EVALUATING THE MID-SECRETORY ENDOMETRIUM APPEARANCE USING HYSTEROSCOPIC DIGITAL VIDEO SUMMARIZATION

Wilson Gavião ¹ and Jacob Scharcansi ²,¹

Universidade Federal do Rio Grande do Sul, Instituto de Informática Caixa, Postal 15064, 91501-970, Porto Alegre, RS, Brasil

Abstract

Often, it is necessary to evaluate the mid-secretory endometrium appearance in Gynecology. For this purpose, hysteroscopic videos have been used, and are of fundamental importance nowadays for diagnosis/prognosis of several uterine pathologies. These videos are continuous (non-interrupted) video sequences, usually recorded in full. However, only a few segments of the recorded videos are relevant for diagnosis/prognosis, and need to be evaluated and referenced later. This paper proposes a new technique to identify clinically relevant segments in diagnostic hysteroscopy videos and, consequently, to find images that present the best view of the endometrium details (e.g. glandular openings and vascularization). Our method produce a rich and compact video summary which supports fast video browsing. This method is based on an extension of known properties of the singular value decomposition (SVD), and it is adaptive, in the sense that it minimizes the need of parameter adjustments. 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.

Key words: medical image analysis, video summarization, hysteroscopies, gynecology

Preprint submitted to Elsevier Science 26 August 2005
1 Introduction

Hysteroscopy is a surgical procedure in which a gynecologist uses a small lighted telescopic instrument (hysteroscope), to diagnose and treat several uterine disorders. Using fiber-optic technology, the hysteroscope transmits an image of the uterine canal and cavity to a television monitor, allowing the gynecologist to guide the instrument into the endometrial cavity. There are two types of hysteroscopy, namely, diagnostic and operative hysteroscopy. Diagnostic hysteroscopy is performed to examine the uterus for signs of normalcy or abnormality. Operative hysteroscopy is performed to treat a disorder after it has been diagnosed. Our focus is on diagnostic hysteroscopy.

In hospital 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. Therefore, after each hysteroscopic video is recorded, a further evaluation is done by browsing its contents, and selecting representative frames that support the diagnosis/prognosis. Usually, the relevant frames are described in the patient records for future reference. This phase tends to be significantly longer than the hysteroscopic examination itself.

Frames of relevant video segments provide an unobstructed view of important details of the reproductive system (see Figure 1 at top and middle rows). The video segments whose frames are corrupted by lighting effects (e.g. highlights), or af-

Email addresses: wgaviao@inf.ufrgs.br (Wilson Gavião),
jacobs@inf.ufrgs.br (Jacob Scharcanski).
1 The authors thank CNPq - National Research Council - for support during the development of this work.
fected by biological features like mucus secretion (as exemplified in Figure 1 at the bottom row), can not be used for diagnosis/prognosis, and do not need to be further evaluated.

The morphologic, hormonal, biochemical and ultrastructural changes of the mid-secretory endometrium are essential for the embryo implantation and pregnancy progress. The study and knowledge of endometrial modifications are essential for defining an implantation window and, consequently, embryo attachment. One of the most important and evident physiologic event that occur during this period is the expression of glandular openings [1]. These structures are related to the reproductive status and prognosis.

Hysteroscopy has been utilized to evaluate the endometrium functional modifications [2,1], by observing the distribution of vascular and glandular characteristics (endoscopic aspect).

Therefore, in order to characterize hysteroscopic videos in terms of the occurrence of glandular openings and other findings, 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 this paper, we propose a diagnostic hysteroscopy video summarization method. The proposed technique selects frames that provide useful views of the uterus details based on statistical principles, underlying some known properties of the singular value decomposition (SVD). Afterwards, the specialists may confirm the frames selected in terms of the presence or absence of important structures (e.g. glandular openings). This work is part of a research effort to identify and quantify glandular openings objectively in hysteroscopic images. According to gynecologists, a small number of frames suffice in general to qualify and quantify the occurrence of glan-
dular openings, and other details in a hysteroscopic video. The Appendix provides insight about techniques used to further process the selected frames, to identify and quantify glandular openings.

