World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:12, No:2, 2018

Key Frame Based Video Summarization via Dependency Optimization

Janya Sainui

***Abstract*—**As a rapid growth of digital videos and data communications, video summarization that provides a shorter version of the video for fast video browsing and retrieval is necessary. Key frame extraction is one of the mechanisms to generate video summary. In general, the extracted key frames should both represent the entire video content and contain minimum redundancy. However, most of the existing approaches heuristically select key frames; hence, the selected key frames may not be the most different frames and/or not cover the entire content of a video. In this paper, we propose a method of video summarization which provides the reasonable objective functions for selecting key frames. In particular, we apply a statistical dependency measure called quadratic mutual informaion as our objective functions for maximizing the coverage of the entire video content as well as minimizing the redundancy among selected key frames. The proposed key frame extraction algorithm finds key frames as an optimization problem. Through experiments, we demonstrate the success of the proposed video summarization approach that produces video summary with better coverage of the entire video content while less redundancy among key frames comparing to the state-of-the-art approaches.

***Keywords*—**Video summarization, key frame extraction, dependency measure, quadratic mutual information, optimization.

I. Introduction

DUE to the advances of multimedia technologies and networking, the number of digital videos is increasing rapidly. Consequently, there is a need to manage videos in an efficient and effective way in order to provide users a quickly video browsing and retrieval, as well as reduce storage. One of the most evolving research areas, which play an important role in this regard, is video summarization, as it provides a concise representation of the video content [1], [2]. Video summaries are generally performed by two different approaches: static and dynamic video summaries [3], [4]. Static video summary contains a collection of a small but meaningful number of silent frames known as key frames, while dynamic video summary contains a collection of important short video clips. Both approaches have their own advantages; however, in this paper, we focus on the static video summarization which is to extract a set of key frames from a video.

Two important properties of key frames are that (1) key frames should represent the entire content of a video and (2) each key frame should be as much different from each other as possible [5]–[10]. Typically, key frame selection algorithms assume that a video has been segmented into shots, and then a small number of key frames are extracted from each shot. A naive approach to key frame selection is to simply use the first frame of each shot [11]. However, the first frame is normally not stable and does not necessarily capture the entire visual content of a video shot. To improve the coverage, one may additionally select the middle and the last frames of a video shot. Otherwise, key frames may be selected by clustering- or sequential-based approaches. Clustering-based approach [4], [8], [12]–[14] groups the similar frames of a video shot into m clusters, and then chooses typically one frame that is closest to cluster centroid as a key frame. This approach requires a number of clusters or some thresholds to control the density of each cluster. However, the quality of selected key frames heavily depends on the required parameters. In practice, if the number of given clusters is not fit to data, the selected key frames are not necessary the most m different frames. Sequential-based approach [6], [9], [15], [22] measures the difference between the current frame and the last frame (or the last key frame), and then compares with a predefined threshold; if the difference exceeds the threshold, the current/middle frame will be selected as a new key frame. Nevertheless, selecting only the first frame or the middle frame from each segment is heuristic in nature. Moreover, this approach heavily depends on the predefined threshold that is not easy to tune in practice, and hence only one key frame may not be enough to represent the visual content of each video segment.

As mentioned above, the existing approaches lack the objective functions for both minimizing the redundancy among key frames and maximizing the coverage of the entire video content. In this paper, we propose an information theoretic based key frame selection approach for video summarization, which selects key frames by using a statistical independence measure as our objective functions for minimizing the information shared among key frames as well as maximizing the coverage of the visual content of an original video. Particularly, we employ *quadratic mutual information* (QMI) [23] to select m frames so that QMI over m key frames is minimized, while QMI between the set of selected key frames and the set of sampling frames is maximized. By this way, the selected key frames will be as much different from each other as possible, at the same time, cover the entire content of the original video. We evaluate the proposed video summarization on videos from the Open Video Project database and also on videos from YouTube. The experimental results demonstrate the capacity of the proposed method to produce video summaries that outperforms the competing approaches. The main contributions of this paper are a new proposal of an efficient and effective key frame based video summarization which has the ability to select key frames from

J. Sainui is with the Department of Computer Science, Faculty of Science, Prince of Songkla University, Hat Yai, 90112, Thailand (e-mail: janya.s@psu.ac.th).

arbitrary videos, the proposed key frame extraction algorithm finds key frames as an optimization problem, and the objective functions for minimizing the redundancy among key frames as well as maximizing the coverage of the entire video content are presented.

The rest of this paper is organized as follows. In Section II, related work in static video summarization is reviewed. The proposed key frame based video summarization is described in Section III. Experimental results and comparison with some of existing approaches are given in Section IV. In Section V, conclusion is provided.

