Content-Based Diagnostic Hysteroscopy Summaries for Video Browsing

*Wilson Gavião and Jacob Scharcanski
Instituto de Informática
Universidade Federal do Rio Grande do Sul
Avenida Bento Gonçalves, 9500.
Porto Alegre, RS, Brazil 91501-970
{wgaviao,jacobs}@inf.ufrgs.br

Abstract

In hospital practice, several diagnostic hysteroscopy videos are produced daily. These videos are continuous (non-interrupted) video sequences, usually recorded in full. However, only a few segments of the recorded videos are relevant from the diagnosis/prognosis point of view, and need to be evaluated and referenced later. This paper proposes a new technique to identify clinically relevant segments in diagnostic hysteroscopy videos, producing a rich and compact video summary which supports fast video browsing. Also, our approach facilitates the selection of representative key-frames for reporting the video contents in the patient records. The proposed approach requires two stages. Initially, statistical techniques are used for selecting relevant video segments. Then, a post-processing stage merges adjacent video segments that are similar, reducing temporal video over-segmentation. Our preliminary experimental results indicate that our method produces compact video summaries containing a selection of clinically relevant video segments. These experimental results were validated by specialists.

1 INTRODUCTION

In human reproduction health, diagnostic hysteroscopy is becoming a popular method for assessing and visualizing important regions of the female reproductive system (e.g. cervical channel, uterine cavity, tubal ostea and endometrial characteristics). Diagnostic hysteroscopy is performed by gynecologists with a small lighted telescopic instrument (hysteroscope). During an examination, the hysteroscope transmits an image sequence (i.e. video) to a TV monitor, while the gynecologist guides the instrument to visually assess, diagnose and treat different uterine disorders.

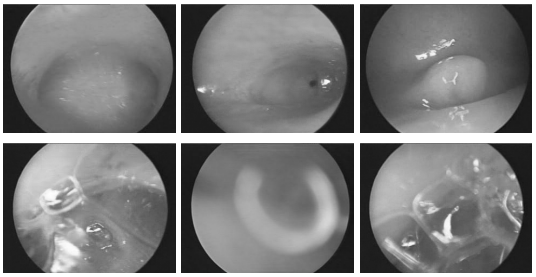

Figure 1. Illustration of frames selected as relevant by our method, showing unobstructed views of the regions of interest (top row). In the bottom row are illustrated some frames discarded by our method, characterized by regions with mucus, and other undesired features.

In practice, several diagnostic hysteroscopy videos are produced daily. Each diagnostic hysteroscopy lasts 1.5-2 minutes, generating a continuous (non-interrupted) video sequence. Usually, the video sequences are recorded in full for further evaluation and reference. However, only portions of the recorded videos are relevant from the diagnosis/prognosis point of view, and need to be evaluated and referenced later. The frames of relevant video segments provide an unobstructed view of important details of the reproductive system (see Figure 1). The video segments whose frames are corrupted by lighting effects (e.g. highlights), or affected by biological features like mucus secretion (as exemplified in Figure 1), can not be used for diagnosis/prognosis, and do not need to be further evaluated.

After each hysteroscopic video is recorded, a further evaluation is done by browsing it, and selecting representative frames that support the diagnosis/prognosis. Usually, the

*The authors thank CNPq - Brazilian Research Council - for financial support.

relevant frames are described in the patient records for future reference. This phase tends to be significantly longer than the hysteroscopic examination itself.

Therefore, a summarization method that provides fast video browsing can be useful in the daily practice. The time required for video browsing and content description can be optimized, while providing a rich hysteroscopy summary for the patient records. Besides, browsing examination details based on the summary can be faster and more accurate than the usual manual frame selection.

In our proposed scheme, specialists would be able to access a video summary based on a few chosen key-frames, e.g. in cases of normal uterus appearance, or, when signs of abnormality are present, they would be able to access more key-frames (and their associated video segments) to describe such cases in detail for the patient records. This paper presents the first steps towards this goal, proposing statistical techniques to identify clinically relevant segments in diagnostic hysteroscopy videos, and their associated key-frames. This work is part of a research effort to provide adaptive endoscopic video summaries, either for fast video browsing and/or inclusion in the electronic patient records, following a hierarchical video representation approach [6].

