KEY FRAME EXTRACTION FROM CONSUMER VIDEOS USING EPITOME

$C.T.Dang$[1],*Student Member IEEE,*$M.Kumar$[2], *Senior Member IEEE, and* $H.Radha$[1],*Fellow IEEE*

[1]Electrical and Computer Department, Michigan State University
dangchin,radha@egr.msu.edu
[2]Eastman Kodak Company, Rochester, Newyork
mrityunjay.kumar@kodak.com

ABSTRACT

Key frame extraction algorithms select a subset of the most informative frames from videos. Key frame extraction finds applications in several broad areas of video processing research such as video summarization, video indexing, and prints from video. In this paper, an image epitome [1][2] based method to extract key frames from unstructured consumer videos is presented. In the proposed approach, we exploit image epitome to measure dissimilarity between frames of the input video. The dissimilarity scores are further analyzed using a min-max approach to extract the desired number of key frames from the input video. The proposed approach does not require shot(s) detection, segmentation, or semantic understanding. A comparison of the results obtained by this method with the ground truth agreed by multiple judges clearly indicates the feasibility of the proposed approach.

Index Terms— Key frame extraction, image epitome, consumer videos, video analysis.

1. INTRODUCTION

Key frame extraction algorithms select a subset of the most informative frames from videos [35]. Key frame extraction finds applications in several broad areas of video processing research such as video summarization, creating chapter titles in DVDs, video indexing, and prints from video. Conventional key frame extraction approaches can be loosely divided into two groups: (i) shot-based, and (ii) segment-based. In shot-based key frame extraction, the shots of the original video are first detected, and then one or more key frames are extracted from each shot [6]. In segment-based key frame extraction approaches, a video is segmented into higher-level video components, where each segment or component could be a scene, an event, a set of one or more shots, or even the entire video sequence. Representative frame(s) from each segment are then selected as the key frames [4].

Existing key frame approaches, shot-based as well as segment-based, are usually suitable for structured videos such as news and sports videos. However, they are sub-optimal for consumer videos as these videos are typically captured in an unconstrained environment and record extremely diverse contents. Moreover, consumer videos often lack a pre-imposed structure, which makes it even more challenging to detect shots or segment such videos for key frame extraction [7][8].

Because of the enormous number of pixels present in videos, as a rule of thumb, first, feature vectors can be computed from video frames and then these features can be processed to extract key frames. Due to diverse nature and lack of structure in consumer videos, it is challenging to identify (set of) features that represent the entire consumer video space adequately. Intuitively, features that preserve low-level information present in video frames (texture, edge, etc.) can be very useful in determining key frames from consumer videos. In addition, efficient image processing/computer vision models can be to process feature vectors to extract key frames. Typically, clustering-based models are used to extract key frames from features [9-10]. In clustering-based model, frames having similar features are grouped together and one or more frames from each cluster are selected to generate the desired number of key frames. However, clustering based models suffer from inter- and intra-cluster variability. Furthermore, clustering-based approaches are very similar to shot- and segment-based approaches and their limitations for dealing with consumer videos have already been discussed above.

In this paper, we proposed an image epitome [1][2] based method to extract key frames from unstructured consumer videos. In the proposed approach, we use image epitome as a feature vector and apply an information divergence based distance measure on the feature vector to measure dissimilarity between frames of the input video. Epitome is significantly smaller as compared to the original image and yet preserves important visual information (texture, edge, color, etc.) of the input image. Furthermore, epitome has been shown to be shift and scale invariance and effective in terms of modeling the spatial structure [11]. The dissimilarity scores are further analyzed using a min-max approach to extract the desired number of key frames from the input video.

This paper is organized as follows. Section 2 reviews image epitome representation. In Section 3, the proposed

key frame extraction algorithm is described, while Section 4 presents benchmarking results comparing to ground truth data. Finally, concluding remarks are given in Section 5.