II. Related Works

Many existing methods developed so far require some prior information about the content of video frames based on high-level sematic features such as motions, activities, and objects [16]–[21]. Although this approach was demonstrated to work well, it is applicable only when prior knowledge on motions, activities, objects, etc. When such prior knowledge is not available, methods based on low-level index feature are more useful. By this approach, the features like colors and textures of frames are considered to compute the difference between two frames, and then clustering-based or sequential-based algorithms can be applied to select key frames. In this paper, we focus on the low-level index feature so that it can be applied to arbitrary videos. Some of the existing methods based on low-level index feature found in the literature are reviewed as follows.

Zhuang et al. [12] presented a key frame selection algorithm based on unsupervised clustering. This method uses the color histogram in HSV color space to evaluate the similarity between frames, and requires a threshold to control the density of clustering. Before a new frame is classified into a certain cluster, the similarities between the current frame and the centroid of the existing clusters are computed. These values are compared with a given threshold, and then the current frame will be added into the closest cluster. If the current frame is not closed enough to the existing clusters, a new cluster is then formed. The key frame selection is considered only to the clusters that are larger than the average size of all clusters, then a frame which is closest to the considered cluster centroid is selected as a key frame.

Mundur et at. [13] proposed a clustering-based technique for producing video summarization. They used the Delaunay Triangulation (DT) for clustering video frames. The method starts by sampling the frames from the input video sequence. The color histogram in HSV color space is extracted from each of sampling frames to form a matrix, then Principal Component Analysis (PCA) is applied to reduce the dimension of the matrix. After that, DT is performed on the lower dimension matrix. Finally, for each cluster, a frame that is nearest its cluster center is chosen as a key frame.

Furini et at. [14] presented STIll and MOving Video Storyboard (STIMO), a video summarization technique designed to produce on-the-fly video storyboard, which allows user to select the storyboard length and the maximum time to wait to get the storyboard. The method is based on clustering algorithm where the histogram in the HSV color space is extracted to represent each video frame. This method first computes the pairwise distance of consecutive frames and compares with a given threshold to obtain the number of clusters. Next, the Furthest-Point-First (FPF) algorithm with the Generalized Jaccard Distance (GJD) is applied to cluster the video frames. After the key frames are extracted, some meaningless frames are removed from the produced summary.

Ejaz et al. [9] introduced a technique for key frame extraction based on sequential approach. The method starts by sampling the frames from the input video sequence. Three adaptive frame different measures based on the correlation of RGB color space, color histogram of HSV color space, and moments of inertia are combined to compute the difference between the last key frame and the current frame. If the difference exceeds given thresholds, the current frame will be selected as a new key frame. Finally, the meaningless frames (e.g., totally back/white frames, faded frames) are eliminated. After generation of the set of key frames, the redundancy is further reduced by removing those key frames which are very similar to each other.

Almeida et al. [15] presented VIdeo Summarization for ON line application (VISON) that operates directly in the compressed domain and allows user interaction. For each frame of an input sequence, the histogram in the HSV color space of DC images is extracted. They adopted the Zero-mean Normalized Cross Correlation (ZNCC) as the distance function. The ZNCC between consecutive frames is compared with some thresholds to produce subsequences of similar frames, and then the middle frame from each subsequence is chosen as a key frame. Finally, the selected frames are filtered in order to avoid possible redundancy or meaningless frame in the video summary.

As reviewed above, most of the existing approaches select key frames in a heuristic way. For example, the clustering-based approach selects the frames that are closest to cluster centers, the sequential-based approach selects the first or the middle frame. If the parameters required by these approaches are not suitable, then the selected key frames may not be proper key frames. Especially for a video with containing a lot of short shots, we found that the existing approaches lost many representative frames. To address this issue, in this paper, we propose a novel video summarization technique which extracts key frames by minimizing the information shared among key frames as well as maximizing the coverage of the entire content of video resulting in the improved quality of video summaries comparing to the competing approaches.

III. Proposed Method

The goal of our proposed approach is to automatically select a maximum number of key frames with minimum redundancy from a video. In other words, we expect that the set of selected key frames will preserve all of the information of the video content, while each key frame is as much different from each other as possible.