The majority of the video summarization techniques presented in the literature, propose methods for video parsing and key-frame identification considering the way production videos are created. This is achieved in general by reducing inter-frame redundancy, and by parsing videos in the traditional video units, like shots and scenes [5, 4, 1, 3, 8]. However, diagnostic hysteroscopy videos are produced as continuous sequences, and it is not straightforward modelling them in terms of these traditional video units. Therefore, unfortunately, we did not find appropriate published works to compare to our approach.

This paper is organized as follows. Our approach is detailed in Sections 2 and 3. An overview of the proposed method and the experimental evaluation of our video summarization approach are presented in Sections 4 and 5, respectively. The concluding remarks and ideas for future work are presented in Section 6.

2 THE PROPOSED METHOD

A video hysteroscopy is generally performed in four distinct phases (or steps). In each phase, specific examination goals are achieved, as described next [2]:

- *Uterine cavity*: when the internal cervical orifice is passed, the uterine cavity is examined. First, a panoramic view is performed, and then the examination proceeds with the identification and examination of both tubal orifices. Figure 6(a), on the bottom, shows some images captured during the panoramic view phase;
- *Left (or right) tubal orifice* examination;
- *Right (or left) tubal orifice* examination. Figure 1 (image on the middle of first row) illustrates a image captured in the course of tubal orifice phase;
- *Uterine fundus*: the optical system approaches the uterine fundus to visualize and examine its endometrial characteristics. Figure 6(a) illustrates, on the middle, some images captured during this phase;

When the specialist is performing a diagnostic video hysteroscopy, he/she guides the hysteroscope seeking relevant clinical findings. Little time is spent observing clinically irrelevant areas, but most examination time is spent examining areas that may be relevant for the diagnosis/prognosis. When the relevant areas are found, the specialist focuses the micro camera on the region of interest, or moves it slowly to also examine its surroundings. Therefore, clinically relevant video segments tend to have similar frames (i.e. are static, or redundant, video segments). This is verified in all phases of a diagnostic hysteroscopy examination. This is a fundamental hypothesis for our video summarization approach, which was confirmed experimentally, as detailed next.

In order to estimate activity in video segments, several methods can be used [1]. In this work, we use the distance between color histograms H_i belonging to adjacent video frames X_i and X_{i+1}. The adopted histogram distance metric $D(H_i, H_{i+1})$ is the Jeffrey divergence [7]. Conceptually histograms are empirical probability distributions, which should be compared by a distance measure for probability distributions (e.g. Jeffrey divergence). The Jeffrey divergence was chosen because it provides the best results, considering a set of other histogram difference metrics, such as the histogram intersection and Minkowski distance [7].

$$D(H(X_i), H(X_{i+1})) = \sum_j H_j(X_i) \log \frac{H_j(X_i)}{\overline{H}(j)} \quad + \\ + \quad H_j(X_{i+1}) \log \frac{H_j(X_{i+1})}{\overline{H}(j)} \quad , \quad i \in [1, N-1] \quad (1)$$

where $H_j(X_i)$ and $H_j(X_{i+1})$ are histogram entries corresponding to the histogram bin j, for the successive frames X_i and X_{i+1}; $\overline{H}(j) = [H_j(X_i) + H_j(X_{i+1})]/2$ is the mean histogram; and N is the number of frames in the video.

Small $D(H(X_i), H(X_{i+1}))$ values correspond to small differences between adjacent frame histograms (i.e. the frames are similar). Therefore, small $D(H(X_i), H(X_{i+1}))$ values occur in redundant (i.e. static) video segments, and large divergence values occur in less static video segments. Consequently, $D(H(X_i), H(X_{i+1}))$ can be used as a redundancy measure for each video frame X_i.