2. IMAGE EPITOME REVIEW

An image epitome E of size $p \times q$ is a condensed version of the corresponding input image X of size $M \times N$ where $p \ll M$, $q \ll N$[1][2]. Let $Z = \{Z_k\}_1^P$ be the patch level representation of X, i.e., is the set of all possible patches from X . The epitome (E) corresponding to X is estimated using Z and represents the salient visual contents of X effectively. More specifically, epitome E is derived by searching for a set of patches in E that corresponds to the set Z based on Gaussian probability distribution. The patches in E are defined by a set of mapping, $T = \{T_k\}_1^P$, which shows a displacement between two patches X and E respectively. Assuming distribution at each epitome location to be Gaussian, the conditional probability for mapping patches in epitome to set of patches in an image is defined as:

$$p\left(Z_k|T_k,E\right) = \prod_{i \in S_k} N(z_{i,k}; \mu_{T_k(i)}, \phi_{T_k(i)}) \quad (1)$$

$$p\left(\{Z_k\}_1^P \mid \{T_k\}_1^P, E\right) = \prod_{k=1}^{p} p(Z_k|T_k,E) \quad (2)$$

In which $\left\{\mu_{T_k(i)}, \phi_{T_k(i)}\right\}$, mean and variance of a Gaussian distribution, are parameters stored in one epitome coordinate that is mapped to pixel i in Z_k. Solving the maximum likelihood problem leads to expectation maximization algorithm. In the expectation-step, given the current epitome E and $Z = \{Z_k\}_1^P$, the set of mappings is specified by optimizing (2), searching for every allowed correspondences. The multiple patch mappings allow one pixel in epitome to be mapped onto numerous pixels in the larger image. In the maximization-step, given the new set of mappings, mean and variance at each location, e.g. location u, are calculated [1]:

$$\mu_u = \frac{\sum_i \sum_k [u=T_k(i)] z_{i,k}}{\sum_i \sum_k [u=T_k(i)]} \quad (3)$$

$$\phi_u = \frac{\sum_i \sum_k [u=T_k(i)] (z_{i,k}-\mu_u)^2}{\sum_i \sum_k [u=T_k(i)]} \quad (4)$$

$$[P] = \begin{cases} 1 & if Pis true \\ 0 & otherwise \end{cases} \quad (5)$$

3. IMAGE EPITOME DISSIMILARITY MEASUREMENT AND MIN-MAX ALGORITHM

3.1. Image epitome dissimilarity measurement

Measuring perceptual or visual dissimilarity between images is an important research area and finds its applications in many image processing and computer vision problems

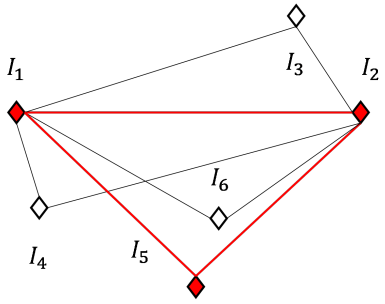


Fig.1.These red points (I_1, I_2 and I_5) will be selected using the min-max algorithm

including key frame extraction. Selecting feature(s) or descriptor(s) that describe visual content of images effectively is crucial for image dissimilarity measurement. Motivated by this, we use epitome representation of an image as feature to compute image dissimilarity since epitome is significantly smaller as compared to the original image and yet preserves important visual information (texture, edge, color, etc.) of the input image. Furthermore, epitome has been shown to be shift and scale invariant and effective in terms of modelling the spatial structure [11]. The detailed description of epitome based image dissimilarity is provided next.

