining a keyframe summary vary dramatically depending on the type of the video. Egocentric videos and life-logging photo streams are especially difficult to summarise because of the large variability within the content of the units (events) [8]. This calls for tailor-made methods for summary evaluation. One component of the evaluation protocol is the choice of alternative methods to compare against. Typical choices for such baseline methods are Random (R), Uniform (U), and Mid-Event (ME). For R and U, the number of frames K must be fixed in advance. For R, K frames are randomly picked from the video regardless of their temporal position. For U, the video is split into K segments of equal length and the middle frame in each segment is taken for the summary. The Mid-Event summarisation method requires that the video is already split into temporally coherent units (events), either by an user or by an automatic method. The middle frame (time-wise) is chosen to summarise this event. These three baseline methods have been widely used (almost exclusively) as the rival methods in evaluating a proposed summary: Random (R) [3], [9], [10], Uniform (U) [3], [5], [10]–[12], and Mid-Event (ME) [12]–[15]. Arguably, these baselines are quite easy to beat. A new summarisation method is naturally expected to rate better in comparison to these baselines. However, an experiment confined only to R, U and ME still leaves open the question of how the new method compares to the state of the art.

Here we propose a new *baseline* summarisation method termed Closest-to-Centroid (CC) which is meant to serve as a competitor stronger than R, U and ME. The CC approach has been used in the past either as a baseline or as a part of the new method proposed within the respective study [1], [3], [4], [10], [16]–[21]. Here we develop CC into a baseline keyframe selection method by choosing among a large variety of feature descriptors, thereby ensuring that CC is a higher quality summary compared to U and ME (R is not taken forward because it is deemed to be the weakest baseline anyway). In order to evaluate the merit of the keyframe summaries we design a generic matching protocol.

The rest of the paper is organised as follows. Section II introduces the proposed baseline method. The feature spaces are discussed in Section III. Our experiment with the UTE egocentric video database [12][1] is presented in Section IV. Finally, Section V gives our conclusions.

II. Closest-to-Centroid baseline

The information required by the R and U baseline methods is only the number of frames in the video / photo stream. This is why R and U have been widely used in the evaluation parts of many studies. The ME method requires knowledge of the units to be represented in the summary (events, shots,

This work was done under project RPG-2015-188 funded by The Leverhulme Trust, UK. We also thank the financial support from the São Paulo Research Foundation - FAPESP (grant 2016/06441-7) and the Brazilian National Council for Scientific and Technological Development - CNPq (grant 423228/2016-1).

978-1-5386-1842-4/17/$31.00 ©2017 IEEE

[1]http://vision.cs.utexas.edu/projects/egocentric/

scenes, segments, etc.). Segmenting the video into such units is a difficult task in its own right, even more so for egocentric and life-logging data [8].

Our proposal requires a further assumption. The frames of the video must be described in some feature space. Let $V = \langle f_1, \ldots, f_N \rangle$ be the video data, where each frame is indexed by its time tag, and is represented by a feature vector an n-dimensional space, $\mathbf{x}(f_i) \in \mathbb{R}^n$. (To simplify notation, we will use just $\mathbf{x}_i$ to represent frame f_i) . Let $I_k \subset \{1, 2, \ldots, N\}$ be the index set of consecutive time tags identifying event k from the total of K events, $k = 1, \ldots, K$. The baseline model proposed here is to return the frame closest to the centroid for each event. We refer to the events as "clusters" although they may not form a conventional cluster structure in $\mathbb{R}^n$. Formally, the summary is the collection of ordered indices $J = \langle j_1, \ldots, j_K \rangle$ where

$$j_k = \arg\min_{m \in I_k} \{d(\mathbf{x}_m, \mathbf{c}_k)\}, \tag{1}$$

$d(.,.)$ is a chosen distance metric in $\mathbb{R}^n$, and

$$\mathbf{c}_k = \frac{1}{|I_k|} \sum_{j \in I_k} \mathbf{x}_j$$

is the centroid of cluster (event) k.