This paper is structured as follows: Section 2 presents details about hysteroscopic videos and their contents. Also, we contextualize our work with respect to the current literature in the field. Section 3 describes the some SVD properties used in our approach. Section 4 presents the proposed methodology to summarize hysteroscopic videos. The experimental results are discussed in Section 5. Finally, our conclusions are presented in section 6.

2 Hysteroscopic Videos and Related Work

In general, there are four distinct phases (or steps) in a video hysteroscopy. In each phase, specific examination goals are achieved, as described next [3]:

- **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 1, on the top left, shows a typical frame of the panoramic view phase;

- **Left (or right) tubal orifice** examination;

- **Right (or left) tubal orifice** examination. Figure 1 (middle of first row) illustrates a typical frame captured in the tubal orifice phase;

- **Uterine fundus**: the optical system approaches the uterine fundus to examine its endometrial characteristics. Figure 1 illustrates, on the middle row, typical frames captured during this phase;

Usually, the qualification (and quantification) of glandular openings is based on video segment frames of the *Uterine fundus phase* considered relevant (from the diagnosis/prognosis point of view). Some glandular openings are presented as an
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 [4–11]. 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.

3 Properties of the SVD Relevant to our Work

Let \( A \) be a matrix whose \( i \)th column represent a color histogram \( H_i \), calculated from frame \( A_i \) [12,9,13]. Thus, matrix \( A \) is temporal representation of the video color information. Given an \( m \times n \) matrix \( A \), where \( m \) is the number of color features (histogram bins), \( n \) is a number of frames, and \( m \geq n \), the SVD of \( A \) is defined as:

\[
A = U \Sigma V^T \tag{1}
\]

where \( U = [u_{ij}] \) is an \( m \times n \) column-orthonormal matrix, whose columns are the left singular vectors; \( \Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_n) \) is an \( n \times n \) diagonal matrix whose diagonal elements are non-negative singular values sorted in descending order; and \( V = [v_{ij}] \) is an \( n \times n \) orthonormal matrix whose columns are the right singular vectors.

If \( \text{rank}(A) = q \) we may view the SVD as mapping the \( m \)-dimensional raw feature space, spanned by the color histograms of video frames, into a \( q \)-dimensional (reduced) feature space with linearly-independent dimensions. Each column vector
$A_i$ of $A$, representing the color histogram of frame $i$, is mapped to a $q$-dimensional column vector $\psi_i = [v_{i1} \ v_{i2} \ \cdots \ v_{iq}]^T$ of the matrix $V^T$.

The Euclidean norm square of $\psi_i$ is defined as:

$$\|\psi_i\| = \sqrt{\sum_{j=1}^{\text{rank}(A)} v_{ij}^2}$$  \hspace{1cm} (2)

Since $V$ is orthonormal, if $\text{rank}(A) = n$ then $\|\psi_i\| = 1$, where $i = 1, 2, \ldots, n$ [9].

In this work we explore two properties of the SVD [12,9]:

**Property 1:** Let $B = [B_1, B_2, \ldots, B_n]$ be a $m \times n$ matrix where $m \geq n$ and $\text{rank}(B) = n$, i.e., the columns of $B$ are linearly independent. Let $B_0$ be a $m$-dimensional column vector that is linearly independent of the vectors $B_i$ in $B$, and let $B_D$ be a $m \times d$ matrix obtained by replicating the column vector $B_0$ $d$ times. Thus,

$$B_D = [B_0, B_0, \ldots, B_0] \quad \text{and} \quad \text{rank}(B_D) = 1.$$  \hspace{1cm} (3)

Let $A'$ be a $m \times (n + d)$ matrix that is constituted by the column vectors of $B$ and $B_D$ arranged in any order, for example:

$$A' = [B_1 \ldots \widehat{B_1^j} \ldots B_d^j \ldots B_n].$$  \hspace{1cm} (4)

Now, let the SVD of $A'$ be given by Eq. 1 and $V'^T = [\psi_1 \ \widehat{\phi_1} \ \phi_d \ \cdots \ \psi_n]$ be the corresponding right singular vector matrix obtained. It can be shown that the Euclidean norm square $\|\phi_j\|^2 = 1/d$, where $j = 1, 2, \ldots, d$ and $d$ is the number of times that $B_0$ is replicated in $A'$ (see [12] for proof).