The framework of our proposed approach is illustrated in Fig. 1. Given an input video, the sequence of frames

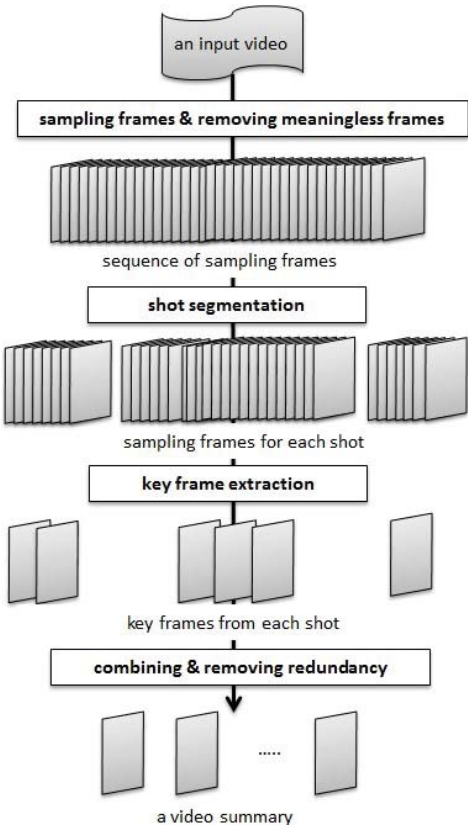


Fig. 1 Framework of proposed approach

is extracted, and then the number of frames is reduced by presampling approach and removing the meaningless frames. Next step is to segment the sequence of sampling frames into shots. After that, key frame extraction is performed for each shot to select a suitable number of key frames. Finally, the selected key frames from each shot are merged to produce a video summary. In the following subsections, each of steps is explained in more detail.

A. Preprocessing

As a video consists of a sequence of frames that contain a lot of redundancies, it is not necessary to perform key frame extraction on all of the video frames in practice. Generally, the frames are pre-sampled from a video by selecting one out of fixed length frames. In this work, we select one out of 30 frames, which give a sample of 1 frame-per-second (fps) for a video with 30 fps. In addition, meaningless frames (e.g., the dark frames due to fade-in/fade-out effects) are ignored if the standard deviation of the brightness of a frame is very low. This step leads to reduce the number of frames that will be processed afterwards.

B. Frame Different Measures

Accurately evaluating the similarities among video frames is important for key frame extraction. In this paper, we apply a statistical independence criterion called *Quadratic mutual information* (QMI) as similarity measure. The color histogram difference is also incorporated to ensure the differences between frames. In this subsection, we describe these two criteria that will be used through this paper for evaluating the differences among video frames.

1) Quadratic Mutual Information (QMI): QMI [23] is a variant of mutual information (MI) based on L_2 distance which is more robust against outliers than MI. Unlike correlation, QMI allows to capture higher-order correlation for more than two variables simultaneously. Thus, it is more reliable than the square error and the cross correlation.

Let $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(m)})^\top$ be an m-dimensional random variables defined on $\mathcal{X} \subset \mathbb{R}^m$. QMI for $\boldsymbol{x}$ is defined as

$$\mathrm{QMI}(x^{(1)}, \ldots, x^{(m)}) := \int \big(p(\boldsymbol{x}) - q(\boldsymbol{x})\big)^2 \mathrm{d}\boldsymbol{x}, \quad (1)$$

where $p(\boldsymbol{x})$ denotes the joint probability density of $x^{(1)}, \ldots, x^{(m)}$:

$$p(\boldsymbol{x}) = p(x^{(1)}, \ldots, x^{(m)})$$

and $q(\boldsymbol{x})$ denotes the product of marginal densities $\{p_k(x^{(k)})\}_{k=1}^m$:

$$q(\boldsymbol{x}) = \prod_{k=1}^{m} p_k(x^{(k)}).$$

Here, we assume that we are given a set of samples

$$\{\boldsymbol{x}_i \mid \boldsymbol{x}_i = (x_i^{(1)}, \ldots, x_i^{(m)})\}_{i=1}^n,$$

which are independently drawn from a joint probability distribution with density $p(\boldsymbol{x})$. We then use least-squares QMI (LSQMI) detailed in [24], [25] as our QMI estimator ($\widehat{\mathrm{QMI}}$) for capturing the dependency among frames; the high value indicates the high similarity between two frames and vise versa.

2) Histogram Frame Different Measure: Color histograms have been very popular for selecting key frames because it is simple and robust against small change in camera motions.

Let $H^{(\boldsymbol{x})}$ and $H^{(\boldsymbol{x}')}$ be the color histogram corresponding to frames $\boldsymbol{x}$ and $\boldsymbol{x}'$, respectively. To capture the histogram difference between frames $\boldsymbol{x}$ and $\boldsymbol{x}'$, we use the histogram intersection [26] and it is defined as:

$$d_H(\boldsymbol{x}, \boldsymbol{x}') := 1 - \frac{1}{P \times Q} \sum_{b=1}^{B} \min(H^{(\boldsymbol{x})}(b), H^{(\boldsymbol{x}')}(b)), \quad (2)$$

where $P \times Q$ is the size of a frame and B denotes the number of bins. The range of this value is $[0, 1]$. The higher value indicates the higher difference between two frames. On the other hand, the small value indicates that two frames are similar in color content.