The CC approach has been widely used either as the sole selection method, as a component thereof, or as a baseline, sometimes under different names. For example, if d is the Euclidean distance, it can be easily shown that the *minimum distance* method of Bolaños et al. [10] is, in fact CC.

III. Feature spaces

A crucial component of any keyframe selection method is the chosen feature space. Following the literature, we consider two groups: features which are meant to describe the *content* of the frame, and features used to evaluate its *quality*. Note that the two groups are not completely non-intersecting; they likely share low-level features. Here we are interested in the former group.

The content type feature spaces can be further divided into low-level (context-free) and high-level (context-involved or semantic). Quite often, the original feature space is transformed further through Principal Component Analysis (PCA).

The boundary between low-level and high-level features is somewhat blurred as many feature extraction methods are designed with a view to enable detecting semantic content. A perfect example are feature spaces extracted through deep learning neural networks (e.g., Convolutional Neural Networks (CNN)). In some studies, CNN output, taken before the last fusion layer, is classed as low-level, while in others, as high-level. In any case, CNN is the leading feature extraction method for video summarisation [5], [10], [22], and therefore we include it in our experiments.

The more context-involved the feature space is, the less useful it is likely to be for a baseline method with wide applicability. This is why we chose for our study a wide variety of mostly low-level features as summarised in Table I.

The colour descriptors are as follows: *Auto Colour Correlogram* (ACC) [23], *Colour and Edge Directivity Descriptor* (CEDD) [24], *Colour Layout Descriptor* (CLD) [25], *Fuzzy Colour and Texture Histogram* (FCTH) [26], *Fuzzy Opponent Histogram* (FOH) [27], *GIST* [28], *HSV Colour Histogram* (HSV^{ch}), *Joint Composite Descriptor* (JCD) [29], *RGB Colour Histogram* (RGB^{ch}) [30], *RGB Colour Moments* (RGB^{cm}), *Scalable Colour Descriptor* (SCD) [25]. For encoding shape information, we use the *Pyramid of Histogram of Oriented Gradients* (PHOG) [31]. The descriptors for encoding texture properties are: *Edge Histogram Descriptor* (EHD) [25], *Gabor* features [32], *Local Binary Patterns* (LBP) [33], *Rotation Invariant Local Binary Patterns* (LBP^{riu2}) [33], *Tamura* features [34].

The HSV^{ch} features refer to a colour histogram computed only from the hue value (H) of the HSV colour space after its uniform quantization into 32 colour bins. The RGB^{cm} colour moment features were extracted as follows: each frame was divided uniformly into a 3-by-3 grid of blocks and then we computed the mean and the standard deviation for each block and each colour (9 blocks $\times$ 3 colour $\times$ 2 statistics = 54 features). For extracting the GIST features, we used the Lear's GIST implementation[2]. All the other descriptors were extracted using the LIRE library[3] [35]. In addition to such descriptors, we considered four other descriptors also provided in the LIRE library, named as *Basic Features* (BF), *Jpeg Coefficient Histogram* (JCH), *Joint Histogram* (JH), and *Luminance Layout Descriptor* (LLD).

Also, we evaluated a mid-level representation based on visual dictionaries, called *Fisher Vectors* (FV) [36], which encodes local features as visual words. To create the visual dictionary, local patches were extracted with a Hessian-affine detector and described by SIFT descriptors [37], which were reduced using Principal Component Analysis (PCA) and then used to create a codebook with 64 visual words learned by Gaussian Mixture Models (GMM). A global representation of a video frame is obtained by accumulating the residual vectors. The difference of each reduced SIFT descriptor and the mean vector of the Gaussian distribution assigned to each visual word was calculated. These differences were concatenated into a single feature vector, which was subsequently power-law normalised and then L_2-normalised. The GMM computation and FV encoding were performed using the Yael library[4] [38].

