Real-Time Segmentation and Recognition of Surgical Tasks in Cataract Surgery Videos

Article *in* IEEE Transactions on Medical Imaging · December 2014

DOI: 10.1109/TMI.2014.2340473 · Source: PubMed

Gwenole Quellec
French Institute of Health and Medical Research
111 PUBLICATIONS **4,021** CITATIONS

Mathieu Lamard
Université de Bretagne Occidentale
122 PUBLICATIONS **3,418** CITATIONS

Real-time segmentation and recognition of surgical tasks in cataract surgery videos

Gwénolé Quellec Mathieu Lamard Béatrice Cochener Guy Cazuguel

IEEE Trans Med Imaging — in press

Abstract

In ophthalmology, it is now common practice to record every surgical procedure and to archive the resulting videos for documentation purposes. In this paper, we present a solution to automatically segment and categorize surgical tasks in real-time during the surgery, using the video recording. The goal would be to communicate information to the surgeon in due time, such as recommendations to the less experienced surgeons. The proposed solution relies on the content-based video retrieval paradigm: it reuses previously archived videos to automatically analyze the current surgery, by analogy reasoning. Each video is segmented, in real-time, into an alternating sequence of idle phases, during which no clinically-relevant motions are visible, and action phases. As soon as an idle phase is detected, the previous action phase is categorized and the next action phase is predicted. A conditional random field is used for categorization and prediction. The proposed system was applied to the automatic segmentation and categorization of cataract surgery tasks. A dataset of 186 surgeries, performed by ten different surgeons, was manually annotated: ten possibly overlapping surgical tasks were delimited in each surgery. Using the content of action phases and the duration of idle phases as sources of evidence, an average recognition performance of $A_z = 0.832 \pm 0.070$ was achieved.

Index terms— cataract surgery, real-time video analysis, content-based video retrieval, conditional random fields

1 Introduction

During eye surgeries, the surgeon wears a binocular microscope and the output of the microscope can be recorded. It is now common practice to record every surgical procedure and to archive the resulting videos for documentation purposes [1]. Therefore, it may now be possible to automatically monitor the surgery. Such a tool would be useful to communicate information (about the tools or the patient) to the surgeon in due time, typically at the beginning of a new surgical task. In the particular case of new surgeons, the system could also provide recommendations on how to best perform the current or the next task, based on the experience of their peers in similar surgeries (similar patients, similar implants, etc.). To allow such a communication system, surgical tasks need to be detected in real-time during the surgery.

In recent years, a few systems were presented for the automatic recognition of surgical tasks or gestures, assuming a known temporal segmentation of the tasks or gestures. Two approaches were evaluated for the automatic classification of surgical gestures in video clips of minimally invasive surgery [2]: one is based on a linear dynamical system, the other is based on the Bag-of-Words (BoW) model [3, 4]. Combining these two approaches leads to the same performance as gesture classification based on kinematic data [2]. If visual cues are combined with kinematic data, using multiple kernel learning, classification performance is pushed even further [5, 6]. We have presented a solution for the automated recognition of cataract surgery tasks [7]: short subsequences are characterized in the video stream and video subsequences with similar structures are searched in a video archive. The proposed feature vectors are unchanged by variations in duration and temporal structure among the target surgical tasks.

A second group of systems was presented for the automatic temporal segmentation of surgical tasks or gestures, given the full surgical video. Such a system was proposed for the automatic segmentation of a surgical task into a sequence of gestures, in laparoscopic videos [8]. The system relies on sparse Hidden Markov Models (HMM), whose observations are sparse linear combinations of elements from a gesture-specific dictionary of basic surgical motions. Another system was presented for the automatic temporal segmentation of surgical tasks, also in laparoscopic videos [9]. During the training phase, tool usage is analyzed to perform dimension reduction on visual features, using canonical correlation analysis. At the end of the surgery, the video is registered to a manually segmented average surgery, using Dynamic Time Warping (DTW) or an HMM. A similar system based on tool usage rather than visual cues was also presented [10]. A system was proposed for the automatic segmentation and recognition of surgical gestures using both visual cues and kinematic data, using a combined Markov/semi-Markov Conditional Random Field (CRF) [6]. Finally, one system was presented for the automatic temporal segmentation of surgical phases in microscope videos using DTW or an HMM: the visual content of images is described by color histograms, Haar-based features and SIFT descriptors, among other features [11].

Unlike the first group of systems, the proposed system does not assume a known segmentation of the surgical tasks. Unlike the second group of methods, it does not need the full video in order to temporally segment the surgical tasks: the segmentation is performed as the video is recorded. As proposed in a previous paper [7], the Content-Based Video Retrieval (CBVR) paradigm is used in this purpose [12]. CBVR aims at finding, inside a video collection, videos or video segments that are similar to a query video. In eye clinics archiving a video for every eye surgery, large video collections are indeed quickly available. We propose to reuse these large archives to segment and categorize surgical tasks, by analogy reasoning, during a new video-monitored surgery. In speech recognition, analogy reasoning has been used with HMMs to model temporal evolution [13]. In this paper, we propose to model the temporal evolution of the surgery by a CRF, a statistical modeling method often used for structured prediction as an alternative to HMMs [14]. CRFs have been used for action recognition [15, 16] and surgical gesture recognition [6]. Some solutions rely on visual cues extracted from one frame [15], the so-called Markov CRF model. Other rely on visual cues extracted from fixed-length video segments [16], the so-called semi-Markov model [17]. The combination of both approaches has also been proposed [6]. A novel CRF-based design, specific to surgical task segmentation and categorization, is presented in this paper.