C. Shot Segmentation

Before performing key frame extraction, we first segment a video (i.e., sequence of sampling frames) into shots. Note that each shot may have more than one key frame. To do so, let n be the number of sample size (e.g., the number of pixels in each frame) and N be the number of sampling frames. For $i = 1, \ldots, n$ and $k = 1, \ldots, N$, let $\boldsymbol{x}_i^{(k)} \in \mathbb{R}^d$ be a

Algorithm 1 shot segmentation algorithm

Input: Feature vectors of video frames $V = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$ and a threshold, τ.
Output: k shots, $\{V'^{(1)}, \ldots, V'^{(k)}\}$.

1: $k \leftarrow 1$
2: **for** $i = 2, \ldots, N$ **do**
3: compute d_{QMIH} (4) between $\boldsymbol{x}^{(i-1)}$ and $\boldsymbol{x}^{(i)}$
4: **if** $(d_{QMIH} > \tau)$ **then**
5: $k \leftarrow k + 1$
6: $V'^{(k)} \leftarrow \boldsymbol{x}^{(i)}$
7: **else**
8: $V'^{(k)} \leftarrow \{V'^{(k)} \cup \boldsymbol{x}^{(i)}\}$
9: **end if**
10: **end for**

d-dimensional feature vector of the i-th sample (i.e., the i-th pixel) at the k-th frame. Let V be the set of d-dimensional feature vectors of sampling frames:

$$V = \{\boldsymbol{x}_i^{(1)}, \ldots, \boldsymbol{x}_i^{(N)} | \boldsymbol{x}_i^{(k)} \in \mathbb{R}^d\}_{i=1}^n.$$

As described in Algorithm 1, a video represented by V is segmented into k shots, i.e., $\{V'^{(1)}, \ldots, V'^{(k)}\}$. Following this algorithm, the similarity between consecutive frames is computed, and shot changed is defined if its distance is higher than a given threshold τ.

In order to determine the significant change between frames, we linearly combine $\widehat{\mathrm{QMI}}$ with the color histogram difference. However, the $\widehat{\mathrm{QMI}}$ is not bounded; its range is from 0 to ∞, so it is difficult for comparison. We here normalize it and define the distance between frames $\boldsymbol{x}$ and $\boldsymbol{x}'$ based on $\widehat{\mathrm{QMI}}$ as

$$d_{QMI}(\boldsymbol{x}, \boldsymbol{x}') := 1 - \frac{\widehat{\mathrm{QMI}}(\boldsymbol{x}', \boldsymbol{x})}{\max\left(\widehat{\mathrm{QMI}}(\boldsymbol{x}', \boldsymbol{x}'), \widehat{\mathrm{QMI}}(\boldsymbol{x}, \boldsymbol{x})\right)}. \quad (3)$$

We assume that the values of $\widehat{\mathrm{QMI}}(\boldsymbol{x}', \boldsymbol{x}')$ and $\widehat{\mathrm{QMI}}(\boldsymbol{x}, \boldsymbol{x})$ are larger than that of $\widehat{\mathrm{QMI}}(\boldsymbol{x}', \boldsymbol{x})$ as they are the relationship between itself. Two distance functions d_H (2) and d_{QMI} (3) are linearly combined as

$$d_{QMIH} := w d_{QMI} + (1 - w) d_H, \quad (4)$$

where w is the weight for controlling the balance between the two criteria; though this paper, we set w to 0.5 so that the two criteria are equally important. Through this paper, we use d_{QMIH} for considering the differences between two frames.

D. Key Frame Extraction

Here, for each shot represented by a subsequence of sampling frames,

$$V' = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N')}\},$$

Our goal is to select m frames,

$$S = \{f^{(1)}, \ldots, f^{(m)}\},$$

from V' with minimum redundancy as well as maximum the coverage of the entire content of that shot as key frames.

Algorithm 2 key frames extraction algorithm

Input: Feature vectors of each subsequence of video frames $V' = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N')}\}$
Output: Selected key frames $S = \{f^{(1)}, \ldots, f^{(m)}\}$.