For the *Convolutional Neural Networks* (CNN) we used MatConvNet [39]. The 4096 deep features were extracted right before the classification (soft-max) layer, from the response of the Fully Connected layer (FC7) of the CNN. The runner-up

[2]The Lear's GIST implementation is available at: https://lear.inrialpes.fr/src/lear_gist-1.2.tgz (As of March 2017)

[3]The LIRE library is available at: http://www.lire-project.net (As of March 2017)

[4]The Yael library is available at: http://yael.gforge.inria.fr (As of March 2017)

in ILSVRC 2014, known as VGGNet architecture [40], was chosen to train the network. This network contains 16 hidden (Conv/FC) layers.

We also considered a spatio-temporal descriptor to encode motion information, known as *Histogram of Motion Patterns* (HMP) [41].

TABLE I
THE MAIN CHARACTERISTICS OF THE EVALUATED FEATURE REPRESENTATIONS.

Feature Type	Visual Information	Acronym	Size
Low-Level	Colour	1. ACC	1024
		2. CEDD	144
		3. CLD	118
		4. FCTH	192
		5. FOH	576
		6. GIST	960
		7. HSV^{ch}	32
		8. JCD	168
		9. JCH	192
		10. JH	576
		11. RGB^{ch}	512
		12. RGB^{cm}	54
		13. SCD	64
	Texture	14. BF	8
		15. EHD	80
		16. Gabor	60
		17. LBP	256
		18. LBP^{riu2}	36
		19. LLD	64
		20. Tamura	18
	Shape	21. PHOG	630
Mid-Level	Corners and edges	22. FV	4096
High-Level	People and objects	23. CNN	4096
Low-Level	Motion	24. HMP	6075

IV. AN EXPERIMENT WITH THE UTE EGOCENTRIC VIDEO DATABASE

The purpose of this experiment is to identify a feature representation among the chosen 24 representations in Tab. I where CC is markedly better than U and ME. In doing so, we also contribute a method for comparing keyframe summaries based on the visual appearance of the frames.

The assumptions in this experiments are

1) The video has been already segmented into temporally coherent events.
2) One frame per event is selected in the summary.
3) There is a ground truth of representative frames (one per event).

A. Data

The UTE dataset [12] contains 4 videos (each lasting about 3-4 hours) of subjects performing their daily activities such as driving, shopping, attending lectures and eating.[5] The data set is challenging because it contains frequent changes of the illumination and the camera position. The videos were recorded at 15 frames/second with 350 $\times$ 480 resolution per frame. We sub-sampled each video taking one frame per four seconds, thus reducing the number of frames as follows:

- P01 , 3464 frames, 14 events.
- P02 , 4566 frames, 19 events.
- P03 , 2696 frames, 10 events.
- P04 , 4446 frames, 16 events.

Each video was segmented into events using SR-clustering [42][6].

A ground truth summary was constructed for each video. A user picked a frame for each event so that the events are faithfully represented and still discernible within the video.

B. Matching procedure

Our matching procedure is intended to pair two frames *for the same event* with respect to their visual appearance. While there are many possibilities, we chose SURF features [43] on the grey image to match objects and shapes as done before [13], [15], and HSV histograms (following the protocol by De Avila et al. [44]) to match the colour distribution.

Let f_1 and f_2 be the frames being compared. Denote by p_1 and p_2 the number of SURF points of interest in the respective frames. Let m_1 be the number of matches found from f_1 to f_2, and m_2, the number of matches from f_2 to f_1. The matching score from the SURF features is taken to be

$$S_{\text{SURF}} = \frac{m_1 + m_2}{p_1 + p_2}.$$

The two frames are considered matching on SURF features if $S_{\text{SURF}} > \theta_{\text{SURF}}$, where $\theta_{\text{SURF}} \in [0, 1]$ is a threshold.