It implies that a column vector $A_i$, which is linearly independent in $A$, is projected by the SVD onto a vector $\psi_i$, whose distance from the origin of the lower dimensional feature space is one, according to Eq. 2. When $A_i$ has copies $A_i^{(j)}$ (i.e. $A_i$ is
redundant, or linearly dependent), the distance of its projected vector \( \phi_j \) from the space origin decreases. The more copies \( A_i \) has (i.e. the more redundant it is), the smaller are their distances to the space origin [9].

**Property 2**: If \( A \) is a \( m \times n \) matrix of rank \( k \) with a singular value decomposition defined by Eq. 1, then the Euclidean distance between any two column vectors of \( A \) is equal to the Euclidean distance between the corresponding column vectors of \( V^T \) weighted by their corresponding singular values [12]. Consequently, the Euclidean distance between a pair of frame feature vectors \( H_i \) and \( H_j \) [9] is:

\[
D(\psi_i, \psi_j) = \sqrt{\sum_{l=1}^{k} \sigma_l (v_{il} - v_{jl})^2}
\] (5)

where \( \psi_i \) and \( \psi_j \) are the vectors representing frames \( i \) and \( j \) in the obtained SVD feature space, and \( \sigma_l \) are the singular values obtained by the SVD.

4 **Our Proposed Method**

When the specialist performs 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.

At this stage, the reader could ask a reasonable question such as: ”why a physician do not simply press a capture button when useful video segments are found ?”. The
answer is not too obvious. In clinical practice, the diagnostic procedure is based on
the examination of different regions of endometrium. Therefore, the micro camera
is relocated frequently, and often artifacts degrade some of the video frames (or
parts of them), like undesired lighting effects, or the obstruction of visual details
caused by the mucus secretion. Even by recording selectively with a record button,
such artifacts still affect the video quality. Besides, the videos are captured as se-
quences related spatially and temporally to the anatomical regions being examined.
If the video sequences are interrupted, and made too short, spatial and temporal
references may be lost, making it difficult to describe anatomically the regions of
interest in the video sequences. Usually, several sub-sequences can be relevant in a
hysteroscopic video, and our video summarization method can help identify these
useful video segments, saving time and directing the specialist’s attention to these
specific video segments for an objective (quantitative) diagnosis.

In order to estimate activity in video segments, several methods can be used [6].
We estimate if a video segment is static or dynamic (redundant or not) based on
properties of the video in the SVD space, where we measure the similarity between
frames. In this work we use property 1, stated in section 3. Based on Eq. 2 we
compute the Euclidean norm square of each frame feature vector $\psi_i$ in the SVD
space spanned by $V^T$. The feature vectors coming from frames of a redundant
(i.e. static) video segment are projected closer to the origin, than feature vectors
of frames coming from less redundant video segments (which are projected farther
away from the origin). Therefore, small $\|\psi_i\|$ values indicate redundant (i.e. static)
video segments, and large values indicate less static video segments. Consequently,
$\|\psi_i\|$ is used as a redundancy measure for each video frame $i$.

Our approach uses an adaptive threshold $\tau$ to discriminate between static video
segments, characterized by small $\|\psi_i\|$ values, and non-static video frames (i.e.
dynamic segments). It is not trivial to determine the threshold value $\tau$, since the
decision between static or dynamic segments tends to subjective.
We regard $P(\|\psi\|)$ as the probability density estimate of $\|\psi\|$, given all $\|\psi_i\|$, and $i = 1, ..., n$. The threshold $\tau$ is chosen as the mode of $P(\|\psi\|)$ for a given diagnostic hysteroscopic video. In order to have a precise estimate of the mode, we model $P(\|\psi\|)$ by a mixture of Gaussian kernels (see [14] for details). Thus, a video frame $i$ is considered as redundant, and coming from a static (i.e. relevant) video segment, if:

$$\|\psi_i\| \leq \tau$$

(6)

Otherwise, this frame is assumed to be non-redundant, and, consequently, not relevant. Therefore, all adjacent frames satisfying Equation 6 are video segments considered relevant for the video summary.

The different groups of adjacent frames selected with Equation 6 are taken as relevant video segments $S_k$, $k = 1, ..., M$ where $M$ is the number of relevant video segments, and constitute the initial video summary. Each relevant video segment $S_k$ has an associated key-frame $K S_k$. In our approach, a key-frame $K S_k$ is the frame $i \in S_k$ with the smallest $\|\psi_i\|$ value (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 select suitable relevant images that will be used in the glandular openings evaluation stage.