Let E_i be the lexicographical representation of the epitome (i.e., $E_i \in R^{m \times 1}$) corresponding to the i^{th} image I_i. Therefore, the distribution function (represented as f_i) of E_i can be expressed as a linear combination of m Gaussians as given belows:

$$f_i = \frac{1}{m} \sum_{k=1}^{m} N(\mu_k^i, \phi_k^i) \quad (6)$$

Where $N\left(\mu_k^i, \phi_k^i\right)$ is the distribution of the k^{th} element of E_i. The proposed dissimilarity of two images, denoted as I_i and I_j, respectively, is computed as follows:

$$D(I_i/I_j) = D(I_j/I_i) = \frac{1}{2}\left(\int f_i \log \frac{f_i}{f_j} + f_j \log \frac{f_j}{f_i}\right) \quad (7)$$

Note that the proposed dissimilarity measure in eq. (7) exploits well-known Kullback-Leibler divergence [12]. In case of two Gaussian mixtures, there is no closed form solution for eq. (7); hence approximate solution based on unscented transform or Gaussian elements matching is typically employed in practice to solve eq. (7)[13]. In the proposed approach, we use unscented transform-based approach to solve eq. (7) because of the potential overlap between epitomes caused due to temporal correlation present between video frames. The unscented transformation attempts to calculate the statistics of a random variable which undergoes a non-linear transformation. Given a d-dimensional normal random variable x, distribution function $f(x) \sim N(\mu, \Sigma)$ and an arbitrary non-linear function $h(x) : R^d \rightarrow R$, the approximated expectation of function $h(x)$ over $f(x)$ is given by:

$$\int f(x)h(x)dx \approx \frac{1}{2d}\sum_{k=1}^{2d} h(x_k) \quad (8)$$

the set of $2d$ "sigma points" x_k is chosen as follows:

$$x_k = \mu + (\sqrt{d\Sigma})_k \qquad k = 1 \dots d \quad (9)$$

$$x_{k+d} = \mu - \left(\sqrt{d\Sigma}\right)_k \qquad k = 1...d \quad (10)$$

In the case of epitome distribution, two Gaussian mixtures, $f = \sum_{i=1}^{n} \alpha_i N(\mu_{1,i}; \Sigma_{1,i})$ and $g = \sum_{j=1}^{m} \beta_j N(\mu_{2,j}; \Sigma_{2,j})$, we have:

$$d = 3; \qquad k = 1...d \quad (11)$$

$$\int^f \log g \approx \frac{1}{2d}\sum_{i=1}^{n} \alpha_i \sum_{k=1}^{2d} \log g(x_{i,k}) \quad (12)$$

$$x_{i,k} = \mu_{1,i} + (\sqrt{d\Sigma_{1,i}})_k \quad (13)$$

$$x_{i,k+d} = \mu_{1,i} - (\sqrt{d\Sigma_{1,i}})_k \quad (14)$$

3.2. The min-max algorithm for key frame extraction

We now apply the proposed dissimilarity distance into key frame extraction problem. The dissimilarity between every pair of frames in video sequence is measured. Traditionally, clustering methods could be used to segment these frames into clusters and then select the center frame (or medoid) in each cluster as a key frame. However, we believe that clustering is not the best method in the key frame selection problem. Two important criteria that a good set of key frames need to satisfy: i) covering the entire content of video and ii) reducing redundancies between any pair of key frames. Medoids in two clusters might not be two points with a highest distance. It implies that redundancy between them could be higher than two other frames in these clusters.

The min-max approach [14][15] represents a powerful optimization tool, which is used in many disciplines (game theory, statistics, decision theory, etc.), under two opposite constraints. We bring this approach into our algorithm with the two aforementioned constraints (i-ii). Fig. 1 shows an example of selecting 3 key frames based on the min-max algorithm. The first step selects two points with the largest distance (I_1 and I_2) to assure that they contain the highest amount of information. Under the next step, I_5 is selected as the best points because I_3 and I_4 are to close (high redundancies) to I_1 and I_2 respectively. I_6 is clearly not as good as I_5. Table 1 outlines the detailed algorithm.