For the HSV feature space, a 32-bin histogram of the hue value was calculated for each frame. The bin counts were normalised so that the sum was 1 for each histogram. Let $B_j = \{b_{j,1}, \ldots, b_{j,32}\}$ be the normalised histogram for f_j, $j = 1, 2$. The L_1 distance was calculated by

$$D_H = \sum_{i=1}^{32} |b_{1,i} - b_{2,i}|.$$

The two frames are considered matching on HSV features if $D_H < \theta_H$, where $\theta_H \in [0, 2]$ is a threshold.

To ensure that the frames are a true visual match they must be a match on the objects/shapes (SURF) as well as colour (HSV). Because of this conservative rule, we pick threshold values which will allow for a fairly liberal match on each components: $\theta_{\text{SURF}} = 0.05$ and $\theta_H = 0.6$.

To illustrate the matching method, we show in Fig. 1 the results for matching the ground truth and the uniform, mid-event and CC (PHOG) summaries of video P03. The matched frames are highlighted in red.

[5] This benchmark dataset has been used as a sole experimental test bed in many studies on egocentric video summarisation.

[6] https://github.com/MarcBS/SR-Clustering

Finally, the match between the *summaries* can be calculated as the F-measure, which in this case reduces to the proportion of matches. For the examples in Fig. 1, $F = \frac{1}{10} = 0.1$ for U and ME, and $F = \frac{5}{10} = 0.5$ for CC with PHOG features.

C. Results

We identified the CC summary for each feature space, and quantified its proximity to the ground truth using the above matching procedure. Additionally, we prepared three alternative versions for each feature space. We applied PCA and retained components explaining respectively 95%, 90% and 80% of the variability of the data. The the CC summaries were obtained, and the F-measure was calculated for these additional feature spaces. The results are shown in Table II. The higher the values, the better the feature spaces. We have shown for comparison the F-measures for the two baseline methods we contrast CC against: the uniform summary (U) and the mid-event summary (ME). Ideally, all F-values for CC will be higher than those for U and ME.

The results show that many feature spaces lead to CC which matches the ground truth better than U or ME. The effect of PCA is not consistent. Sometimes the F measure increases with the transformation and retaining the fewer features, and sometimes the effect is the opposite, both for the same feature space and different videos (e.g., the Gabor feature space). To show the overall performance of the feature spaces, we averaged the F values across the videos and the 4 variants of each feature space (across the columns of the table). Figure 2 shows the averaged values for the CC baseline method for the 24 feature spaces. The U and ME baselines are represented by horizontal lines as they do not depend on the feature spaces.

With small exceptions, the feature spaces are suitable for the CC baseline as the F-values for CC are higher than those for U and ME. The best feature space in this experiment happens to be PHOG. This can be explained with the fact that the SURF features used as a part of the matching procedure also account for the shapes in the frames. The same argument can be put forward for HSVch. The highly acclaimed CNN feature space showed a modest improvement of CC over U and ME. Note that lower values of the F-measure do not mean that the respective feature space is flawed. The F-values give us grounds for recommending a particular feature space for the CC baseline against which "proper" keyframe selection methods should be compared. Based on the results of this experiment, we recommend 21. PHOG, 1. ACC, 15. EHD, 7. HSVch and 4. FCTH.

V. Conclusion

Here we address one of the most acute problems in video summarisation: automatic evaluation of keyframe summaries. We propose a baseline model, Closest-to-Centroid (CC) and advocate its use instead of the weaker baselines widely used thus far – the Uniform and the Mid-event selections. In addition, we propose an evaluation framework to compare summaries where each event is represented by a single keyframe.

The main limitations of CC and the matching procedure are as follows: the video must be already split into events; the matching procedure addresses only visual similarity between the frames.

Future experiments may refine the choice of a feature space for CC and the parameter values for the matching procedure. The CC can be applied to semantic feature spaces provided that those can be suitably quantified and equipped with a distance metric. To make the CC baseline even more competitive, an image quality component can be added to the closest-to-centroid criterion.

