ORB ACADEMIC PUBLISHER

Int. J. Image Video Process. Theor. App., vol. 2014, p. 1-10.
*Available online at www.orb-academic.org*

**International Journal of Image and Video Processing: Theory and Application**
ISSN: 2336-0992

# Computer-Aided Retinal Surgery using Data from the Video Compressed Stream

Zakarya Droueche[1,2], Gwénolé Quellec[2], Mathieu Lamard[2,3], Guy Cazuguel[1,2], Béatrice Cochener[4], Christian Roux[1]

[1] *Institut TELECOM/ TELECOM Bretagne, Dpt ITI, Brest, F-29200 France*
[2] *Institut National de la Santé et de la Recherche Médicale (INSERM), U650, Brest, F-29200 France*
[3] *University of Bretagne Occidentale, Brest, 29200 France*
[4] *Centre Hospitalier Universitaire Brest, Service d'Ophtalmologie, Brest, F-29200 France*

*Email: Mohammed.droueche@telecom-bretagne.eu*

**ABSTRACT**
This paper introduces ongoing research on computer-aided ophthalmic surgery. We propose a Content-Based Video Retrieval (CBVR) system for surgeons decision aid: given the video stream captured by a digital camera monitoring a surgery, the system retrieves similar annotated video streams in video archives. For comparing videos, we propose to characterize them by features extracted from compression data. First, motion vectors are extracted from the MPEG-4 AVC/H.264 compressed video stream. Second, images are segmented into regions with homogeneous motion vectors, using region growing. Third, region displacements between consecutive frames are tracked, using the well-known Kalman filter, in order to extract features characterizing region trajectories. Other features are also extracted from the residual information encoded in the MPEG-4 AVC/H.264 compressed video stream. This residual information consists of the difference between original input images and predicted images. Once features are extracted, videos are compared using an extension of the fast dynamic time warping to multidimensional time series. In this paper, the system is applied to two medical datasets: a small dataset of 69 video-recorded retinal surgery steps and a dataset of 1,400 video-recorded cataract surgery steps. In order to assess its generality, the system is also applied to a large dataset of 1,707 movie clips with classified human actions. High retrieval scores are obtained on all the three datasets.

**KEYWORDS**
Content-Based Video Retrieval (CBVR), MPEG-4 AVC/H.264 standard, region growing, Kalman Filter, Extended Fast Dynamic Time Warping (EFDTW).


## 1. Introduction

Over the last few years, Content-Based Video Retrieval (CBVR) has become a popular research topic [1, 2, 3]. CBVR analyzes the content of a query video to retrieve videos with a similar content. Video features like color, motion, texture, etc. are extracted and compared to those of other videos in a reference database to find the most similar.

Video-monitored surgery is an increasingly active research field. Several methods have been proposed to identify key surgical events [4], categorize surgical stages [5], detect surgical tools (for augmented reality purposes) [6], or finely analyze regions of interest (through image mosaicing) [7]. Content-based research techniques are expected to help surgeons in their daily practice [8, 9]. They allow surgeons to find relevant information in databases. One potential use is finding, in a reference video database, situations that are similar to an unusual situation that was recorded during the surgery.

In this paper, we focus on improving surgical procedures through CBVR. Information stored in surgical videos that resemble the current surgery monitoring video are expected to help surgeons decisions. Information stored in the associated surgery reports may also be helpful. For instance, they could let the surgeon know what a more experienced fellow worker would do in a similar situation. In that purpose, efficient, automatic search tools are needed. The purpose of the proposed CBVR framework is to analyze the current surgery monitoring video and find similar videos within digital archives (reference video databases).

The setup of the paper is as follows. Section II describes the state of the art of CBVR in the context of video-monitored

surgery. Section III describes the proposed video characterization, which involves extracting video signatures and using these signatures to compare videos. Section IV presents the video datasets used for evaluation and results. We end with a discussion and conclusion in section VI.

## 2. VIDEO MONITORED SURGERY -A STATE OF THE ART-

A few systems have been developed in the context of video monitored surgery (II-A). Some of them, designed for medical training (II-B), are Content-Based Video Retrieval (CBVR) systems.