We investigated experimentally the hypothesis that relevant video segments have redundant (i.e. static) frames, using 10 interpreted diagnostic hysteroscopic videos. Figure 4 shows, on the right, the histogram representing the distribution of adjacent frame distances d of the video segments selected by the specialists for their video summaries. In general, the distances in those relevant segments are smaller than the distances obtained for all adjacent video frames (Figure 4, on the left). The mean μ and standard deviation σ of these distributions (see Table 2) confirm that lower d values are obtained for the video segments clinically relevant, with a smaller dispersion around the mean. Therefore, our preliminary experimental evidence indicates that clinically relevant video segments have redundant (i.e. static) frames, and this is verified in all phases of a diagnostic hysteroscopy examination.

Our approach uses an adaptive threshold τ to discriminate between static video segments, characterized by small inter-frame divergence $D(H(X_i), H(X_{i+1}))$ values, and non-static video frames (i.e. dynamic segments). It is not trivial to determine the threshold value τ, since the decision between static or dynamic segments tends to subjective. Our approach is to decide the value of τ based on probabilistic models for the static and dynamic segment classes.

We regard $P(d)$ as the probability of divergence value d, given all $D(H(X_i), H(X_{i+1}))$, and $i = 1, ..., N-1$. Let h_o be the hypothesis that a given d value characterizes a redundant frame; and h_1 be the hypothesis that d characterizes a non-redundant frame. Therefore, according to Figure 4, the probability of d given that h_1 occurs, namely $P(d|h_1)$, shall increase with d values increasing; consequently, the probability of d given that h_0 occurs, i.e. $P(d|h_0)$, shall decrease with increasing d values.

The accumulated probability $P_C(d)$ is adopted as a model for $P(d|h_1)$:

$$P(d|h_1) \equiv P_C(d) = \sum_{\gamma=0}^{\gamma=d} P(\gamma) \tag{2}$$

where $d \in \{D(H(X_i), H(X_{i+1}))\}$.

The probability model for $P(d|h_0)$ is then:

$$P(d|h_0) = 1 - P(d|h_1) = 1 - \sum_{\gamma=0}^{\gamma=d} P(\gamma) \tag{3}$$

The threshold τ is the d value that makes $P(d|h_0) = P(d|h_1)$, minimizing the error of confirming h_0 when h_1 is true, and vice-versa. Therefore,

$$P(\tau|h_0) = P(\tau|h_1) \tag{4}$$

$$\sum_{\gamma=0}^{\gamma=\tau} P(\gamma) = 1 - \sum_{\gamma=0}^{\gamma=\tau} P(\gamma) \tag{5}$$

and,

$$\sum_{\gamma=0}^{\gamma=\tau} P(\gamma) = \frac{1}{2} \tag{6}$$

From the above discussion, we conclude that a reasonable estimate is $\tau = median\{d\}$. The threshold τ is chosen as the median of the histogram distances for a given diagnostic hysteroscopic video Thus, a video frame X_i is considered as redundant, and coming from a static (i.e. relevant) video segment, if :

$$D(H(X_i), H(X_{i+1})) \leq \tau \tag{7}$$

confirming the hypothesis h_0 for frame i; otherwise, h_1 is confirmed for this frame. Therefore, all adjacent frames satisfying Equation 7 are the video segments considered relevant for the video summary.

According to the proposed scheme, the video is hierarchically summarized. The relevant video segments S_k constitute the initial summary, $k = 1, ..., M$ where M is the number of relevant video segments; the next stage is constituted by the key-frames X^k of the relevant video segments S_k. In our approach, a key-frame X^k is the frame $X_i \in S_k$ with the smallest distance $D(H(X_i), H(X_{i+1}))$ (i.e. the most redundant frame according to this measure). The highest level in our summary is constituted by the key-frames chosen manually by specialists (among the key-frames of the video summary), to describe the video phases in the patient records.

Figure 2 shows, for a particular diagnostic hysteroscopy video, the distances between adjacent frames as vertical bars (see Equation 1). The horizontal axis represents the frames X_i in the temporal video sequence, and the dotted line represents the threshold τ. Consecutive gray bars represent the video segments discarded by the threshold τ, and the consecutive black bars represent the relevant video segments retained. The key-frame occurs in the temporal position corresponding to the smallest black bar within each relevant video segment. The arrows indicate these key-frames.