References

[1] B. T. Truong and S. Venkatesh, "Video abstraction," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 3, no. 1, pp. 3–es, 2007.

[2] M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, "STIMO : STIll and MOving Video Storyboard for the Web Scenario," *Multimedia Tools and Applications*, vol. 46, no. 1, pp. 47–69, 2010.

[3] A. Khosla, R. Hamid, C. J. Lin, and N. Sundaresan, "Large-Scale Video Summarization Using Web-Image Priors," *Proc. IEEE Conf. on Comp. Vision and Pattern Recognition (2013)*, pp. 2698–2705, 2013.

[4] G. L. Priya and S. Domnic, "Shot based keyframe extraction for ecological video indexing and retrieval," *Ecological Informatics*, vol. 23, pp. 107–117, 2014.

[5] A. Lidon, M. Bolaños, M. Dimiccoli, P. Radeva, M. Garolera, and X. G. i Nieto, "Semantic summarization of egocentric photo stream events," *arXiv:1511.00438*, 2015.

[6] G. Liu, X. Wen, W. Zheng, and P. He, "Shot boundary detection and keyframe extraction based on scale invariant feature transform," in *Proc. 8th IEEE/ACIS Int. Conf. on Comp. and Information Science (ICIS)*, 2009, pp. 1126–1130.

[7] A. G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *Journal of Visual Communication and Image Representation*, vol. 19, no. 2, pp. 121–143, 2008.

[8] A. G. D. Molino, C. Tan, J. H. Lim, and A. H. Tan, "Summarization of egocentric videos: A comprehensive survey," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 1, pp. 65–76, 2017.

[9] S. Chowdhury, P. McParlane, M. S. Ferdous, and J. Jose, "My day in review: Visually summarising noisy lifelog data," in *Proc. 5th ACM Int. Conf. on Multimedia Retrieval*, 2015, pp. 607–610.

[10] M. Bolaños, R. Mestre, E. Talavera, X. Giró i Nieto, and P. Radeva, "Visual summary of egocentric photostreams by representative keyframes," in *Proc. IEEE Int. Multimedia and Expo Workshops*, 2015, pp. 1–6.

[11] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Proc. IEEE Comp. Society Conf. on Comp. Vision and Pattern Recognition*, 2013, pp. 2714–2721.

[12] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," *Proc. IEEE Comp. Society Conf. on Comp. Vision and Pattern Recognition*, pp. 1346–1353, 2012.

[13] P. Ratsamee, Y. Maei, A. Jinda-Apiraksa, M. Horade, K. Kamiyama, M. Kojima, and T. Arai, "Keyframe selection framework based on visual and excitement features for lifelog image sequences," *Int. Journal of Social Robotics*, vol. 7, no. 5, pp. 859–874, 2015.

[14] A. R. Doherty, D. Byrne, A. F. Smeaton, G. J. F. Jones, and M. Hughes, "Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs," in *Proc. 2008 Int. Conf. on Content-based Image and Video Retrieval CIVR*, 2008, pp. 259–268.

[15] A. Jinda-Apiraksa, J. Machajdik, and R. Sablatnig, "A Keyframe Selection of Lifelog Image Sequences," *Proceedings of MVA 2013 IAPR Int. Conf. on Machine Vision Applications*, pp. 33–36, 2013.

[16] A. Hanjalic and H. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 8, pp. 1280–1289, Dec 1999.

[17] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proc. Int. Conf. on Image Processing (ICIP)*, vol. 1, 1998, pp. 866–870.