In this paper, we focus on cataract surgery, which is the most common eye surgery [18]. An

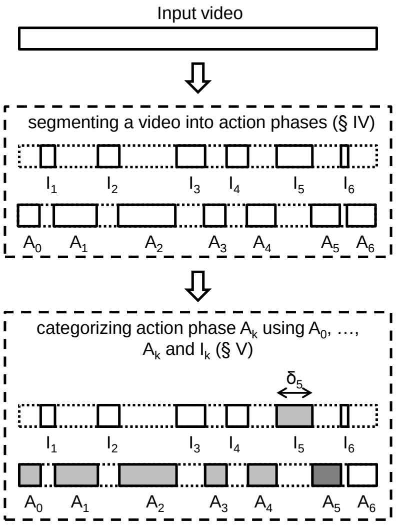

Input video
segmenting a video into action phases (§ IV)
I_1 I_2 I_3 I_4 I_5 I_6
A_0 A_1 A_2 A_3 A_4 A_5 A_6
categorizing action phase A_k using A_0, …, A_k and I_k (§ V)
δ_5
I_1 I_2 I_3 I_4 I_5 I_6
A_0 A_1 A_2 A_3 A_4 A_5 A_6

Figure 1: Method outline. Idle phases are indicated by an upper-case 'I' letter. Action phases are indicated by an upper-case 'A' letter. $k = 5$ in this example.

algorithm has been proposed for the automatic segmentation of cataract surgery videos into surgical phases [19]: it specializes a more general system [11] to this particular surgery, notably by extracting visual features inside the pupil only. However, that algorithm does not allow real-time recognition of the surgical tasks or phases: as mentioned above, the segmentation can only be performed when the surgical video is available in full, i.e. after the end of the surgery. To our knowledge, this paper is the first attempt to segment and recognize eye surgical tasks in real-time.

2 System Overview

In the proposed framework, summarized in Fig. 1, each video is segmented into an alternating sequence of idle phases and actions phases. An idle phase is a phase of the surgery where no clinically-relevant motions are visible. An action phase features one high-level surgical task, or possibly several surgical tasks in a row. To detect idle phases, the system accepts short video subsequences as input and retrieves the most similar video subsequences in a reference dataset (§4).

Then, the segmented action phases are categorized with respect to the appearing surgical tasks. In that second part of the framework, the system accepts video segments as input (the detected action phases) and retrieves the most similar video segments in the reference dataset. A Conditional Random Field (CRF) is used to model the temporal evolution of the action phases. The use of CRFs in the proposed context (adaptive-length video segments, dual action-idle phase segmentation, CBVR) is novel and implies specific designs:

1. The probability that a given surgical task occurs during a given action phase depends on a manual labeling of the most similar action phases within a reference dataset (§5.2.1). The

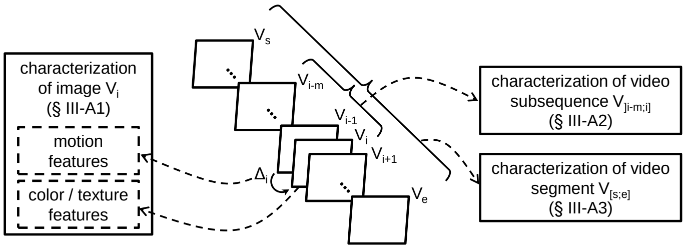

characterization of image V_i (§ III-A1)
motion features
color / texture features
V_s
V_{i-m}
V_{i-1}
V_i
V_{i+1}
V_e
Δ_i
characterization of video subsequence $V_{]i-m;i]}$ (§ III-A2)
characterization of video segment $V_{[s;e]}$ (§ III-A3)

Figure 2: Characterizing video segments

search for similar action phases depends:

(a) either exclusively on the visual content of the query action phase and of the reference action phases (§5.2.1),

(b) or also on the visual content of all action phases preceding the query and reference action phases in their respective surgical videos (§5.2.3).

2. The probability to switch from one surgical task to another, when moving from one action phase to the next, depends on the duration of the separating idle phase (§5.3).

The core CBVR engine was presented in details in a previous publication [7]: it is summarized in the following section. The remaining sections are novel and therefore described in more details.

3 Real Time Content-Based Video Retrieval Framework

This section summarizes the underlying CBVR part of the method. First, it explains how to characterize an arbitrary long segment of a video, such as a high-level surgical task (an incision, a rhexis, a hydrosissection, etc.), using low-level features (§3.1). Then, it explains how to compare two short video subsequences (§3.2), or two video segments (§3.3), using their low-level characterizations.

3.1 Characterizing Video Segments

A video essentially consists of a sequence of images, or *frames*. Let V be a video. Let $V_{[s;e]}$ denote a segment of V starting with frame V_s and ending with frame V_e. First, each frame V_i in that segment, $s \leq i \leq e$, is characterized individually (§3.1.1). Then, the characterization of V_i is combined with the characterization of the $m-1$ preceding frames in V (§3.1.2): $V_{]i-m;i]}$ is referred to as a (fixed-length) video subsequence. Finally, all video subsequence characterizations are combined to describe the video segment $V_{[s;e]}$ as a whole (§3.1.3). The workflow is summarized in Fig. 2. Note that each step in this processing chain is performed in real time.

3.1.1 Image Characterization

Motion features are extracted from the optical flow Δ_i between the previous frame V_{i-1} and the current frame V_i. To compute the optical flow, strong corners are first detected in V_{i-1}. The optical flow is computed at each detected strong corner with the Lucas-Kanade iterative method [20]. This optical flow contains a mixture of clinically-relevant motion information (surgical tool motion, eye motion, etc.) and camera motion information (camera translation, camera rotation, zoom, etc.). Camera motion, modeled as an affine transformation, is first estimated and subtracted [21]. Simple motion features (amplitude, direction and position histograms [7]) are then extracted from the residual motion, which is suspected to be clinically-relevant. Finally, color and texture features are extracted in each color plane of the current frame V_i through a wavelet analysis [22].

3.1.2 Video Subsequence Characterization

In this paper, videos have one frame every 40 milliseconds (§6). Because 40 milliseconds are too short to characterize an elementary surgical tool motion, frames are not analyzed independently. Instead, short video subsequences, consisting of the current frame plus the last few preceding frames, are analyzed as a whole. Given a video subsequence v, a C-dimensional feature vector $f(v) = \{f_c(v), c = 1..C\}$ is extracted.

3.1.3 Video Segment Characterization

Finally, to characterize $V_{[s;e]}$, an arbitrary long segment of a video, the idea is to describe how the feature vectors extracted during that video segment are distributed. Each feature vector component c is described independently, using the cumulative distribution function (CDF) of all $\{f_c(v), v \in V_{[s;e]}\}$ values [7]. Let $F_c(V_{[s;e]})$ denote the c^{th} CDF.

3.2 Comparing Video Subsequences

Two video subsequences u and v are compared using a weighted Euclidean distance:

$$d(u,v) = \sqrt{\frac{1}{\sum_{c=1}^{C} \lambda_c} \sum_{c=1}^{C} \lambda_c [f_c(u) - f_c(v)]^2} \tag{1}$$

where $\lambda_c, c = 1..C$, are feature weights tuned to fill the semantic gap between low-level feature vectors and the high-level concept of semantic distance between manually-delimited video segments [7]. One set of weights is used for idle phase segmentation and another one is used for action phase categorization.