For a particular diagnostic hysteroscopy video, Figure 2 shows a series of $\|\psi_i\|$ values associated to frames as vertical bars (please, see Equation 2). The horizontal axis represents the frames $i$ in the temporal video sequence, and the dotted line represents the threshold $\tau$. Consecutive gray bars represent the video segments discarded by the threshold $\tau$, and the consecutive black bars represent the relevant video segments retained. The smallest black bars within each relevant video segment indicate their corresponding key-frame, and the arrows point to these key-frames.
The extracted key-frames based on the adaptive threshold $\tau$ can be very similar (i.e. redundant). This occurs because short relevant video segments $S_k$ and $S_{k+1}$ are usually located temporally close to each other in the continuous video sequence. As a result, we may obtain video over-segmentation. The three relevant video segments on the left in Figure 2 illustrate this problem. To address the video over-segmentation problem we present next a post-processing stage.

4.1 POST-PROCESSING

In order to reduce video over-segmentation and key-frame redundancy, two adjacent relevant video segments $S_k$ and $S_{k+1}$ are merged if the following two conditions are satisfied:

- Let $\|\psi_{S_k}\|$ and $\|\psi_{S_{k+1}}\|$ be the average $\|\psi\|$ value of the video segments $S_k$ and $S_{k+1}$, respectively. Let $\|\psi_{S_x}\|$ be the average $\|\psi\|$ value of an irrelevant video segment that is temporally located in between $S_k$ and $S_{k+1}$. Thus, the first condition to merge $S_k$ and $S_{k+1}$ is:

$$\frac{\|\psi_{S_k}\| + \|\psi_{S_{k+1}}\| + \|\psi_{S_x}\|}{3} \leq \tau.$$ 

(7)

Satisfying this condition requires that after merging the video segments $S_k, S_x$ and $S_{k+1}$, the resulting video segment still can be considered static;

- Another condition must be satisfied to minimize the chance of merging video segments, and connecting boundary frames that are not similar (i.e. merging video segments that are visually distinct).

We compute the adjacent frame distances $D(\psi_i, \psi_{i+1})$ within all relevant video segments $S_k$ (see Eq. 5). The accumulated probability of $D(\psi_i, \psi_{i+1})$ values is denoted as $P_C(D(\psi_i, \psi_{i+1}))$. Let $\psi^{k}_{\text{Last}}$ and $\psi^{k+1}_{\text{First}}$ be the last and the first feature frame vectors of segments $S_k$ and $S_{k+1}$, respectively, in the SVD space. Thus, for a given diagnostic hysteroscopic video, the segments $S_k$ and $S_{k+1}$ are merged
if:

\[ P_C(\psi_{\text{Last}}^k, \psi_{\text{First}}^{k+1}) \leq \nu \]  

(8)

In all our experiments, we set \( \nu = 0.99 \).

Based on the discussion above, two adjacent relevant video segments \( S_k \) and \( S_{k+1} \) are merged if Equations 7 and 8 are satisfied.

## 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 digitized 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 [6]. 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 a color histogram for each block. These nine histograms are then concatenated, constituting a vector with \( m \) elements, where \( m = 1206 \). Therefore, each frame \( i \) of a hysteroscopy video is represented by a column vector \( A_i \) in
Figure 3 provides an indication of the locations of relevant, 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 2. More compact video summaries are built by merging adjacent video segments according to Equations 7 and 8, at same time, medical information loss is minimized. Table 2 shows our preliminary results in the 4th. and 5th. columns; these results are further detailed in Figures 4.

The specialists selected from 1 to 4 segments for each video, with an average duration of 3.5 seconds per segment. Therefore, our method actually detects a larger set of relevant video segments (i.e. some false positives), comparing to the selection done by specialists. However, the summaries provided by our method tend to not generate false negatives (i.e. discard relevant video segments as they were irrelevant), and tend to be complete, containing images that represent all phases of the hysteroscopy.

The most promising result is that all segments selected by the specialists had an intersection with video segments provided by our summarization approach. We obtained 100% recall of the relevant segments, but lower precision. However, the obtained summaries are very compact, as shown in Table 2.  