1: **for** $i = 1, \ldots, N$ **do**
2: compute $\widehat{\mathrm{QMI}}(\boldsymbol{x}^{(i)}, V')$
3: **end for**
4: $f^{(1)} \leftarrow \mathrm{argmax}_{\boldsymbol{x}^{(i)} \in V'} \widehat{\mathrm{QMI}}(\boldsymbol{x}^{(i)}, V')$
5: $S \leftarrow \{f^{(1)}\}$
6: remove frames that are similar to $f^{(1)}$ from V'
7: **repeat**
8: **for** each remaining $\boldsymbol{x}^{(i)}$ in V' **do**
9: compute $\widehat{\mathrm{QMI}}(S \cup \boldsymbol{x}^{(i)})$
10: **end for**
11: $f' \leftarrow \mathrm{argmin}_{\boldsymbol{x}^{(i)}} \widehat{\mathrm{QMI}}(S \cup \boldsymbol{x}^{(i)})$
12: **if** $\widehat{\mathrm{QMI}}(S \cup f', V')$ is maximized **then**
13: $S \leftarrow \{S \cup f'\}$
14: remove frames that are similar to f' from V'
15: **end if**
16: **until** $\widehat{\mathrm{QMI}}(S, V')$ is maximum, or V' is empty.

We propose the key frame extraction algorithm described in Algorithm 2. Following our algorithm, the first key frame, $f^{(1)}$, is selected by finding a maximizer of $\widehat{\mathrm{QMI}}$ between $f^{(1)}$ and V'. By this way, the first key frame is the frame that is the most coverage of the visual content of that shot. In line 6, the frames related to the first key frame are removed from V', as they are not necessary to be considered anymore. Hence, the number of frames to be processed later is reduced. Next, a candidate key frame f' that minimizes $\widehat{\mathrm{QMI}}$ between $f^{(1)}$ and f' is selected. If $\widehat{\mathrm{QMI}}$ between $\{f^{(1)}, f'\}$ and V' is maximized, f' is selected as a new key frame. Then, in line 14, the frames related to f' are removed from V'. The algorithm is iteratively for finding $f^{(1)}, \ldots, f^{(m)}$ with minimal $\widehat{\mathrm{QMI}}$, where it terminates if $\widehat{\mathrm{QMI}}$ between S and V' is maximum, or V' is empty. At the end, the set of m key frames is obtained. Notice that, in lines 6 and 14, two frames are regarded as being similar if the value of d_{QMIH} (4) between them is less than 0.6.

E. Postprocessing

At the final process, the key frames selected from each shot are merged to produce a summary for a video. However, the summary may still contain redundancies because similar content may exist in several segments. To do this, each key frame in the summary is compared with every other key frames based on d_H (2); the frames having 80% similarities (almost having the same visual content) with any other frames in the summary are removed.

IV. Experiments and Results

A. Setup

We compared the proposed method with DT [13], STIMO [14], and VISON [15] including the video summaries from the

Open Video Project website (OV) [27] . The key frames of VISON can be seen at [29], [30], while those of DT, STIMO, OV, and the ground truth key frames are available at [28]. For our approach, RGB pixel intensities with size 160×120 were extracted for calculating $\widehat{\text{QMI}}$, and the color histogram was set to 32 bins; 16 bins for hue component and 8 bins for each of the saturation and intensity components. The thresholds τ in Algorithm 1 was set at 0.5.

We used two datasets provided by [8] for evaluation, as the ground truth key frames and the results of selected key frames obtained from some existing approaches are available. The first dataset is collected from the Open Video Project database. All videos are in the MPEG-1 format (30 fps, 352×240 pixels). The selected videos are distributed among several genres (documentary, education, ephemeral, history, and lecture) and their duration varies from 1 to 3 mins. The videos in the second dataset are obtained from YouTube. They comprise news, sports, commercials, tv-shows, and home videos with durations varying from 1 to 10 mins.

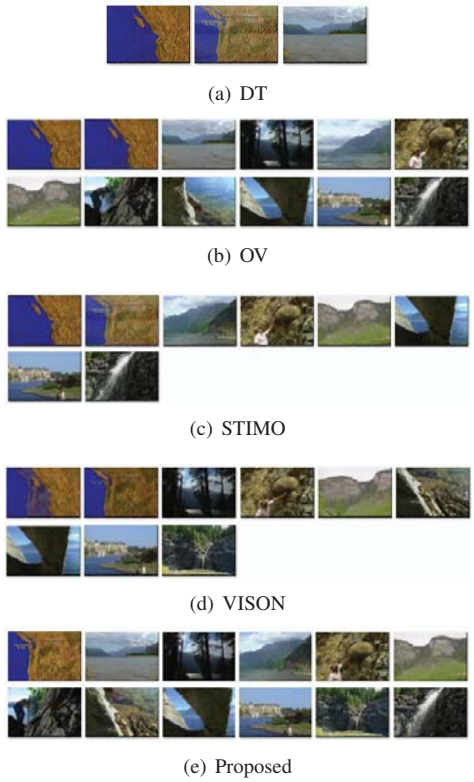
B. Illustrative Examples

The goal of key frame based video summary is to extract a maximum number of key frames while the redundancy of information in these key frames is minimal. Here, we illustrate the video summaries obtained from different approaches using two videos from the Open Video Project database and two videos from the YouTube, in order to show you that our proposed method finds key frames with the highest quality of summaries.