Algorithm 1 The min-max algorithm

Inputs:the number of key frames T, the total number of frames N, video sequence $V = \{I_1, I_2, \dots, I_N\}$.
Outputs: Key frames sequence $S = \{f_1, f_2, \dots, f_T\}$
Initialization: $S = \emptyset$
Do for i from 1 to $(N-1)$
Do for j from $(i+1)$ to N
Compute $D\,(I_i/I_j)$ using epitome-based dissimilarity
End
End
Detect two first key frames:
$\{f_1, f_2\} = \underset{I_i, I_j}{Arg\max}\, D\,(I_i/I_j)$
Update:$S = \{f_1, f_2\}$,$n = 2$,$V/S = \{g_1, g_2, \dots, g_{N-n}\}$;
V/S be the remaining set of frames without key frames.
Repeat
Do for i from 1 to $(N-n)$
$a_i = \min\{D(g_i/f_k), k = 1, 2, \dots, n\}$
$imax = \underset{i}{\text{Arg}\max}\, a_i$
End
Update:$n = n+1$;$f_n = g_{imax}$;$S = S \cup \{f_n\}$;
$V/S = \{g_1, g_2, \dots, g_{N-n}\}$
Until $n = T+1$
Return $S = \{f_1, f_2, \dots, f_T\}$

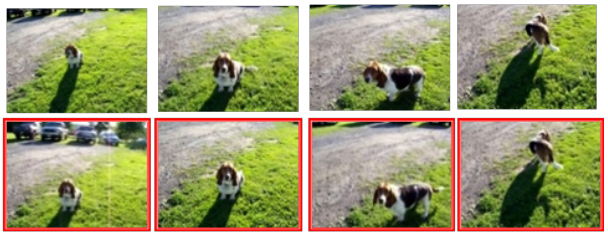
4. EXPERIMENTAL RESULTS

The results of the proposed key frame extraction algorithm are presented in this section. A database having 100 video clips selected from a large database of over 3000 video clips [7] was used to validate the proposed approach. These 100 video clips were captured using Kodak EasyShare C360 and V550 zoom digital cameras, with frame size 640×480 and frame rates ranging from 24 to 30 fps. These video clips were intended to capture a wide range of events typically encountered in consumer video clips such as outdoor activities, natural sceneries, trips, sports, weddings, etc.

The proposed method was compared with the ground truth agreed by multiple human judges [7]. To establish the ground truth, three human judges were asked to independently browse the video clips and provide the key frames. Photographers who actually captured the videos were not selected as the judges. The key frames estimated by the three judges were reviewed in a group session with a fourth judge (arbitrator) to derive final key frames for each of the 100 video clips. The number of key frames was determined by the human judges based on the representativeness and quality of the corresponding video clips.

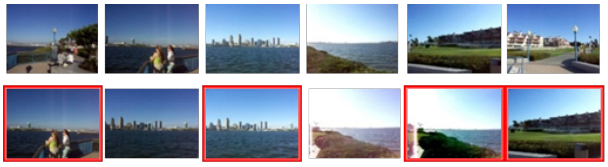
Due to page constraints, only three sets of results are presented in this paper. For each video clip in the database, the number of key frames determined by the judges was considered as the ground truth number of key frames for the corresponding video. Therefore, the proposed approach was executed to generate exactly the same number of key frames as those determined by the judges. Also, for simplicity, input video frames were rescaled to size 160×120. Furthermore, epitome and patch sizes considered in this experiment were 10×10, and 4×4, respectively. Comparisons of the proposed approach (second row) with the ground truth determined by multiple judges (first row) are provided in Fig. 2(a)-(c). The

input video corresponding to the key frames in Fig. 2(a) was 15.75 sec long and was captured at 24 fps. The video had a significant amount of perspective change. The ground truth



(a) *Happydog*

(b) *BusTour*



(c) *SkyLinefromOverlook*

Fig.2.Key Frames from ground truth (1^{st} row) and the proposed approach (2^{nd} row)

number of key frames was 4. Visually the proposed approach appears to be similar to the ground truth. Furthermore, the proposed approach did not require any video segmentation or semantic understanding.