The main disadvantage of our method is that it is not able to discard short redundant segments.

\[ \text{Precision} = \frac{\#R_{\text{segments}}}{\#R_{\text{segments}} + \#I_{\text{segments}}} \]

where \( \#R_{\text{segments}} \) and \( \#I_{\text{segments}} \) are the number of relevant and irrelevant video segments detected, respectively.

2 Precision was measured as the ratio \( \frac{\#R_{\text{segments}}}{\#R_{\text{segments}} + \#I_{\text{segments}}} \), where \( \#R_{\text{segments}} \) and \( \#I_{\text{segments}} \) are the number of relevant and irrelevant video segments detected, respectively.
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. For all videos, our method achieved a mean summarization rate around 1.8% (see the 4th column of Table 2), retaining at least one frame from each relevant video segment selected by the specialists. Therefore, our summaries provide adequate choices to help specialists searching fast for good video segments to evaluate medical findings.

A qualitative and quantitative evaluation of endometrial features (e.g. glandular openings distribution and vascularization) is based on key-frames of relevant video segments of the *Uterine fundus* phase. A method for detecting and quantifying glandular openings, and evaluating qualitatively their distribution, is presented in the Appendix (Section 7).

### 6 Conclusions

The hysteroscopic appearance of uterus is of great importance for gynecologists. However, the obtained hysteroscopic videos usually contain a lot of irrelevant data, and in general, after searching the entire video sequence for findings, experts tend to select only the reduced number of frames where these findings occur, and use them to support his/her diagnosis/prognosis. In this paper, we propose a technique for hysteroscopic video summarization, which extracts from a video sequence the segments (and their key-frames) that are likely to be relevant from the diagnosis and/or prognosis point of view. This approach can reduce the expert’s visual search for relevant findings to a small set of video segments, where these findings are likely to occur.

Our preliminary experiments with a set of hysteroscopic videos were promising, and have indicated that our method can provide results comparable to the visual
inspection of the same videos by gynecologists.

In the continuation of this work, we intend to refine the features extracted from the video frames, in order to reduce false positives, and obtain more compact video summaries of all hysteroscopy phases.

7 Appendix - Detection of Glandular Openings

Once adequate key-frames are found by the hysteroscopic video summarization technique, these images are used as input to detect glandular openings automatically, reducing the subjectivity in uterine fundus appearance evaluation.

Glandular openings appear in uterine fundus video frames as spots relatively brighter than the background. To automatically detect these spots is challenging, because they usually are heterogeneous in brightness, and do not contain only one brightness maximum. Given an image $I$, we propose a method based on morphological concepts: grayscale reconstruction $\rho_{I}(J)$, regional minimum $RMIN(I)$ and extended maxima transformations $EMAX_{h}(I)$ [15,16].

Ideally, each regional maximum region corresponds to a glandular opening. However, image artifacts, like highlight effects and spots with weak contrast interfere. In order to detect only relevant spots and eliminate artifacts, we impose constraints on regional maxima. Initially, a threshold is used to remove regions (i.e. artifacts) with weak contrast. Thus, each candidate region exceeds a local contrast threshold $h$ w.r.t. the background, eliminating spurious regional maxima whose difference between the maximum intensity value (i.e. top of the spot) and the background intensity value (i.e. base of the spot) is less than $h$. This is achieved by computing the extended maxima transform of $I$.

To eliminate highlight effects, we use an intensity value threshold $m$ determined in a set-up stage, before operation. The spots containing pixels with intensity higher
than \( m \), are assumed highlights (i.e. are unlikely to correspond to any biological structure of the mid-secretory endometrium). We observed experimentally that pixel intensities in highlights can be approximated by a half-Gaussian distribution. The value of \( m \) is set as the distribution outlier, i.e. any \( n > m \) belongs to the highlights half-Gaussian distribution with 95\% of probability. Thus, in practice, the selection of \( h = 3 \) and \( m = 200 \) provided satisfactory results in all hysteroscopy videos tested.