The first example is the summaries for the *Exotic Terrane, segment 08* video from the Open Video Project database. Its duration is 1.21 minutes. This video is not complex; it contains camera motions like panning and zooming, where the object movement in the video is a little bit. Fig. 2 shows the summaries obtained from different approaches. The results indicate that DT approach is the worst for this video, as it lost a lot of information. On the other hand, our proposed method outperforms others. More specifically, our method produced the summary with the maximum number of key frames, where each of them is different from each other. STIMO and VISON produced summaries that are better than DT, but worse than OV and our proposed method. The summary obtained by OV is also well; however, the redundant key frame is contained.

The second example is the summaries for *A New Horizon, segment 03* video from the Open Video Project database. This video is quite long (3.29 mins), and contains a lot of short shots. The results are shown in Fig. 3, confirming that our proposed method produced the summary with higher quality than other methods. As you can see, other methods, especially DT and OV, lost many informative frames, whereas our method tends to keep all visual content of the video.

Fig. 4 is the third example of video summaries for a news video from YouTube. The duration of this video is 1.49 minutes. The content changes in this video are clear, and visual content of each shot is quite different. So both our method and VISON perform well. In other words, all visual content of the video are preserved by both methods.



(a) DT (b) OV (c) STIMO (d) VISON (e) Proposed

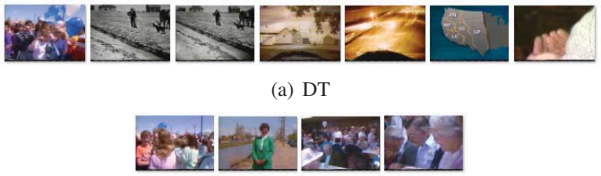
Fig. 2 Video summaries from different approaches for the *Exotic Terrane, segment 08* video

The last example is the summaries for a sport (football) video from YouTube. The duration of this video is 6.21 minutes, and its content is more complex than the second example. It is also difficult to judge that which frames are key frames, as the the objects and camera motions in the video are always moved. However, as results shown in Fig. 5, you can see that our method preserves the information of the video as well, whereas VISON lost many representative key frames.

As examples above, these can confirm the performance of the proposed method over other competing approaches. More specifically, our proposed approach can extract key frames from both easy and complex videos with less redundancy of key frames but high coverage of the entire content of a video. Thanks to LSQMI whichs provides several advantages for evaluating the differences/similarities among video frames.

C. Quantitative Evaluation

1) Evaluation Criteria: In the literature, to objectively evaluate the quality of selected key frames, the most straightforward way is to see whether a selected key frame matches a ground truth key frame. Note that a ground truth key frame can be matched at most one selected key frame. The numbers of matched and non-matched key frame are used for evaluation. Here, we use the standard measures: Precision, Recall, and F1-score, for evaluation, and they are defined as follows:



(a) DT

(b) OV

(c) STIMO

(d) VISON

(e) Proposed

Fig. 3 Video summaries from different approaches for *A New Horizon, segment 03* video

(a) VISON

(b) Proposed

Fig. 4 Video summaries from different approaches for a news video

$$\text{Precision} := \frac{n_{\text{mAS}}}{n_{\text{AS}}},$$
$$\text{Recall} := \frac{n_{\text{mAS}}}{n_{\text{US}}},$$
$$\text{F1-score} := 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

where n_{mAS} is the number of matching key frames, n_{AS} is the number of selected key frames from automatic selection, and n_{US} is the number of key frames from user selection (ground truth). We also use the compression ratio (CR) [20], which indicates how well a set of selected key frames represents the entire video. It be computed by the number of video frames that are related to selected key frames divided by the number of total frames of the original video as

$$\text{CR} := \frac{|\bigcup_{f^{(k)} \in S} C_k|}{|V|}.$$

In our evaluation, $C_k = \{\boldsymbol{x}^{(i)} \in V | d_{\text{QMIH}}(f^{(k)}, \boldsymbol{x}^{(i)}) < 0.6\}$, $V = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$ is the set of sampling frames from the original video, $S = \{f^{(1)}, \ldots, f^{(m)}\}$ is the set of selected key frames.