3.3 Comparing Video Segments

When two distributions f and g are described by their CDF, respectively F and G, they can be compared using the two-sample Kolmogorov-Smirnov statistic $D(F,G)$ [23]:

$$D(F,G) = \sup_{x \in \mathbb{R}} |F(x) - G(x)| \tag{2}$$

Therefore, two video segments $U_{[s';e']}$ and $V_{[s;e]}$ are compared using the following distance:

$$D(U_{[s';e']}, V_{[s;e]}) = \sum_{c=1}^{C} \lambda_c D(F_c(U_{[s';e']}), F_c(V_{[s;e]})) \tag{3}$$

4 Segmenting a Video into Action Phases

This section describes how videos are automatically segmented into a sequence of alternate idle phases and action phases. An idle phase is a phase of the surgery where no clinically-relevant motions are visible (4.1). Ideally, an action phase features a high-level surgical task. It may also feature one part of a surgical task only. In some cases, it might also feature two consecutive surgical tasks with no interruption in-between.

4.1 Detecting Idle Phases

In order to detect idle phases in a surgical video U, the nearest neighbors of each video subsequence $u \in U$ are searched in a reference dataset using the weighted Euclidean distance (see equation 1). Videos in that reference dataset have been manually-segmented into an alternate sequence of idle and action phases as described in section 6.3. If most neighbors come from idle phases, according to the human reference standard, then the current video subsequence likely is in an idle phase as well. Therefore, an idle-probability is defined for each video subsequence as the percentage of neighbors coming from idle phases. The relevance of this idle-probability relies on the video analyzer's ability to distinguish between camera motions and clinically-relevant motions (§3.1.1). The weights λ_c in equation 1, $c = 1..C$, are trained at the video segment level (see section 3.3), using the manual segmentation of reference videos into idle and action phases.

To segment idle phases automatically, the idea is to detect whenever the idle-probability becomes higher, or conversely lower, than some cutoff $\tau_p \in [0, 1]$. If the idle-probability function is noisy, a median filter (with a window size n) should be applied to the idle-probability function beforehand. Moreover, idle phases with a duration lower than a predefined threshold $\tau_\delta \in \mathbb{R}^+$ are ignored.

4.2 Alternate Sequence of Idle and Action Phases

On output, the idle phase detector provides a temporal segmentation of each surgical video: it indicates time intervals $[s_k, e_k]$, $k = 1..K$, likely to be idle phases. Let I_k denote the k^{th} idle phase and let $\delta_k = e_k - s_k$ denote its duration.

Once idle phases are detected, action phases are defined dually: an action phase is defined as a time interval delimited by two consecutive idle phases. Let A_k, $k = K_f..K_l$, denote the action phase directly following idle phase I_k. If the video starts with an idle phase, then $K_f = 1$, otherwise $K_f = 0$. If the video ends with an idle phase, then $K_l = K - 1$, otherwise $K_l = K$.

5 Categorizing Action Phases

This section describes how each action phase A_k is automatically categorized with respect to the high-level tasks that the surgeon performed during that phase. Let T_i, $i = 1..N$, denote the high-level tasks that surgeons usually perform during a surgery (incision, rhexis, hydrodissection, etc.). The probability $p_{k,i}$ that a task of type T_i (or a T_i-task for short) was performed during A_k is estimated using a conditional random field (CRF), as described hereafter.

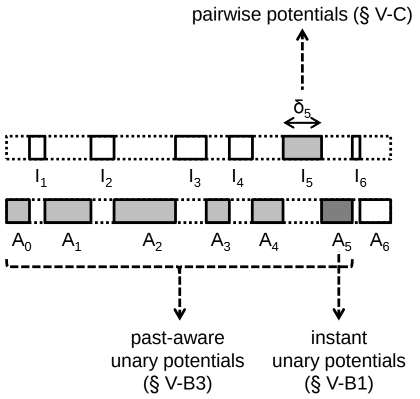


Figure 3: Unary and pairwise potentials

5.1 Conditional Random Fields

Let $\mathbf{A} = (A_k)_{k=K_f..K_l}$ denote an input sequence of action phases and let $\mathbf{y} = (y_k)_{k=K_f..K_l}$ denote an automatic segmentation of $\mathbf{A}$ into surgical tasks, where each element y_k is a label from the fixed set $V = \{T_i, i = 1..N\}$.

5.1.1 Markov Property

a CRF models $Pr(\mathbf{y}|\mathbf{A})$ using a Markov random field. The label variables $\mathbf{y}$, conditioned on the observed variables $\mathbf{A}$, obey the Markov property with respect to a directed graph $G = (V, E)$ [14]:

$$\begin{aligned} &Pr(y_i = T_i|\mathbf{A}, y_j = T_j, T_i \neq T_j) \\ = &Pr(y_i = T_i|\mathbf{A}, y_j = T_j, (T_i, T_j) \in E) \end{aligned} \tag{4}$$

5.1.2 Unary and Pairwise Potentials

unlike inference in HMMs, inference in CRFs relies on any number of feature functions or potentials that do not necessarily have a probabilistic interpretation. Besides, these potentials can inspect the entire input sequence $\mathbf{A}$ at any point during inference. Let $\Psi_k^u(y_k, \mathbf{A}), u = 1..U$, denote U potentials expressing the relevance of assigning the task label $y_k \in V$ to the k^{th} action phase; these potentials are referred to as unary potentials. Let $\Psi_{k-1,k}^v(y_{k-1}, y_k, \mathbf{A}), v = 1..V$, denote V potentials expressing the relevance of switching from task label $y_{k-1} \in V$ to task label $y_k \in V$, $(y_{k-1}, y_k) \in E$, when moving from the $(k-1)^{th}$ to the k^{th} action phase; these potentials are referred to as pairwise potentials. Unary and pairwise potentials are defined in sections 5.2 and 5.3, respectively (see Fig. 3).

5.1.3 Log-linear Model

a CRF can be written as a log-linear model with unary and pairwise potentials [14]:

$$\begin{aligned} Pr(\mathbf{y}|\mathbf{A},\boldsymbol{\lambda},\boldsymbol{\mu}) \propto \exp\Bigg(& \sum_{u=1}^{U} \lambda_u \sum_{k=K_f}^{K_l} \Psi_k^u(y_k,\mathbf{A}) \\ & + \sum_{v=1}^{V} \mu_v \sum_{k=K_f+1}^{K_l} \Psi_{k-1,k}^v(y_{k-1},y_k,\mathbf{A}) \Bigg) \end{aligned} \tag{5}$$

where a vector of weights $\boldsymbol{\lambda} = (\lambda_u)_{u=1..U}$ over the unary potentials and a vector of weights $\boldsymbol{\mu} = (\mu_v)_{v=1..V}$ over the pairwise potentials are learnt from training data by maximum log-likelihood estimation.