We start by detecting the background around glandular openings. Therefore, two regions are identified: regional minima \( RMIN(I) \), and intersections \( D_{ij} \) (when existing) between bright spots \( M^i \) and \( M^j \), \( i \neq j \). These intersections are defined using the relative support concept proposed in [17]. By definition, the relative support of a maximum \( M^i \), \( \text{supp}(M^i) \), is the smallest dilation of \( M^i \) which has a nonempty intersection with \( RMIN(I) \). The method to obtain the relative support is [17]:

\[
\begin{align*}
M_o^i &= M^i \\
M_k^i &= M_{k-1}^i \oplus H \\
k_o^i &= \inf \{ k \in \mathbb{N} \mid M_k^i \cap RMIN(I) \neq \emptyset \} \\
\text{supp}(M^i) &= M_{k_o^i}^i
\end{align*}
\]

(9)

where \( H \) is the basic structuring element. This is analogous to a front propagation starting at \( M \) and stopping when it meets a minimum. Next, we define the smallest intersection \( D_{ij} \) between the maxima \( M^i \) and \( M^j \) (\( i \neq j \)) as:

\[
\begin{align*}
\arg \min_{a,b} \{ a \in [1, \ldots, k_o^i], b \in [1, \ldots, k_o^j] \mid M_a^i \cap M_b^j \neq \emptyset \}, \quad \text{where} \quad i \neq j \\
D_{ij} &= \text{supp}(M_a^i) \cap \text{supp}(M_b^j), \quad \text{if} \quad \text{supp}(M_a^i) \cap \text{supp}(M_b^j) \neq \emptyset \quad (10) \\
D_{ij} &= \emptyset, \quad \text{if} \quad \text{supp}(M_a^i) \cap \text{supp}(M_b^j) = \emptyset
\end{align*}
\]
Thus, the set \( D(EMAX_h(I)) \) of smallest intersections \( D_{ij} \) among the maxima of \( I \) (computed by \( EMAX_h(I) \)) is defined as the union of all smallest intersection regions:

\[
D(EMAX_h(I)) = \bigcup_{i \neq j} D_{ij}.
\]  

(11)

The regional minima \( RMIN(I) \) provide seeds to detect the background in the neighborhood of each glandular opening. These background regions are detected by a grayscale reconstruction, using regional minima as markers \( \rho(I(RMIN(I))) \). This procedure obtains local background regions, and exclude the glandular openings.

Glandular openings are detected by subtracting the reconstructed image from the original, i.e. \( I - \rho(I(RMIN(I))) \). The resulting image is binarized using a threshold at graylevel 1, so every spot brighter than the local background is detected. Fig. 5(b) illustrates the obtained results. To individuate these spots, we compute the smallest intersection regions \( D(EMAX_h(I)) \), and their union with the set of regional minima is used to obtain a marker image \( I(RMIN(I) \cup D(EMAX_h(I))) \) (it is zero outside of the union set). Finally, the glandular openings spot detection procedure is defined as:

\[
I - \rho(I(RMIN(I) \cup D(EMAX_h(I)))).
\]  

(12)

The obtained results are illustrated in Fig. 5(c), showing that bright spots were individuated.

The method is outlined below, as a sequence of processing steps:

1. Perform median and mean filter on the images (frames), which are denoted by \( I \) in what follows;
2. Compute \( \rho(I(RMIN(I) \cup D(EMAX_h(I)))) \) performing grayscale reconstruction, and using the union set of regional minima and smallest intersec-
tions $I(RMIN(I) \cup D(EMAX_h(I)))$ as a marker;

(3) Compute $I_r = I - \rho_I(I(RMIN(I) \cup D(EMAX_h(I))))$;

(4) Obtain regions containing glandular openings by thresholding $I_r$ at gray level 1;

(5) Discard regions containing highlights (i.e. with pixel intensities in $I$ higher than $m$).

References


Table 1
Manual summarization of the videos by specialists.