Based on the adaptive threshold τ, the extracted key-frames can be very similar (i.e. redundant), as illustrated in Figure 6(a). This occurs because short relevant video segments S_k and S_{k+1} usually are located temporally close to each other in the continuous video sequence. The three relevant video segments on the left in Figure 2 illustrate this problem. Therefore, we present a post-processing step in the next section.

3 POST-PROCESSING

In order to reduce the frame redundancy when selecting key-frames, we propose to merge consecutive relevant video segments that are temporally close. Therefore, two

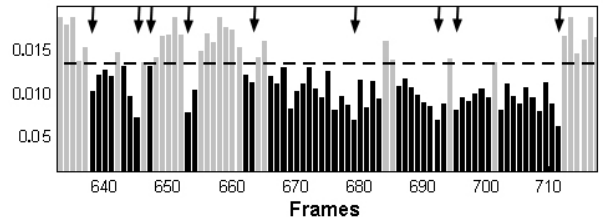

Figure 2. Diagram showing adjacent frame distances $D(H(X_i), H(X_{i+1})$ as bars. Horizontal axis represents each frame X_i in the temporal sequence of the video, and the vertical axis represents $D(H(X_i), H(X_{i+1})$. Dotted line is the adaptive threshold τ. Gray bars represent video segments discarded and black bars represent video segments selected. The smallest black bar within each selected video segment denotes their corresponding key-frame. The arrows indicate these key-frames.

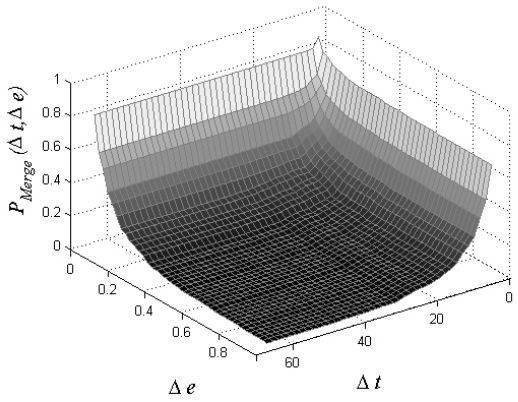

Figure 3. $P_{Merge}(\Delta t, \Delta e)$ for a particular histeroscopic video.

consecutive relevant video segments S_k and S_{k+1} are merged based on features of their respective key-frames X^k and X^{k+1}:

- Δt: temporal distance between the key-frames X_i^k and X_j^{k+1}, defined as

$$\Delta t(X_i^k, X_j^{k+1}) = |i - j| \quad , \quad i, j \in [1, N-1], \quad (8)$$

- Δe: difference in color statistics between key-frames X^k and X^{k+1}, namely $\Delta e(X^k, X^{k+1})$, which is represented by the Jeffrey distance histogram in Figure 2 (see Equation 1).

In this context, key-frames X^k and X^{k+1} from consecutive relevant video segments S_k and S_{k+1} presenting small Δt and Δe values tend be more similar visually, and are more likely to be redundant. Therefore, we compact more the video summaries, minimizing loss of medical information, by merging redundant video segments S_k and S_{k+1}, and forming larger video segments, as described next.

Let $P(\Delta t, \Delta e)$ be the joint probability of Δt and Δe values for all $\Delta t(X^k, X^{k+1})$ and $\Delta e(X^k, X^{k+1})$, $k = 1, ..., M$. Then, the accumulated probability $P_C(\Delta t, \Delta e)$ is :

$$P_C(\Delta t, \Delta e) = \sum_{\varphi=\Delta t_{min}}^{\varphi=\Delta t} \sum_{\psi=\Delta e_{min}}^{\psi=\Delta e} P(\varphi, \psi) \quad (9)$$

Similar to Equation 9, the degree of confidence that two consecutive relevant video segments S_k and S_{k+1} should be merged is given by :

$$P_{Merge}(\Delta t, \Delta e) = 1 - P_C(\Delta t, \Delta e) \quad (10)$$

Therefore, the confidence that two relevant video segments should be merged increases with Δt and Δe values decreasing, because they are more likely to be similar. High P_{Merge} values (i.e., values near to one) indicate that only video segments whose key-frames are temporally/visually very similar should be merged. We leave P_{Merge}, namely ξ, as a parameter to be set by specialists. Figure 3 illustrates $P_{Merge}(\Delta t, \Delta e)$ for a particular histeroscopic video.