5.1.4 Online Inference

inference in a CRF is usually performed with a forward-backward algorithm [24]. We are interested in online inference in a CRF, i.e. inference using present and past information only. So a forward algorithm is used to estimate the probability $p_{k,i}$ that a T_i-task was performed during action phase A_k:

$$\left\{ \begin{array}{lcl} \log(p_{0,i}) & = & \sum_{u=1}^{U} \lambda_u \Psi_0^u(T_i,\mathbf{A}) \\ \log(p_{k,i}) & = & \max\limits_{j} \Bigg[\log(p_{k-1,j}) + \sum\limits_{u=1}^{U} \lambda_u \Psi_k^u(T_i,\mathbf{A}) \\ & & + \sum\limits_{v=1}^{V} \mu_v \Psi_{k-1,k}^v(T_j,T_i,\mathbf{A}) \Bigg] \end{array} \right. \tag{6}$$

5.2 Unary Potentials

Remember that potentials can inspect the entire input sequence $\mathbf{A}$ at any point during inference. Note, however, that we are interested in online inference, so in our case, the definition of potentials can only rely on present and past information. Two sets of unary potentials are defined: the first $\frac{U}{2}$ potentials rely on current observations (§5.2.1), the last $\frac{U}{2}$ potentials rely on all observations since the beginning of the surgery (§5.2.3). All rely on a nearest neighbor search and analogy reasoning.

5.2.1 Instant Unary Potentials

a first set of unary potentials relies on the content of action phase A_k. The idea is to compare the digital content of A_k with that of other action phases in a manually-interpreted reference dataset (§6.2). The distance measure presented in section 3.3 is used to find the nearest neighbors of A_k in the reference dataset. Then, multiple estimators $p_{k,i}^{(n)}$ of probability $p_{k,i}$, $n = n_1, ..., n_{\frac{U}{2}}$, are determined by analogy reasoning (§5.2.2). Potentials are analog to logarithms of probabilities (see equation 5), so the instant unary potentials are defined as follows:

$$\Psi_k^u(T_i,\mathbf{A}) = log\left(p_{k,i}^{(n_u)}\right), u = 1..\frac{U}{2} \tag{7}$$

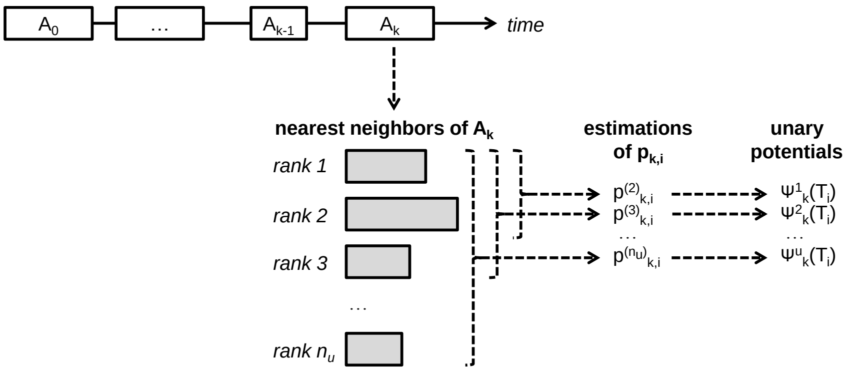


Figure 4: Analogy reasoning

5.2.2 Analogy Reasoning

if a T_i-task was performed in most of the nearest action phases, then the surgeon likely performed a T_i-task in A_k as well. Let n_u denote the number of nearest neighbors retrieved per query. The conditional probability $P_{n_u}(T_i|n')$ that a T_i-task was performed during an action phase, given that the surgeon performed a T_i-task in n' neighbors out of n_u, is estimated by the frequency observed in a training set (§6.2).

During inference, if we observe that the surgeon performed a T_i-task in n' neighbors out of n_u, then $p_{k,i}$ can be estimated by $P_{n_u}(T_i|n')$. Let $p_{k,i}^{(n_u)}$ denote the estimation of $p_{k,i}$ obtained when n_u nearest neighbors are retrieved per query. We don't know what the optimal value for n_u is, but since CRFs allow multiple unary potentials, several estimations of $p_{k,i}$ can be used, by varying the value of n_u: $n_u = n_1, ..., n_{\frac{U}{2}}$. The process is summarized in Fig. 4.

5.2.3 Past-Aware Unary Potentials

the second set of unary potentials is similar to the first one, except that the distance metric is modified in order to take past and present observations $A_l, l = K_f..k$, into account. The nearest action phases are searched in the manually-interpreted reference dataset with respect to the Dynamic Time Warping (DTW) semi-distance [25], where the atomic distance between action phases is still the one presented in section 3.3. Specifically, to assess the relevance of some action phase $A'_{k'}$ in a reference video, the DTW semi-distance between $\{A_l, l = K_f..k\}$ and $\{A'_l, l = K'_f..k'\}$ is computed. $\frac{U}{2}$ new unary potentials are defined using the same analogy reasoning as above (§5.2.2): $\Psi_k^u(T_i, \mathbf{A}), u = \frac{U}{2}+1..U$ (see equation 7). Note that only past observations are taken into account to define those past-aware unary potentials: past label variables are not, in accordance to the Markov property (§5.1.1).

5.3 Pairwise Potentials

In section 5.2, each action phase A_k was automatically categorized using its own digital content, as well as the content of preceding action phases. However, the content of idle phases was ignored.

5.3.1 On the Relevance of Idle Phases

by definition, an idle phase I_k does not contain clinically-significant information (§4.1). So, strictly speaking, its content does not convey useful information. However, its duration δ_k may indicate how the preceding and the following action phases, respectively A_{k-1} and A_k, are linked. For instance, if I_k is very short, the surgeon likely did not have time to put his/her tool down and take another tool, so A_{k-1} and A_k likely belong to the same category. Similarly, if I_k is very long, it may indicate that something wrong happened, so the surgeon may start the same task over.