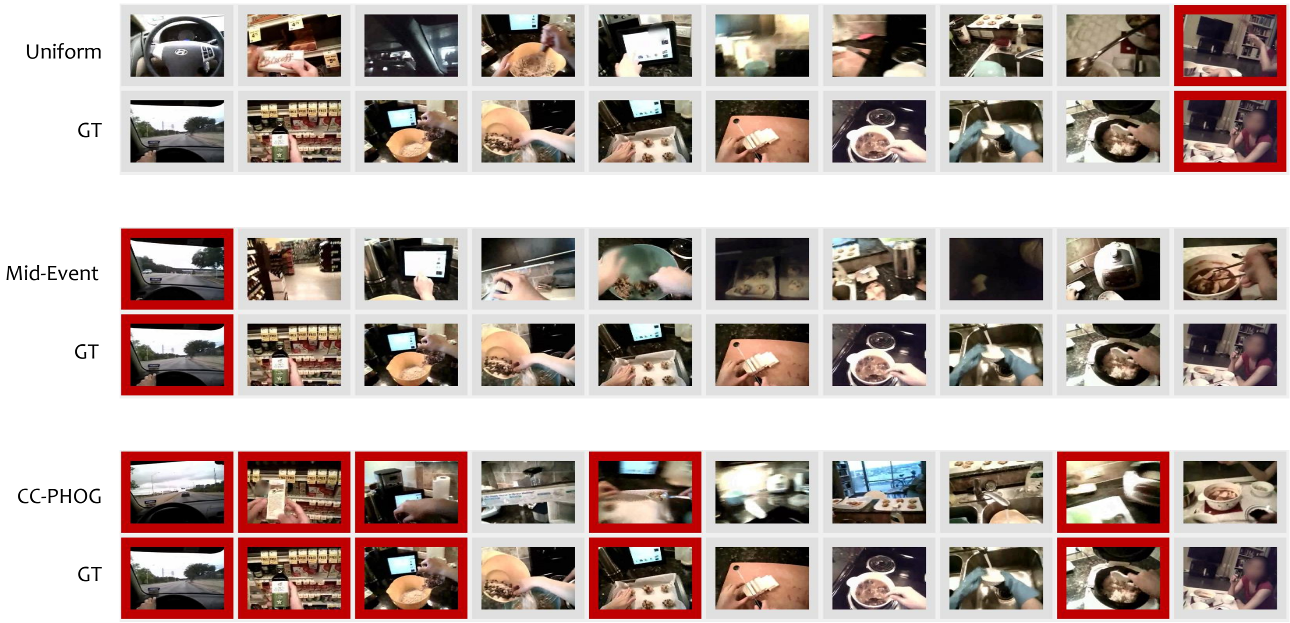

Fig. 1. Illustration of the results from the matching procedure on the 10 events for video P03.

TABLE II
F-MEASURE (IN%) FOR THE 4 VIDEOS FOR THE U, ME AND CC SUMMARIES WITH RESPECT TO THE GROUND TRUTH.

	Features	P01				P02				P03				P04			
		Org	P95	P90	P80	Org	P95	P90	P80	Org	P95	P90	P80	Org	P95	P90	P80
1	ACC	36	36	36	36	21	11	11	11	10	10	10	10	50	44	50	44
2	CEDD	14	14	14	36	11	11	11	11	10	10	10	10	50	44	50	44
3	CLD	7	7	7	7	16	11	11	5	0	0	0	0	19	25	19	19
4	FCTH	14	14	14	21	5	5	16	16	40	30	20	10	38	50	44	44
5	FOH	14	14	14	14	0	5	0	11	0	0	10	10	38	38	38	31
6	GIST	21	14	21	7	0	0	0	0	10	10	0	10	31	31	31	31
7	HSVch	29	29	21	29	11	11	16	16	30	40	10	20	38	31	38	31
8	JCD	21	21	21	21	16	21	21	21	0	0	0	20	56	44	44	44
9	JCH	21	21	7	0	5	11	11	11	20	20	30	0	56	50	38	31
10	JH	14	7	14	14	16	16	16	16	10	10	10	0	56	38	25	31
11	RGBch	29	29	21	21	5	0	0	0	10	10	10	10	25	19	25	19
12	RGBcm	14	14	14	7	16	21	21	16	10	10	20	20	50	31	38	31
13	SCD	21	14	21	7	5	5	5	21	0	0	20	10	38	44	25	44
14	BF	21	14	14	14	16	16	21	21	10	10	10	10	38	44	44	44
15	EHD	29	29	21	21	16	16	16	16	20	20	10	10	50	44	44	50
16	Gabor	21	21	21	21	5	5	0	0	20	10	10	10	19	25	25	25
17	LBP	14	21	29	29	11	16	16	11	10	10	10	10	44	38	44	38
18	LBPriu2	21	14	14	14	32	21	5	5	10	0	0	10	38	19	19	31
19	LLD	14	7	7	21	11	11	16	16	0	0	0	0	19	13	19	25
20	Tamura	29	14	36	21	5	5	11	11	0	0	10	10	38	25	25	25
21	PHOG	29	29	29	29	11	16	5	0	50	50	40	40	38	44	44	50
22	FV	29	21	21	21	0	16	16	16	20	20	20	20	44	31	31	31
23	CNN	7	7	7	21	0	0	5	5	20	20	20	0	38	38	44	38
24	HMP	21	14	14	0	0	0	5	11	20	20	10	10	44	31	38	38
	Uniform	7				16				10				13			
	Mid-event	7				11				10				25			