The videos shown in Fig. 2(b) and (c) were 22.63 and 23.38 sec long, respectively, and were captured at 24 fps. Both videos were taken outdoors with a significant amount of change in perspective and brightness. The ground truth number of key frames was 5 and 6, respectively. For video in Fig. 2(b), the proposed approach missed the second key frame partially and the fourth key frame completely. Quantitatively speaking, the proposed approach extracted 3.5 out of 5 key frames correctly. Similarly, for video in Fig. 2(c), the proposed approach extracted 4 out of 6 key frames correctly as it completely missed the first and the sixth key frames.

5. CONCLUSTIONS AND FUTURE WORK

In this paper, an image epitome based framework for extracting key frame from consumer video is proposed. In the proposed approach, image epitome in conjunction with Kullback-Leibler divergence is used to measure dissimilarity between frames of the input video. The dissimilarity scores are further analyzed using a min-max approach to extract the desired number of key frames from the input video. In contrast to shot- and segment-based key frame extraction algorithms, the proposed approach does not require shot(s) detection, segmentation, or semantic understanding. Extensive experimental results clearly indicate the feasibility of the proposed approach. Future work will focus on validating the proposed approach on larger scale video datasets. Quantitative evaluation of the quality of the extracted key frames will also be carried out in the future.

6. REFERENCES

[1] N. Jojic, B. Frey, and A. Kannan, Epitomic analysis of appearance and shape, In *Intl. Conf. on Computer Vision (ICCV)*, 2003.

[2] V. Cheung, B. Frey, and N. Jojic, Video epitomes, In *Proc. CVPR*, 2005.

[3] B. T. Truong and S. Venkatesh, Video abstraction: a systematic review and classification, in *ACM Trans*. Multimedia Comput. Commun. Appl., vol. 3, no. 1, pp. 3es, Feb. 2007.

[4] Z. Rasheed and M. Shah, Detection and representation of scenes in videos, *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1097-1105, Dec. 2005

[5] N. Dimitrova, T. McGee, and H. Elenbaas, Video keyframe extraction and filtering: a keyframe is not a keyframe to everyone, in *Proc. CIKM*, Mar. 1997, pp. 113-120.

[6] S. Uchihashi and J. Foote, Summarizing video using a shot importance measure and a frame-packing algorithm, in *IEEE ICASSP*, 1999, vol. 6, pp. 3041-3044.

[7] J. Luo, C. Papin, and K. Costello, Toward extracting semantically meaningful key frames from personal video clips: From humans to computers, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 2, pp. 289-301, Feb. 2009.

[8] K. Costello and J. Luo, First-and third-party ground truth for key frame extraction from consumer video clips, in *Proc. SPIE* 6492, 64921N (2007).

[9] Y. Zhuang, Y. Rui, T. S. Huang and S. Mehrotra, Adaptive key frame extraction using unsupervised clustering, In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, 1998.

[10] Vasileios Chasanis, Aristidis Likas and Nikolaos Galatsanos, Video rushes summarization using spectral clustering and sequence alignment, Proceedings of the 2nd *ACM TRECVid* Video Summarization Workshop, 2008.

[11] Kai Ni, Anitha Kannan, Antonio Criminisi, and John Winn, Epitomic Location Recognition, *IEEE Trans. On Pattern Analysis And Machine Intelligence*, Vol. 6, no. 1, Jan. 2007

[12] S. Kullback, Information theory and statistics, *John Wiley and Sons*, NY 1959.

[13] Goldberger, Shiri Gordon, and Hayit Greenspan, "An efficient image similarity measure based on approximations of KL-divergence between two gaussian mixtures," in *Proceedings of ICCV* 2003, Nice, October 2003, vol. 1, pp. 487-493.

[14] J. von Neumann, Zur Theorie der Gesellschaftsphiele, Math. Ann. 100 (1928) pp. 295-3

[15] Jafarpour S, Cevher V, Schapire R.E, A game theoretic approach to expander-based compressive sensing, 2011 *IEEE International Symposium on Information Theory Proceedings (ISIT)*, August 2011, page(s):464-468.