5.3.2 Transition Categories

we assume that the probability of transition between two consecutive action phases A_{k-1} and A_k depends on τ_k, the duration of the idle phase separating A_{k-1} and A_k. All (A_{k-1}, A_k) couples in the training set are grouped in m_v categories of equal cardinal, with respect to τ_k: $c_1^{(v)}, c_2^{(v)}, ..., c_{m_v}^{(v)}$, $|c_1^{(v)}| = |c_2^{(v)}| = ... = |c_{m_v}^{(v)}|$. Then, the probability $P_{m_v}(T_i, T_j|c_l^{(v)})$ to switch from task $A_{k-1} = T_i$ to task $A_k = T_j$, if τ_k is in the $c_l^{(v)}$ category, is estimated by the frequency observed among all training (A_{k-1}, A_k) couples in that category.

During inference, if we observe that τ_k is in the $c_l^{(v)}$ category, then the probability of transition from task $A_{k-1} = T_i$ to task $A_k = T_j$ can be estimated by $P_{m_v}(T_i, T_j|c_l^{(v)})$. Let $p_{k,i,j}^{(m_v)}$ denote the estimated probability of transition obtained when m_v duration categories are used.

5.3.3 Idle Phase Dependent Pairwise Potentials

we don't know what the optimal value for m_v is, but since CRFs allow multiple pairwise potentials, several estimations of the transition probabilities $p_{k,i,j}^{(m_v)}$ can be used, by varying the value of m_v: $m_v = m_1, ..., m_V$. The pairwise potentials are defined as follows:

$$\Psi_k^v(T_i, T_j, \mathbf{A}) = log\left(p_{k,i,j}^{(m_v)}\right), v = 1..V \tag{8}$$

If $m_v > 1$, these pairwise potentials depend on observations (τ_k) and are therefore non-stationary. In the particular case $m_v = 1$, the transition probabilities become stationary, and the resulting pairwise potentials are equivalent to those proposed by Tao et al. [6].

5.4 CRF Implementation

The Wapiti library [26] was used for training a linear-chain CRF. The Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) Quasi-Newton algorithm was used to learn the weight vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ (§5.1.3). The library was modified to allow online inference in a CRF as described in section 5.1.4.

6 CRAB Dataset

A dataset of 186 videos from 186 consecutive cataract surgeries was collected at Brest University Hospital (Brest, France) between February and July 2011. 153 patients were involved: 33 patients had cataract surgery performed on both eyes, 120 had cataract surgery performed on one eye only. Prior to each surgery, demographic data (age, sex and race) and contextual data (presence of diabetes, deafness, inflammatory disease, small pupil size, etc.) were collected and deidentified.

6.1 Video Recording

Surgeries were performed by 10 different surgeons in two different operating rooms (67 in operating room 1, 119 in operating room 2). In operating room 1, videos were recorded with a CCD-IRIS device (Sony, Tokyo, Japan) and a DSR-20MDP videocassette recorder (Sony, Tokyo, Japan). In operating room 2, videos were recorded with a MediCap USB200 video recorder (MediCapture, Philadelphia, USA). They were stored in MPEG2 format, with the highest quality settings, in operating room 1 and in DV format in operating room 2.

6.2 Training and Test Set

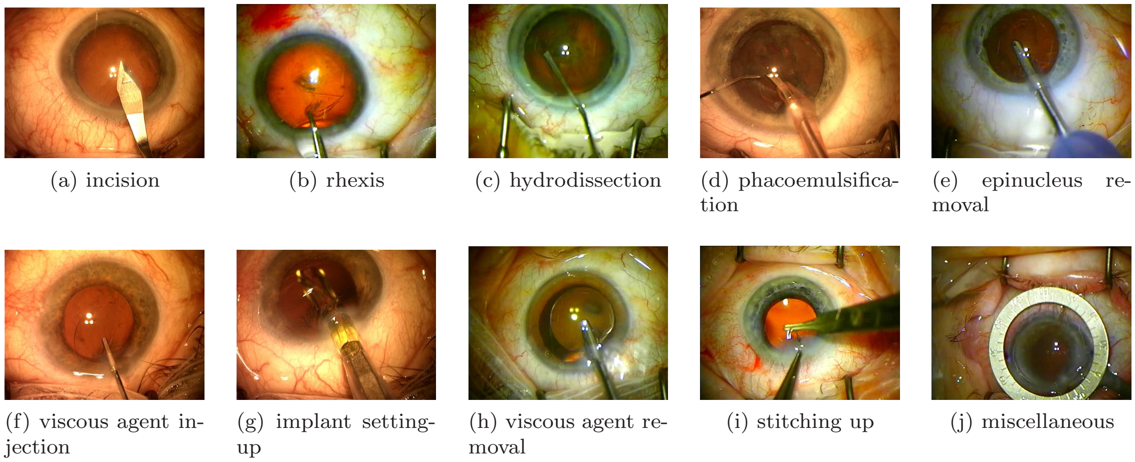
The CRAB dataset was divided into two subsets, referred to as S_1 and S_2, with equal distribution of sex and age. When a patient had cataract surgery performed on both eyes, the associated videos were both assigned to the same subset. Except for the above-mentioned conditions, the dataset was shared out randomly among both subsets. Subset S_1 (93 videos) was used for training and subset S_2 (93 videos) was used for testing. Subset S_1 was also used as reference dataset for analogy reasoning (§5.2.2).

6.3 Idle Phase Segmentation

In order to train the idle phase detector, a temporal segmentation was provided by cataract experts for each idle phase in a random selection of 10 videos from the training set. Let $S_0 \subset S_1$ denote this subset. For each video in S_0, two cataract experts indicated all dates of tool appearance and disappearance from the field of view. Note that the tool used by the surgeon's assistant to moisten the eye was not taken into account since it is not relevant to follow the temporal evolution of the surgery: the eye is moisturized at regular intervals throughout the surgery. The agreement between the two resulting idle phase (no surgical tools visible) / action phase (at least one tool visible) segmentations was 98.5% at frame level. Both segmentations were combined as follows: a frame was assigned to an idle phase if both experts assigned it to an idle phase.

6.4 High-level Surgical Task Segmentation

In each video from the CRAB dataset, a temporal segmentation was provided by one cataract expert. For each surgical task, he indicated the date of first appearance of one tool related to this task into the field of view. Similarly, he indicated the date of last disappearance of one of these tools from the field of view [7]. These segmentations are only used for assessing task categorization, so they do not have to be very accurate at frame level. The following surgical tasks were temporally segmented in videos: incision, rhexis, hydrodissection, phacoemulsification, epinucleus removal, viscous agent injection, implant setting-up, viscous agent removal and stitching up (see Fig. 5). A miscellaneous category was created to account for optional surgical phases: iris retractor setting-up, iris retractor removal, angle measurement, landmark tracing, etc. Miscellaneous surgical phases were also temporally segmented. To assess inter-expert agreement, surgical tasks were temporally segmented by a second cataract expert in each video from S_0: the agreement between both expert segmentations was 97.9% at frame level.