The combination of Δt and Δe values satisfying a given ξ value ($\xi \subset [0, 1]$) is not unique (see Figure 3). However, for a given ξ value, there will be a unique pair of maximum Δt and Δe values leading to ξ :

$$\arg\max_{\Delta t, \Delta e} \{P_{Merge}(\Delta t, \Delta e) = \xi\} \quad (11)$$

Thus, for a given video, two consecutive relevant video segments S_k and S_{k+1} are merged if their key-frames X^k and X^{k+1} satisfy :

$$\begin{cases} \Delta t(X^k, X^{k+1}) \leq \Delta t \quad \wedge \quad \Delta e(X^k, X^{k+1}) \leq \Delta e, \\ \text{where} \quad \arg\max_{\Delta t, \Delta e} \{P_{Merge}(\Delta t, \Delta e) = \xi\} \end{cases} \quad (12)$$

As mentioned before, the video summary is constituted by one key-frame per obtained video segment. Depending on the parameter ξ, more compact video summaries will be produced for fast video browsing. In the next section we present a complete overview of our proposed method.

4 OVERVIEW OF THE PROPOSED METHOD

Our method is outlined below, and consists of the following processing steps:

1. Compute the color histogram distances $D(H(X_i), H(X_{i+1}))$ (see equation 1) between adjacent frames X_i and X_{i+1} for $i = 1, ..., N-1$. N is the number of frames in the video;

2. Compute the adaptive threshold $\tau = median\{d\}$ as the median of the histogram distances d;

3. Compute the relevant video segments S_k ($k = 1, ..., M$ where M is the number of relevant video segments):

   (a) All adjacent frames satisfying Equation 7 are the video segments considered relevant for the video summary;

4. Compute the key-frames X^k:

   (a) X^k is the frame $X_i \in S_k$ with the smallest distance $D(H(X_i), H(X_{i+1}))$ (i.e. the most redundant frame according to this measure);

5. Merge consecutive relevant video segments S_k and S_{k+1}:

   (a) Compute the temporal distances Δt and the differences in color statistics Δe between consecutive key-frames X^k and X^{k+1}, according to Equations 8 and 1 respectively;

   (b) Compute the joint accumulated probability of Δt and Δe values for all $\Delta t(X^k, X^{k+1})$ and $\Delta e(X^k, X^{k+1})$, according to Equation 9;

   (c) Compute $P_{Merge}(\Delta t, \Delta e)$ according to Equation 10;

   (d) Set ξ ($\xi \subset [0, 1]$). Low values produce more compact video summaries (i. e., less redundant video summaries);

   (e) Two consecutive relevant video segments S_k and S_{k+1} are merged if their key-frames X^k and X^{k+1} satisfy the Equation 12;

6. Based on the new arrangement of relevant video segments S_k, where $k = 1, ..., F$ and $F \leq M$, repeat step 4 to compute the new set of key-frames;

7. Select a useful key-frame visually, with the help of a specialist, and retain only the selected video frames to store in the patient records.

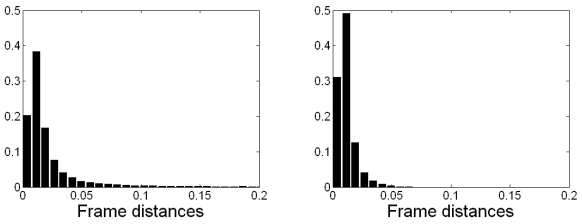

Figure 4. Histograms of adjacent frame distances computed for all videos. On the left, histogram for all frames; On the right, histogram of adjacent frame distances for the relevant video segments selected by the specialists.

5 EXPERIMENTAL RESULTS

We implemented our method in Matlab, and conducted experiments in ten interpreted hysteroscopy videos (namely, v1, ..., v10). These videos were tape recorded at 30 frames per second, and digitalized in AVI format. Among the ten videos, two were taken from patients presenting signs of abnormality. Two different specialists evaluated the videos, without any knowledge of our results, and their evaluation was compared with the results obtained by our method. For every video, the specialists selected the video segments they considered relevant and, according to them, would be enough extracting just one representative frame (i.e. key-frame) from each of these segments. We left the quantity of video segments to be chosen freely by the specialists. A summary of the manual summarization of the videos is described in Table 1.