(a) VISON

(b) Proposed

Fig. 5 Video summaries from different approaches for a sport video

2) Results for the Open Video Project Dataset: We evaluated the quality of key frames selection for 40 videos from the open video database. Each video has 5 ground truth sets created by 5 different users. In other words, 200 ground truths were used for evaluation. The results of Precision, Recall, F1-score, and CR are summaried in Table I, showing that the proposed method overall outperforms other methods. More specifically, our method is the best in term of Recall and CR. These values tell you that the proposed method lost fewer informative frames than compared methods. For Precision and F1-score, our method is better than or comparable to other methods.

3) Results for the YouTube Dataset: We selected 30 videos from YouTube. Following the same rules as the previous dataset, each video has 5 ground truth sets created by 5 different users. In other words, in this dataset, 150 ground truth sets were used for evaluation. We compared our proposed approach with VISON, where its summaries can be seen at [30]. All the videos, the ground truth key frames are available at [28]. The evaluation results are summarized in Table II. As same as the results from the open video project dataset, the proposed method is better than VISON in term of Recall and CR, and is comparable in term of Precision and F1-score.

V. Conclusion

In this paper, we propose a new technique for video summarization based on key frame extraction. Our proposed

TABLE I
AVERAGES (AND STANDARD ERRORS IN THE BRACKETS) OF DIFFERENT CRITERIA OVER 40 VIDEOS FOR OPEN VIDEO PROJECT DATABASE. THE BEST AND COMPARABLE METHODS BY THE T-TEST AT THE SIGNIFICANCE LEVEL 5% ARE DESCRIBED IN BOLDFACE

Criterion	DT	OV	STIMO	VISON	Proposed
Precision	**0.72 (0.015)**	0.67 (0.015)	0.60 (0.012)	**0.72 (0.014)**	**0.71 (0.014)**
Recall	0.55 (0.018)	0.74 (0.017)	0.75 (0.015)	0.85 (0.011)	**0.90 (0.009)**
F1-score	0.60 (0.016)	0.66 (0.016)	0.64 (0.010)	**0.75 (0.010)**	**0.77 (0.009)**
CR	0.65 (0.04)	0.79 (0.032)	0.83 (0.023)	0.93 (0.01)	**0.96 (0.008)**

TABLE II
AVERAGES (AND STANDARD ERRORS IN THE BRACKETS) OF DIFFERENT CRITERIA OVER 30 VIDEOS FOR YOUTUBE DATABASE. THE BEST AND COMPARABLE METHODS BY THE T-TEST AT THE SIGNIFICANCE LEVEL 5% ARE DESCRIBED IN BOLDFACE

Criterion	VISON	Proposed
Precision	**0.053 (0.016)**	**0.56 (0.018)**
Recall	0.71 (0.017)	**0.77 (0.013)**
F-score	**0.58 (0.015)**	**0.62 (0.014)**
CR	0.90 (0.019)	**0.94 (0.009)**

approach first segments a video stream into shots. This step not only reduces time complexity for next step, but also improves the quality of the final set of video summaries. Next, for each shot, one or more frame depending on the content complexity of the shot are selected as key frames. Lastly, the selected key frames from each shot are merged. To select key frame from each shot, the problem is solved by minimizing QMI among selected key frames as well as maximizing QMI between the set of selected key frames and the set of sampling frames in that shot. Because the use of QMI, it makes the possibility to simultaneously measure the differences among m key frames, as well as to measure the dependency between variables with different dimension of feature vectors. Unlike most of the existing key frame extraction methods, our algorithm finds key frames in the optimization manner. Through the experiments, we demonstrate that the proposed method selects key frames that represent the entire content of video with less redundancy better than the compared approaches.