(a) incision (b) rhexis (c) hydrodissection (d) phacoemulsification (e) epinucleus removal

(f) viscous agent injection (g) implant setting-up (h) viscous agent removal (i) stitching up (j) miscellaneous

Figure 5: High-level surgical tasks

7 Training

The following parameters need to be trained:

1. the parameters of the video subsequence analyzer (m, $(\lambda_c)_{c=1..C}$, etc., §3.1),
2. the idle phase detection parameters (τ_p and τ_δ, §4.1),
3. the number of unary potentials and their parameters ($U, n_1, ..., n_{\frac{U}{2}}$, §5.2),
4. the number of pairwise potentials and their parameters ($V, m_1, ..., m_V$, §5.3).

The first group of parameters was tuned in the training set as described in our previous paper [7]. The second group of parameters was determined by a Free-response Receiver Operating Characteristic (FROC) analysis in the training subset S_0 (§8.1). The remaining parameters were optimized through a grid search: each tested tuple of parameter values was assessed by a two-fold cross-validation in the training set.

8 Results

8.1 Idle Phase Segmentation (see Fig. 6)

The idle phase detector was trained by 5-fold cross-validation in S_0 (§6.3). It was evaluated in terms of sensitivity (what percentage of true idle phases was detected?) and in terms of false positive rate (how many false idle phases were detected per video?). The sensitivity and the False Positive Rate (FPR) were measured for different cutoffs τ_p on the idle-probability and different lower bounds τ_δ on the duration of an idle phase. The resulting FROC curve is reported in Fig. 7.

Based on the FROC curve, the following parameters were selected to evaluate action phase categorization: $\tau_p = 0.7$, $\tau_\delta = 0.8s$.

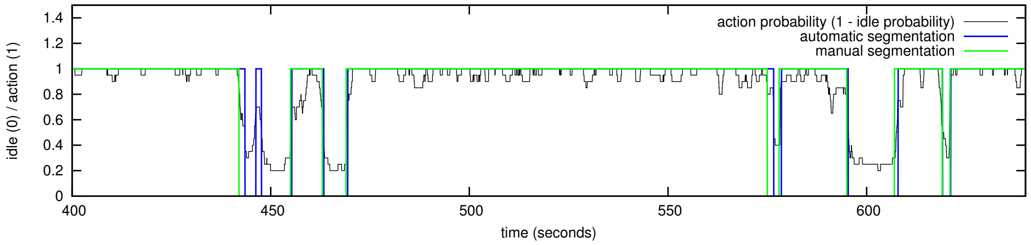


Figure 6: Automatic idle phase segmentation in one surgical video portion. The automatic segmentation is compared to the manual segmentation (§6.3).

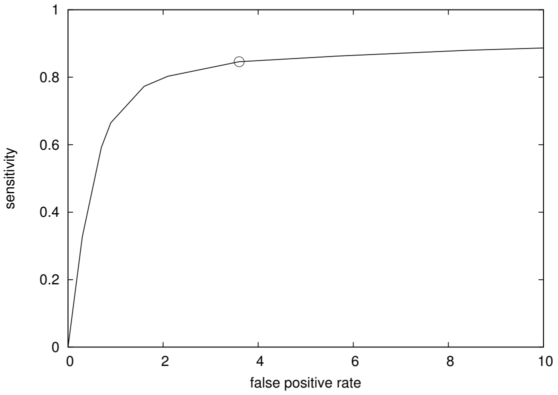


Figure 7: Free-response receiver operating characteristic of the idle phase detector. The circle (FPR=3.6, sensitivity=0.846) indicates the selected settings.

Table 1: Action phase categorization performance.

surgical task	ManS (§8.2.1)	UniS (§8.2.2)	HMM (§8.2.3)	PropWI (§8.2.4)	Prop (§5)
incision	0.741	0.967	0.795	0.906	0.943
rhexis	0.878	0.855	0.785	0.824	0.850
hydrodissection	0.762	0.781	0.668	0.854	0.883
phacoemulsification	0.923	0.795	0.805	0.856	0.891
epinucleus removal	0.969	0.682	0.721	0.808	0.840
viscous agent injection	0.561	0.569	0.560	0.737	0.722
implant setting-up	0.703	0.679	0.750	0.782	0.810
viscous agent removal	0.729	0.675	0.727	0.761	0.768
stitching up	0.883	0.763	0.796	0.824	0.863
miscellaneous	0.591	0.762	0.680	0.705	0.748
average	0.774	0.753	0.729	0.806	0.832

Table 2: Comparing algorithms for action phase categorization performance (p-value of a two-tailed paired t-test).

	ManS	UniS	HMM	PropWI	Prop
ManS		*0.6583*	*0.1924*	*0.3957*	*0.1240*
UniS	*0.6583*		*0.3700*	*0.0607*	0.0053
HMM	*0.1924*	*0.3700*		0.0034	0.0003
PropWI	*0.3957*	*0.0607*	0.0034		0.0011
Prop	*0.1240*	0.0053	0.0003	0.0011	

8.2 Action Phase Categorization

Categorization performance is assessed in terms of A_z, the area under the Receiver Operating Characteristic (ROC) curve. The parameters maximizing A_z in the training set are the following: $M_V \in \{1, 2, 4\}$, $n_u \in \{20, 30, 50, 100\}$. Categorization performance in the test set, for the proposed method (Prop) and a few baseline methods described below, is reported in table 1. ROC curves were built using the online inference probabilities defined in equation 6. Statistical difference between the different methods is assessed by two-sided paired t-tests (see table 2). ROC curves obtained with the proposed method (§5) are reported in figure 8. Overall, an accuracy of 79.3% was achieved.

In terms of computation times, a frame rate of 23.5 FPS was achieved. Most of the computation time (80%) was dedicated to video subsequence characterization, and to image characterization in particular (see section 3.1). All computations were performed using one core of an Intel Xeon(R) E5649 processor running at 2.53GHz.