In our experiments, we utilize a color histogram derived from the HSV (hue, saturation and value) color space [1]. In fact, we split the HSV color space in 134 non-uniforms regions (bins) in order to capture the variations in the tones of red more precisely, because these tones are characteristic in hysteroscopies. Also, in order to integrate spatial and color information, we divide each frame in 9 blocks (3x3), and compute for each block a color histogram. These nine histograms are then concatenated, constituting a vector with m elements, where $m = 1206$. Therefore, each frame X_i of a hysteroscopy video is represented by a vector $H(X_i)$.

As mentioned before in Section 2, the relationship between redundant (or static) video segments and the video segments clinically relevant can be established from experimental evidence.

In order to evaluate our summarization approach, we computed for each video the threshold $\tau = median\{d\}$, and selected relevant video segments according to Equation 7.

Figure 5 provides an indication of the locations of rele-

Table 1. Manual summarization of the videos by specialists.

Videos	Number of frames	Number of relevant segments	Number of frames within relevant segments	Number of key-frames
v1	2591	5	469	5
v2	3078	5	384	5
v3	10844	8	363	8
v4	2365	8	278	8
v5	3878	7	382	7
v6	2309	7	255	7
v7	2703	5	602	5
v8	2489	7	384	7
v9	4159	4	222	4
v10	1750	4	342	4

Table 2. Comparison between adjacent frame distances computed from video segments clinically relevant, and adjacent frame distances computed for the entire video.

	Frame distances for entire video	Frame distances for video segments clinically relevant
μ	0.0248	0.0114
σ	0.0442	0.0073

vant, and irrelevant, frames within the temporal sequence of the video, for all hysteroscopic examination phases. Also, Figure 1 depicts some frames selected as relevant by our method, and these frames are confirmed as presenting unobstructed views of the regions of interest. In the bottom row, also are illustrated some frames discarded by our method, characterized by regions with mucus, and other undesired features.

Our summarization results for the ten interpreted hysteroscopic videos are described in Table 3. As expected, in videos containing few regions of interest (i.e., the video segments where the gynecologist spent most of the examination time, generating static video segments), a higher percentage of the video frames was considered relevant by our method. In longer videos, with more regions of interest, more relevant segments and key-frames were selected.

We build more compact video summaries, and minimize medical information loss, by merging adjacent video segments according to Equation 12, forming larger video segments. Table 3 shows our preliminary results in the 4th. and 5th. columns; these results are further detailed in Figure 6.

The specialists selected from 4 to 8 segments from each video, with an average duration of 2 seconds per segment. Our method detected a larger set of relevant video segments (i.e. detected some *false positives*), comparing to the specialists. Perhaps the most promising result is that all segments selected by the specialists had an intersection with video segments provided by our summarization approach, even when ξ was set to a low value, i. e., $\xi = 0.2$ (see section 2). However, for $\xi \leq 0.1$, some videos had relevant video segments erroneously merged and, consequently, these segments generated only one key-frame when there should be more than one.

The main disadvantage of our method is that it is not able to discard short redundant segments appearing within dynamic segments. Therefore, some segments included in the summary could in fact be discarded. Nevertheless, our approach provides relatively compact summaries for fast diagnostic hysteroscopy video browsing, that contain potentially relevant visual information. Considering all videos tested, our method achieved a mean summarization rate around 2.83% for $\xi = 0.2$ (see Table 3 in the 4th column), retaining at least one key-frame from each relevant video segment selected by the specialists. Therefore, our summaries provide adequate choices for fast browsing, and to produce video descriptions for the patient records.

6 CONCLUDING REMARKS

We propose statistical techniques to identify clinically relevant segments in diagnostic hysteroscopy videos, and their associated key-frames, as means to produce rich video summaries for fast browsing. This work also presents experimental evidence that clinically relevant video segments present a significant redundancy, providing the basis of our approach, and this was verified in all phases of diagnostic hysteroscopy examinations.