REFERENCES

[1] A. G. Money, H. Agius, "Video summarization: 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.
[2] Ajmal, Muhammad and Ashraf, Muhammad Husnain and Shakir, Muhammad and Abbas, Yasir and Shah, Faiz Ali, "Video Summarization: Techniques and Classification," Proceedings of the 2012 International Conference on Computer Vision and Graphics, pp. 1–13, 2012.
[3] B. T. Troung, S. Venkatesh,"Video abstraction: a systematic review and classification," ACM Transactions Multimedia Computing, Communications and Applications, Vol. 3, No. 1, 2007.
[4] M. Furini, F. Geraci, and M. Montangero, "VISTO: Visual STOryboard for web video browsing," CIVR, pp.635-641, 2007.
[5] Z. Li, G. M. Schuster, and A. K. Katsaggelos, "MINMAX optimal video summarization," IEEE Trans Circuits Syst. Video Technol., vol.15, no.10, pp.1245-1256, 2005.
[6] C. Panagiotakis, A. Doulamis, and G. Tziritas, "Equivalent key frames selection based on iso-content principles," IEEE Trans. Circuits Syst. Video Technol., vol.19, no.3, pp.447-451, 2009.
[7] G. Guan, Z. Wang, S. Lu, J. D. Deng, and D. D. Feng, "Keypoints-based keyframe selection," IEEE Trans. Circuits Syst. Video Technol., vol.23, no.4, 2013.
[8] S. E. D. Avila, A. B. P, Lopes, L. J. Antonio, and A. d. A. Araujo, "VSUMM: a mechanism designed to produce static video summaries and novel evaluation method," Pattern Recognition Letter, vol.32 (1), pp.56-68, 2011.
[9] N. Ejaz, T. B. Tariq, and S. W. Balik, "Adaptive key frame extraction for video summarization using an aggregating mechanism," Journal of Visual Communication and Image Representation, vol.23, pp.1031-1040, 2012.
[10] N. D. Doulamis, A. D. Doulamis, Y. Avrithis, and S. D. Kollias, "A stochastic framework for optimal frame extraction from MPEG video databases," Comput. Visi. Image Understand., vol.75, no.1-2, pp.3-24, 1999.
[11] A. Nagasaka and Y. Tanaka, "Automatic video indexing and full-video search for object appearances," in Visual Database Systems II, 1992.
[12] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in Proc. IEEE Int. Image Process., pp.866-870, 1998.
[13] P. Mundur, Y. Rao, and Y. Yesha, "Keyframe-based video summarization using Delaunay clustering," International Journal on Digital Libraries (IJDL) 6(2), pp.219-232, 2006.
[14] 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.
[15] J. Almeida, N. J. Leite, and Ricardo da S. Torres, "VISON: VIdeo Summarization for ONline applications," Pattern Recogn. Lett. 33, 4 (March 2012), pp. 397-409, 2012.
[16] Z. Zhao and A. Elgammal, "Information theoretic key frame selection for action recognition," in Proc. Of British machine vision, pp.1-10, 2008.
[17] T. Liu, H. J. Zhang, and F. Qi, "A novel video key-frame-extraction algorithm based on perceived motion energy model," IEEE Trans. Circuits Syst. Video Technol., vol.13, no.10, pp.1006-1013, 2013.
[18] B. Fauvet, P. Bouthemy, P. Gros, and F. Spindler, "A geometrical key-frame selection method exploiting dominant motion estimation in video," in Proc. CIVR, 2004.
[19] W. Wolf, "Key frame selection by motion analysis," in Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc., 1996.
[20] W. Barhoumi and E. Zagrouba, "On-the-fly extraction of key frames for efficient video summarization," AASRI Conference on Intelligent Systems and Control, vol.4, pp.78-84, 2013.
[21] N. Ejaz, I. Mehmood, and S. W. Baik, "Efficient visual attention based framework for extracting key frames from videos," Signal Processing: Image Communication, vol.28(1), pp.34-44, 2013.
[22] H. Zhang, J. Wu, D. Zhong, and S. W. Smoliar, "An integrated system for content-based video retrieval and browsing," Pattern Recognition, vol.30, no.4, pp.643-658, 1997.
[23] K. Torkkoa, "Feature extraction by non-parametric mutual information," J. Machine Learning Research, vol.3, pp.1415-1438, 2003.
[24] J. Sainui and M. Sugiyama, "Direct approximation of quadratic mutual information and its application to dependence-maximization clustering," *IEICE Trans. Inf. & Syst.*, vol.E96-D, no.10, pp.2282-2285, 2013.
[25] J. Sainui and M. Sugiyama, "Minimum dependency key frames selection via quadratic mutual information ", The 10th International Conference on Digital Information Management (ICDIM2015), pp. 148-153, 2015.
[26] M.J. Swain and D.H. Ballard, "Color indexing", International Journal of Computer Vision, 7 (11), pp. 11-32, 1991.
[27] The Open Video Project (Online). Available: http://www.open-video.org (Accessed on 27/10/2015).
[28] Video SUMMarization (Online). Available: https://sites.google.com/site/vsummsite/download (Accessed on 27/10/2015).
[29] (Online). Available: http://www.liv.ic.unicamp.br/ jurandy/vison/VISON Summary.zip (Accessed on 16/05/2016).
[30] (Online). Available: http://www.liv.ic.unicamp.br/∼jurandy/summaries (Accessed on 16/05/2016).