8.2.1 First Baseline Method: Manual Segmentation (ManS)

a comparison with our previous paper on surgical task categorization was provided [7]. In that paper, the surgical tasks were manually delimited and the task was to categorize these manually-

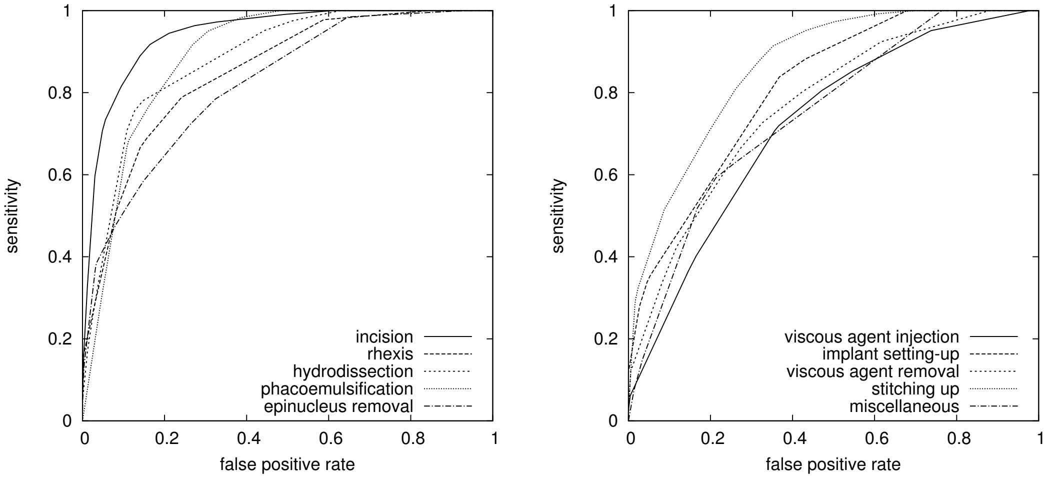


Figure 8: Receiver operating characteristic of action phase categorization (§5)

delimited video segments; the same CBVR engine was used for categorization.

8.2.2 Second Baseline Method: Uniform Segmentation (UniS)

to assess the advantage of idle phase segmentation, a comparison with a semi-Markov CRF [17], where the video is segmented into fixed-length video segments, was provided. The optimal video segment length is searched by two-fold cross-validation on the training set. Besides this different temporal segmentation, the method is identical. Of course, the length of idle phases is not used, since they are not detected. For this baseline method and the following, the task categorization parameters were retrained specifically as described in section 7.

8.2.3 Third Baseline Method: Hidden Markov Model (HMM)

to assess the relevance of CRFs, a comparison with a Hidden Markov Model [19] was provided. A $N \times N$ transition matrix was learned in the training set ($N = 10$): the probability to move from task T_i to task T_j was estimated by the frequency observed in the training set. The Viterbi algorithm was used for finding the most likely sequence of hidden states [27]. In the Viterbi algorithm, the conditional probability $P(state|observation)$ was defined as the average, over n, of all $P_n(T_i|n')$ values (§5.2.2).

8.2.4 Fourth Baseline Method: Without idle phase duration (PropWI)

to assess the relevance of idle phase durations, we tested the proposed solution using a single category of idle phase durations ($V = 1$, $M_1 = 1$, §5.3.3).

8.3 Offline Extension

Without the online inference constraint, the proposed algorithm can be improved by using the forward-backward algorithm [24] and by adding future-aware unary potentials, which are similar to past-aware potentials (§5.2.3), using all observations in reverse chronological order from the end of the surgery to the current observation. An accuracy of 85.3% was obtained with this offline extension.

9 Discussion and Conclusions

A novel framework for the real-time analysis of surgery videos was presented in this paper. It was applied to the automatic segmentation and categorization of cataract surgery tasks. In order to make the most of surgery video archives, the video content was analyzed using the Content-Based Video Retrieval (CBVR) paradigm, both for the online segmentation of surgical tasks and for their characterization. A novel framework, based on Conditional Random Fields (CRFs), was proposed to model the temporal evolution of the surgery.

The temporal segmentation relies on the automatic detection of 'idle phases', during which no relevant motions are detected. A sensitivity of 84.6% was achieved with a false positive rate of 3.6. Since there is an average of 28.7 idle phases per video, it means that the system detects 27.9 idle phases per videos on average $(0.846 \times 28.7 + 3.6)$. It also means that 4.4 'action phases' cannot be separated on average $(28.7 \times (1 - 0.846))$. As soon as an idle phase is detected, the system categories the action phase that just ended. Note that action phases are only defined as the complement to idle phases: multiple surgical tasks may occur during an action phase. The most likely reason would be that the system failed to detect the previous idle phase. But it is not the only reason: the surgeon may finish one task with one hand while getting ready for the next task with the other hand.

Two task categories are rather poorly detected: viscous agent injection and miscellaneous surgical tasks (see Table 1). Regarding viscous agent injection, the reason may be that the injection tool does not move during this task: the only visible motion is the flow of viscous agent, which is not properly detected due to the lack of salient points inside the viscous agent. Regarding miscellaneous tasks, the problem is that they can happen anytime during the surgery (because, by definition, this category groups several tasks together). So, the time modeling proposed in this paper may be counterproductive in this particular case. But if more data was available, then each task listed in the miscellaneous category could be treated independently. Overall, given the real-time constraints, we believe the system performance is quite high: an average area under the ROC curve of $A_z = 0.832 \pm 0.070$ was achieved.

The proposed method works significantly better than the HMM-based equivalent (p = 0.0003, see table 2). The reason probably is that CRFs are more expressive: multiple unary and pairwise potentials can be defined in place of a single probability [14]. Using idle phases to decompose the video into variable-length action phases leads to significantly higher performance than processing a uniform segmentation of the video (p = 0.0053). One reason is that the duration of a given task is highly surgeon- and even surgery-dependent. Besides, some tasks are much shorter than others: viscous agent injection is about ten times shorter than phacoemulsification. So it is hard to find a single video segment duration that is suitable for all tasks. This issue is conveniently solved using an adaptive segmentation. Using the duration of idle phases also pushes performance significantly (p = 0.0011). The reason is that idle phases tend to be longer before the beginning of a new task, because changing tools takes time. Note that the performance is not significantly different from our previous task recognition method that relied on manually-segmented action phases (p = 0.1240). But since the problem tackled is more complex in this paper, this is a positive result.