Experimentation of our method based on the set of ten interpreted hysteroscopy videos was satisfactory, from the specialists point of view. However, our preliminary results indicate that our method tends to produce less compact video summaries, comparing with summaries provided by specialists. A promising result is that when specialists sum-

Table 3. Summarization results obtained by our method.

Videos	Summarization rate before merge	Number of key-frames for browsing before merge	Summarization rate after merge ($\xi = 0.2$)	Number of key-frames for browsing after merge ($\xi = 0.2$)
v1	0.162	421	0.034	91
v2	0.137	422	0.028	89
v3	0.146	1585	0.033	366
v4	0.154	366	0.033	80
v5	0.115	449	0.024	96
v6	0.106	245	0.021	51
v7	0.129	350	0.027	75
v8	0.118	267	0.025	63
v9	0.080	334	0.017	71
v10	0.168	294	0.041	72

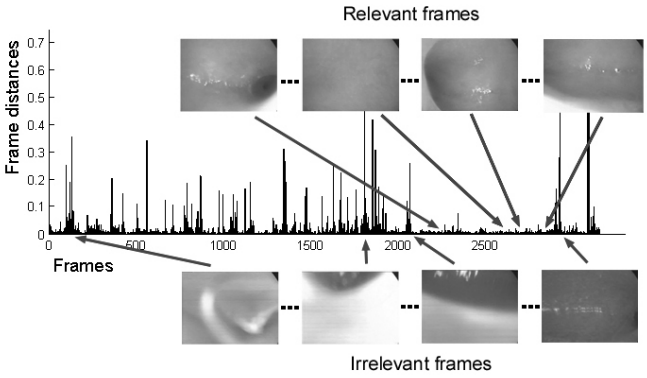

Figure 5. Diagram illustrating relevant and irrelevant frames, and their locations in the video sequence. Relevant frames are associated with smaller adjacent frame distances. Horizontal axis represent each frame X_i in the temporal sequence of the video, and the vertical axis represents $D(H(X_i), H(X_{i+1})$.

marize the same videos manually, they usually select a subset of the video segments provided by our summarization approach.

Future work will concentrate on improving our hierarchical video representation by assigning relevance to the segments in our video summary, and by eliminating spurious redundant frames in dynamic video segments. Besides, we intend test our method in diagnostic endoscopic videos.

References

[1] A. Del Bimbo. *Visual Information Retrieval*. Morgan Kaufmann, San Francisco, USA, 1999.

[2] J. Hamou. *Hysteroscopy and Microcolpohysteroscopy, Text and Atlas*. Appleton and Lange, USA, 1991.

[3] C. Huang and B. Liao. A robust scene-change detection method for video segmentation. *IEEE Trans. Circuits Syst. Video Technol.*, 11(12):1281–1288, dec 2001.

[4] Y. Li, S. Narayanan, and C. Jay Kuo. Content-based movie analysis and indexing based on audiovisual cues. *IEEE Trans. Circuits Syst. Video Technol.*, 14(8):1073–1085, Aug 2004.

[5] T. Liu, H. Zhang, and F. Qi. A novel video key-frame-extraction algorithm based on perceived motion energy model. *IEEE Trans. Circuits Syst. Video Technol.*, 13(10):1006–1013, Aug 2003.

[6] P. Mulhem, G. Gensel, and H. Martin. *Adaptive Video Summarization*, chapter 11, pages 279–298. in Handbook on Video Databases, CRC Press, 2003.

[7] Y. Rubner, J. Puzicha, C. Tomasi, and J. M. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. *Computer Vision and Image Understanding*, 84(1):25–43, Oct 2001.

[8] N. Vasconcelos and A. Lippman. Statistical models of video structure for content analysis and characterization. *IEEE Trans. on Image Processing*, 9(1):3–19, 2000.