### 2.1 Non-CBVR Systems

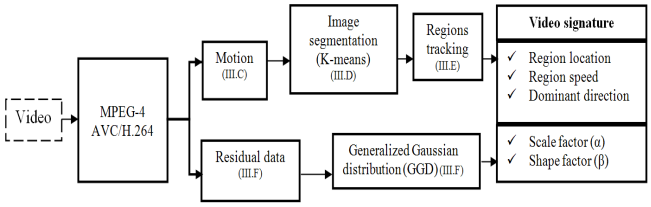


**Figure 1.** *Extraction of video signatures*

Nomany systems have been developed in the context of structure the surgical videos. There has been slow progress in developing tools allowing users to find information in databases. Cao et all introduced a new technique for surgical shot detection [5]. It is based on the detection of therapeutic instruments. First, images are segmented into region using the JSEG algorithm. Then, texture features are extracted from each region in the current image. These features are used for shot segmentation.

In [4], a framework is presented to provide a high-level representation of visual information, which reflects not only the surgery structure but also the underlying semantic and context of the in vivo environment. The goal is to facilitate surgical workflow analysis and understanding. In particular, a content-based data representation for minimally invasive surgery (MIS) data is presented. First, affine-invariant anisotropic region detectors are used. Then, surgical episodes (different events within the sequence) are identified by probabilistic tissue tracking using extended kalman filters (EKF) in the detected regions.

Some systems were designed for augmented reality purposes. A 3-D localization technique for surgical instruments from laparoscopic video sequences is proposed in [6]. It is based on the extraction of relevant 2-D information from images using Sobel edge detector. The goal is to help the development of augmented reality surgical applications. A similar method is studied in [7]. Because of the limited field of view, which can cause navigational difficulties, the tool combines many endoscopic video frames automatically. The motion field between successive frames is estimed by the optical flow. Then, image mosaics are constructed.

In [15], an eye-tracker combined with visual features were used to recognize start and end of one surgical phase in a porcine laparoscopic cholecystectomy.

### 2.2 CBVR Systems

Initially popularized in video surveillance applications [10], the importance and popularity of Content based video retrieval (CBVR) have led to several survey papers, a recent review is given in [11]. Other approaches for video browsing have been proposed in [12]. CBVR recently started developing in other applications. For instance, very few systems have addressed in the issue of medical training. A contented-based multimodal medical image and video retrieval system (CBMVR) is proposed in [13]. The objectif is to let physicians benefit from the mass data of medical images and videos already archived. Some key issues are discussed and a feature representation method named Artificial Potential Field (APF) is presented. A similar scenario is studied in [14], a content-based medical video retrieval is presented, and motion derived from the MPEG4 video stream is used for video representation.

None of these CBVR framework try to find similar videos within digital archives in reel times. In this study, a new CBVR system is presented. Given a video recorded during a surgery, it is intented to help surgeons find similar videos, in the sense that they contain similar surgical episodes. The proposed method does not rely on standard methods, such as the optical flow, to characterize motion in videos. Instead, motion, among other characteristics, is directly extracted from the compressed MPEG stream. The goal is to provide a fast video characterization. Furthermore, to allow fast similarity measurements, a fast dynamic time warping strategy was adopted.

## 3. Description of the Proposed CBVR Framework

### 3.1 Method Outline

In this paper, information is extracted from the compressed 'MPEG-4 AVC/H.264' video stream (§3.2) in order to characterize videos efficiently. Precisely, two categories of motion information are extracted from the stream at regular intervals:

- the first category describes region trajectories,
- the second category characterizes the residual error remaining after motion compensation.

These information are combined into a heterogeneous feature vector or *signature*. Because surgical videos have varying durations, a variable-length vector of signatures is extracted from each video.

To extract motion information, motion-based image sequence segmentation is performed by region growing (§3.4). To estimate region displacements between I-frames (§3.5), and therefore characterize region trajectories, the well-known Kalman filter is used. To characterize residual information (§3.6), the Generalized Gaussian Distribution (GGD) function is used. Figure 1 details the different steps for computing video signatures.

Once signatures have been extracted (§3.7), the next step is to compare videos, using their signatures. Because several features are extracted at each time interval, standard Dynamic Time Warping (DTW) or fast DTW algorithms are not suitable. Therefore, an Extension of Fast Dynamic Time Warping (EFDTW) to multidimensional time series, where each dimension represents a feature, was adopted (§3.8).