It should be noted that the offline extension of the proposed system can be used for video content structuring. Video structuring may be used to generate surgery reports at the end of each surgery [19]. For offline cataract surgery video structuring, Lalys et al. reported an average accuracy of 94% [19], which is clearly higher than the reported performance (85.3%). But it should be noted that the databases were different: Lalys et al. worked with 20 surgeries performed by three surgeons, whereas we worked with 186 surgeries performed by ten surgeons of various experience levels, including 32 surgeries with unusual tasks and 21 unusually long surgeries (longer than half an hour).

This study has one limitation: manual segmentations used for training and testing surgical task recognition was performed by a single expert, even though experts were found to disagree in 2.1% of all frames, in a subset of 10 surgery videos.

As discussed in the introduction section, the proposed video monitoring system was primarily designed to communicate information to the surgeon in due time. In particular, it was designed to send recommendations to the less experienced surgeons at the beginning of each task, or slightly in advance. One day, such an automatic video monitoring system may also be used to communicate with surgical devices directly. The idea would be, for instance, to preventively shut down a device in case of complication.

In conclusion, we have presented a system that can analyze surgery videos in real-time. The key idea is to detect idle phases and to characterize the complementary action phases by analogy reasoning. The system was successfully applied to the automatic segmentation and categorization of cataract surgery tasks, but it is general enough to be applied to other surgeries.

References

[1] B. Raju, N. S. Raju, A. S. Raju, C. P. Sudhakaran, and A. Razak, “Digital video recording and archiving in ophthalmic surgery,” *Indian J Ophthalmol*, vol. 54, no. 1, pp. 53–7, 2006.

[2] B. B. Haro, L. Zappella, and R. Vidal, “Surgical gesture classification from video data,” in *Proc MICCAI’12*, vol. 15, pp. 34–41, 2012.

[3] Z. Harris, “Distributional structure,” *Word*, vol. 10, no. 23, pp. 146–62, 1954.

[4] T. Tamaki, J. Yoshimuta, and M. K. et al., “Computer-aided colorectal tumor classification in NBI endoscopy using local features.,” *Med Image Anal*, vol. 17, no. 1, pp. 78–100, 2013.

[5] L. Zappella, B. Béjar, G. Hager, and R. Vidal, “Surgical gesture classification from video and kinematic data.,” *Med Image Anal*, vol. 17, no. 7, pp. 732–45, 2013.

[6] L. Tao, L. Zappella, G. D. Hager, and R. Vidal, “Surgical gesture segmentation and recognition,” in *Lecture Notes in Computer Science*, vol. 8151, pp. 339–46, 2013.

[7] G. Quellec, K. Charrière, M. Lamard, Z. Droueche, C. Roux, B. Cochener, and G. Cazuguel, “Real-time recognition of surgical tasks in eye surgery videos,” *Med Image Anal*, vol. 18, no. 3, pp. 579–90, 2014.

[8] L. Tao, E. Elhamifar, S. Khudanpur, G. D. Hager, and R. Vidal, “Sparse hidden markov models for surgical gesture classification and skill evaluation,” in *Proc IPCAI’12*, vol. 7330, pp. 167–77, 2012.

[9] T. Blum, H. Feussner, and N. Navab, “Modeling and segmentation of surgical workflow from laparoscopic video,” in *Proc MICCAI’10*, vol. 13, pp. 400–7, 2010.

[10] N. Padoy, T. Blum, S. Ahmadi, H. Feussner, M. Berger, and N. Navab, “Statistical modeling and recognition of surgical workflow,” *Med Image Anal*, vol. 16, no. 3, pp. 632–41, 2012.

[11] F. Lalys, L. Riffaud, D. Bouget, and P. Jannin, “An application-dependent framework for the recognition of high-level surgical tasks in the OR,” in *Proc MICCAI’11*, vol. 14, pp. 331–8, 2011.

[12] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, “A survey on visual content-based video indexing and retrieval,” *IEEE Trans Syst Man Cybern C*, vol. 41, no. 6, pp. 797–819, 2011.

[13] F. Lefèvre, “Non-parametric probability estimation for HMM-based automatic speech recognition,” *Comput Speech Lang*, vol. 17, no. 2–3, pp. 113–36, 2003.

[14] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *Proc ICML’01*, pp. 282–9, 2001.

[15] A. Fathi, A. Farhadi, and J. N. Rehg, “Understanding egocentric activities,” in *Proc ICCV’11*, pp. 407–14, 2011.

[16] Q. Shi, L. Cheng, L. Wang, and A. Smola, “Human action segmentation and recognition using discriminative semi-markov models,” *Int J Comput Vis*, vol. 93, no. 1, pp. 22–32, 2011.

[17] S. Sarawagi and W. W. Cohen, “Semi-markov conditional random fields for information extraction,” in *Proc NIPS’04*, pp. 1185–92, 2004.

[18] X. Castells, M. Comas, M. Castilla, F. Cots, and S. Alarcón, “Clinical outcomes and costs of cataract surgery performed by planned ECCE and phacoemulsification,” *Int Ophthalmol*, vol. 22, no. 6, pp. 363–7, 1998.

[19] F. Lalys, L. Riffaud, D. Bouget, and P. Jannin, “A framework for the recognition of high-level surgical tasks from video images for cataract surgeries,” *IEEE Trans Biomed Eng*, vol. 59, no. 4, pp. 966–76, 2012.

[20] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *Proc Imaging Understanding Workshop*, pp. 121–30, 1981.

[21] P. Bouthemy, M. Gelgon, and F. Ganansia, “A unified approach to shot change detection and camera motion characterization,” *IEEE Trans Circuits Syst Video Technol*, vol. 9, no. 7, pp. 1030–44, 1999.

[22] G. Quellec, M. Lamard, G. Cazuguel, B. Cochener, and C. Roux, “Wavelet optimization for content-based image retrieval in medical databases,” *Med Image Anal*, vol. 14, no. 2, pp. 227–41, 2010.

[23] R. von Mises, *Mathematical Theory of Probability and Statistics.* Academic Press, New York, 1964.

[24] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proc HLT-NAACL'03*, 2003.

[25] S. Salvador and P. Chan, "FastDTW: Toward accurate dynamic time warping in linear time and space," in *KDD Workshop on Mining Temporal and Sequential Data*, pp. 70–80, 2004.

[26] T. Lavergne, O. Cappé, and F. Yvon, "Practical very large scale CRFs," in *Proc ACL'10*, pp. 504–13, 2010.

[27] G. D. Forney, "The viterbi algorithm," *Proc IEEE*, vol. 61, no. 3, pp. 268–78, 1973.