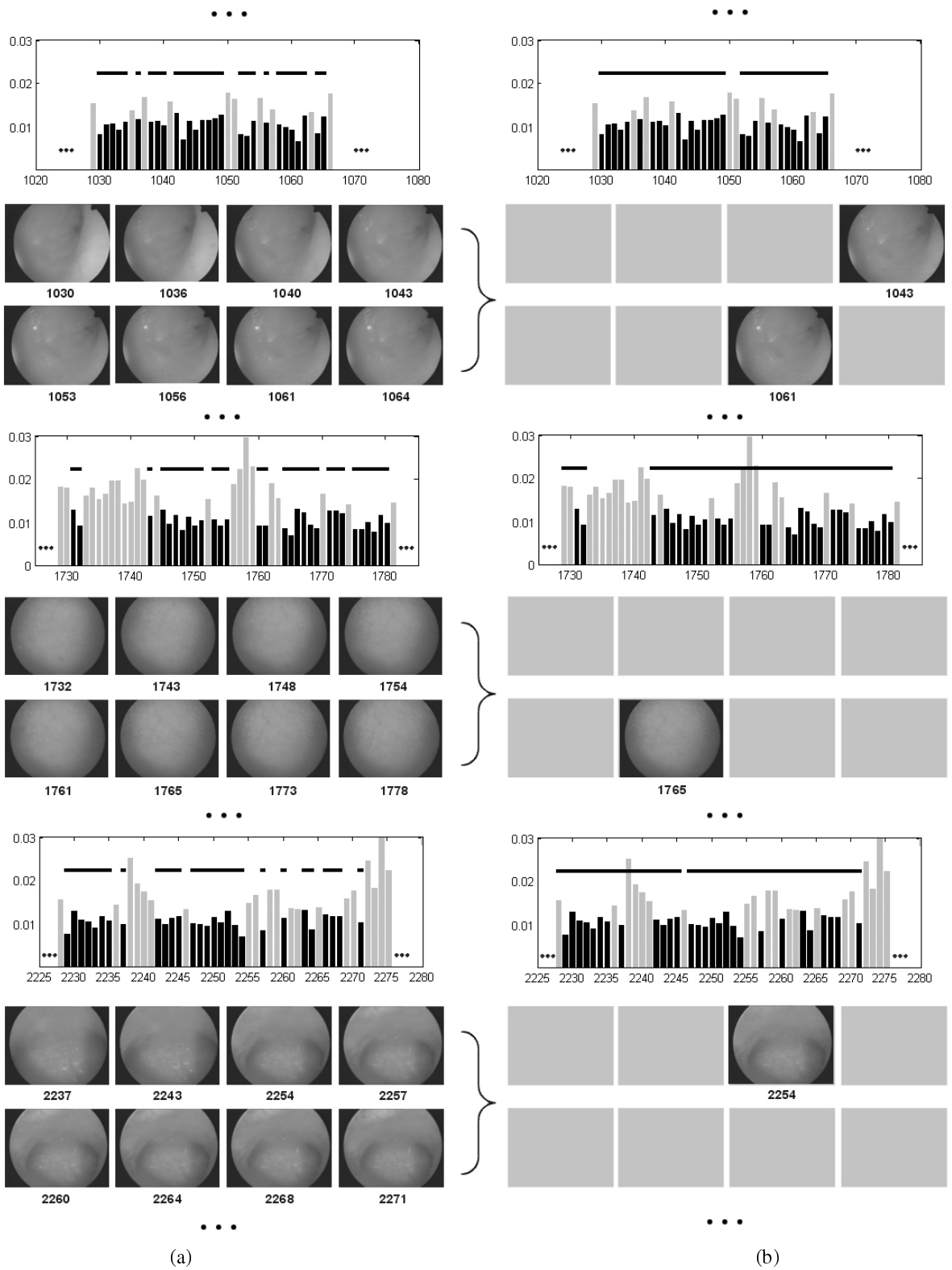

Figure 6. Summarization results for $\xi = 0.2$. Diagrams represent the video segments illustrated in Figure 2, and the horizontal line segments indicate the temporal locations of the relevant video segments; (a) sequence of adjacent video segments (represented by their key-frames) before merging; (b) key-frames of the obtained segments after merge. After merge, the resulting video summary contains 80 key-frames/video segments in total (each video segment is represented by one key-frame).