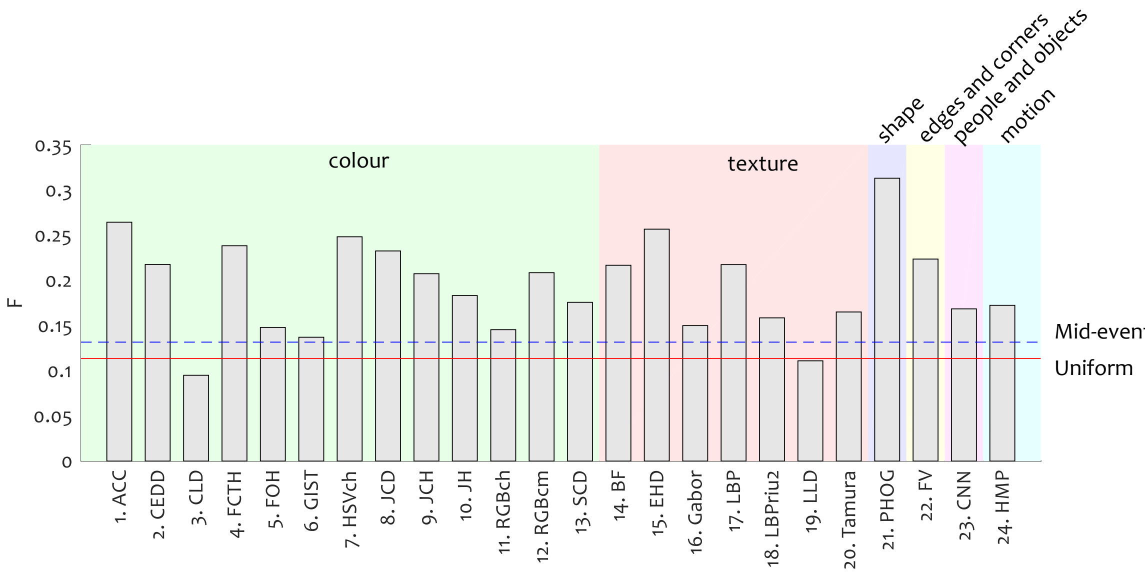

Fig. 2. Averaged F measure comparing for the proposed baseline method (CC) and the ground truth for the 24 feature spaces. The F-values for U and ME are also shown for comparison.

[18] X. Zhu, X. Wu, J. Fan, A. K. Elmagarmid, and W. G. Aref, “Exploring video content structure for hierarchical summarization,” *Multimedia Systems*, vol. 10, pp. 98–115, 2004.

[19] E. Spyrou, G. Tolias, P. Mylonas, and Y. Avrithis, “Concept detection and keyframe extraction using a visual thesaurus,” *Multimedia Tools and Applications*, vol. 41, no. 3, pp. 337–373, 2009.