<table>
<thead>
<tr>
<th>Videos</th>
<th>Number of frames</th>
<th>Number of relevant segments</th>
<th>Number of frames within relevant segments</th>
<th>Number of key-frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>v1</td>
<td>2590</td>
<td>1</td>
<td>280</td>
<td>1</td>
</tr>
<tr>
<td>v2</td>
<td>3075</td>
<td>4</td>
<td>375</td>
<td>4</td>
</tr>
<tr>
<td>v3</td>
<td>2365</td>
<td>4</td>
<td>606</td>
<td>4</td>
</tr>
<tr>
<td>v4</td>
<td>3830</td>
<td>2</td>
<td>93</td>
<td>2</td>
</tr>
<tr>
<td>v5</td>
<td>2700</td>
<td>3</td>
<td>327</td>
<td>3</td>
</tr>
<tr>
<td>v6</td>
<td>2487</td>
<td>1</td>
<td>80</td>
<td>1</td>
</tr>
<tr>
<td>v7</td>
<td>4157</td>
<td>3</td>
<td>186</td>
<td>3</td>
</tr>
<tr>
<td>v8</td>
<td>1748</td>
<td>2</td>
<td>147</td>
<td>2</td>
</tr>
<tr>
<td>v9</td>
<td>3080</td>
<td>1</td>
<td>84</td>
<td>1</td>
</tr>
<tr>
<td>v10</td>
<td>2100</td>
<td>4</td>
<td>347</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 2
Summarization results obtained by our method.

<table>
<thead>
<tr>
<th>Videos</th>
<th>Summarization rate before merge</th>
<th>Number of key-frames for browsing before merge</th>
<th>Summarization rate after merge</th>
<th>Number of key-frames for browsing after merge</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>v1</td>
<td>0.091</td>
<td>238</td>
<td>0.021</td>
<td>55</td>
<td>13 %</td>
</tr>
<tr>
<td>v2</td>
<td>0.021</td>
<td>67</td>
<td>0.006</td>
<td>20</td>
<td>25 %</td>
</tr>
<tr>
<td>v3</td>
<td>0.087</td>
<td>206</td>
<td>0.022</td>
<td>54</td>
<td>31 %</td>
</tr>
<tr>
<td>v4</td>
<td>0.083</td>
<td>319</td>
<td>0.033</td>
<td>128</td>
<td>2 %</td>
</tr>
<tr>
<td>v5</td>
<td>0.050</td>
<td>135</td>
<td>0.015</td>
<td>42</td>
<td>12 %</td>
</tr>
<tr>
<td>v6</td>
<td>0.059</td>
<td>149</td>
<td>0.019</td>
<td>49</td>
<td>6 %</td>
</tr>
<tr>
<td>v7</td>
<td>0.051</td>
<td>212</td>
<td>0.021</td>
<td>91</td>
<td>3 %</td>
</tr>
<tr>
<td>v8</td>
<td>0.073</td>
<td>128</td>
<td>0.013</td>
<td>24</td>
<td>25 %</td>
</tr>
<tr>
<td>v9</td>
<td>0.049</td>
<td>153</td>
<td>0.018</td>
<td>58</td>
<td>2 %</td>
</tr>
<tr>
<td>v10</td>
<td>0.054</td>
<td>114</td>
<td>0.012</td>
<td>27</td>
<td>44 %</td>
</tr>
</tbody>
</table>
Fig. 1. Illustration of frames selected as relevant by our method, showing unobstructed views of the regions of interest (top and middle rows). In the bottom row are illustrated some frames discarded by our method, characterized by regions with mucus, and other undesired features.
Fig. 2. Diagram showing the norm $\|\psi\|$ of the frames as bars. Horizontal axis represents each frame $i$ in the video temporal sequence, and the vertical axis represents $\|\psi_i\|$. Dotted line is the adaptive threshold $\tau$. 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.
Fig. 3. Diagram illustrating relevant and irrelevant frames, and their locations in the video sequence. Relevant frames are associated with smaller values of $\|\psi\|$. Horizontal axis represent each frame $i$ in the temporal sequence of the video, and the vertical axis represents $\|\psi_i\|$. 
Fig. 4. Summarization results. Diagrams representing video segments as explained in Figure 2. The horizontal line segments indicate relevant video segments: (a) sequence of key-frames representing adjacent video segments before merging; (b) key-frames of the obtained video segments after merge. The resulting summary after merge contains 54 key-frames/video segments in total (each video segment is represented by one key-frame).
Fig. 5. (a) Grayscale filtered image; (b) Regions with white borders were obtained by the subtraction \( I - \rho(I(RMIN(I))) \); (c) Final glandular openings detection: Final glandular openings segmentation obtained from \( I - \rho_I(I(RMIN(I) \cup D(EMAX_h(I)))) \) are delimited by the black borders; (d) The detected glandular openings (black points) are superimposed to the original image.