[20] X. D. Yu, L. Wang, Q. Tian, and P. Xue, “Multilevel video representation with application to keyframe extraction,” in *Proc. 10th IEEE Int. Multimedia Modelling Conf.*, 2004, pp. 117–123.

[21] L. Herranz and J. M. Martínez, “An efficient summarization algorithm based on clustering and bitstream extraction,” *Proc. 2009 IEEE Int. Multimedia and Expo Workshops (ICME)*, pp. 654–657, 2009.

[22] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, and N. Yokoya, “Video summarization using deep semantic features,” *arXiv:1609.08758*, 2016.

[23] J. Huang, R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, “Image indexing using color correlograms,” in *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR’97)*, 1997, pp. 762–768.

[24] S. A. Chatzichristofis and Y. S. Boutalis, “FCTH: fuzzy color and texture histogram - A low level feature for accurate image retrieval,” in *Proc. Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS’08)*, 2008, pp. 191–196.

[25] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, “Color and texture descriptors,” *IEEE Transactions on Circuits Systems and Video Technology*, vol. 11, no. 6, pp. 703–715, 2001.

[26] S. A. Chatzichristofis and Y. S. Boutalis, “CEDD: color and edge directivity descriptor: A compact descriptor for image indexing and retrieval,” in *Proc. Int. Conf. on Computer Vision Systems (ICVS’08)*, 2008, pp. 312–322.

[27] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, “Evaluating color descriptors for object and scene recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[28] A. Oliva. and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *Int. Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[29] S. A. Chatzichristofis, Y. S. Boutalis, and M. Lux, “Selection of the proper compact composite descriptor for improving content based image retrieval,” in *Proc. IASTED Int. Congress on on Signal Processing, Pattern Recognition and Applications (SPPRA’09)*, 2009, pp. 134–140.

[30] M. J. Swain and B. H. Ballard, “Color indexing,” *Int. Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[31] A. Bosch, A. Zisserman, and X. Muñoz, “Representing shape with a spatial pyramid kernel,” in *Proc. ACM Int. Conf. on Image and Video Retrieval (CIVR’07)*, 2007, pp. 401–408.

[32] B. S. Manjunath and W.-Y. Ma, “Texture features for browsing and retrieval of image data,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, 1996.

[33] T. Ojala, M. Pietikäinen, and T. Mäenpää, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[34] H. Tamura, S. Mori, and T. Yamawaki, “Textural features corresponding to visual perception,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, no. 6, pp. 460–473, 1978.

[35] M. Lux and O. Marques, *Visual Information Retrieval Using Java and LIRE*, ser. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2013.

[36] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, “Aggregating local image descriptors into compact codes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1704–1716, 2012.

[37] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[38] M. Douze and H. Jégou, “The yael library,” in *Proc. ACM Int. Conf. on Multimedia (ACM-MM’14)*, 2014, pp. 687–690.

[39] A. Vedaldi and K. Lenc, “Matconvnet – convolutional neural networks for matlab,” in *Proc. ACM Int. Conf. on Multimedia (ACM-MM’15)*, 2015, pp. 689–692.

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

[41] J. Almeida, N. J. Leite, and R. S. Torres, “Comparison of video sequences with histograms of motion patterns,” in *Proc. IEEE Int. Conf. on Image Processing (ICIP’11)*, 2011, pp. 3673–3676.

[42] M. Dimiccoli, E. Talavera, S. G. Nikolov, and P. Radeva, “SR-Clustering: Semantic regularized clustering for egocentric photo streams segmentation,” *arXiv:1512.07143v1*, 2015.

[43] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, “Speeded-up robust features (SURF),” *Comp. Vision and Image Understanding*, vol. 110, no. 3, pp. 346 – 359, 2008.

[44] 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*, vol. 32, no. 1, pp. 56–68, 2011.