### 3.2 MPEG-4 Structure

According to the MPEG-4 AVC/H.264 international standard, the video stream is composed of Groups Of Picture (GOP), while the GOP are composed of frames. Each frame is divided into Macroblocks (MB), the fundamental unit of the video encoding process. Three types of frames are defined : I-frames (intra-compressed), P-frames (forward predicted) and B-frames (bi-directional predicted). An I-frame is encoded as a single image, with no reference to any past or future frames. Therefore, I-frames don't provide any motion information: motion is only encoded in P-frames and B-frames. A P-frame is encoded relatively to the past reference frame (either an I-frame or a P-frame). For each macroblock, MPEG encodes a motion vector that specifies the correspondences between its blocks and those of the previous frame. For B-frames, the algorithm proceeds in the same way, except that it is encoded relatively to the previous and following images.

### 3.3 Motion extraction

The Joint Model (JM) reference software [16], which has the "MPEG-4 AVC/H.264" video codec [17] implemented, is used to extract motion information. The MPEG video encoder used in this work produces one I-frame every 15 frames approximately (15 frames corresponding to a GOP). Shots (the units of the video recorder) are in accordance with the screen frequency (this is 25 frames per second). Consequently, each shot is bound to include at least one I-frame. I-frames contain the most informative data and they are included at every scene change. To save computation time, while keeping the main content in the shot, we only extract information from macroblocks in the I-frames. Motion information are extracted from each (I-frame, following P-frame) pair (see Fig. 2).

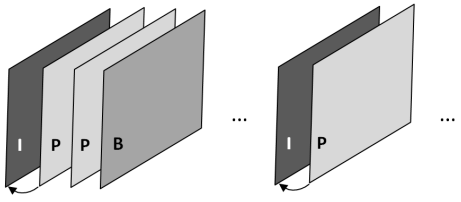


**Figure 2.** *Motion extraction from consecutive I-frames and P-frames*

### 3.4 Motion segmentation

Frames are then segmented into region of similar motion and region displacements are tracked over time. Estimating motion whithin regions offers a way to enforce motion field homogeneity, and may, to some extent, reduce Motion Vector (MV) estimation noise. A region growing strategy is used to find consistent regions. Regions motion segmentation is based on region growing and motion consistency verification using I-frames only. For each I-frame, the segmentation procedure includes the following steps [18]:

- Step 1: The group of blocks with the the most consistent MVs, within the frame, is chosen as the starting point (or seed) for region growing. The consistency of a seed is measured by the total deviation of the MVs from its centroid:

$$D_{seed} = \sum_{j \in seed} \| \overrightarrow{MV}^{seed}_{centroid} - \overrightarrow{MV}_i \| \quad (1)$$

where $seed = \{Center, Top, Left, Right, Bottom\}$ is the set of block locations in the seed pattern (see Fig. 3.a) and $\overrightarrow{MV}^{seed}_{centroid}$ is the centroid MV of the seed. The group of blocks that minimizes $D_{seed}$ is selected as initial seed. Chosing a consistent seed ensures the robustness of the entire segmentation process. Then, a region $R$ is gradually grown by clustering 4-adjacent blocks into the region, as shown in (Fig. 3.b) and described in step 2. Initially, $R = \{seed\}$.

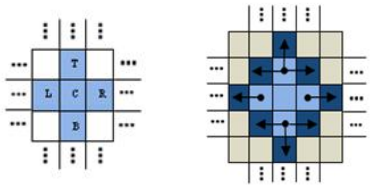


**Figure 3.** *4-adjacency blocks for clustering*

- Step 2: 4-adjacent blocks in the region's outer border are included into $R$ if their motion vector $\overrightarrow{MV}_i$ is sufficiently similar to the prevalent motion inside $R$. To decide if the motion inside a block is sufficiently similar to the region's prevalent motion, the following condition must be realized:

$$\| \overrightarrow{MV}^{R}_{Centroid} - \overrightarrow{MV}_i \| \leq D^{R}_{TH} \quad (2)$$

where $\overrightarrow{MV}^{R}_{Centroid}$ is the motion vector computed at the centroid and $D^{R}_{TH}$ is the average distance between MVs in the region and the centroid's MV, increased by the $D_{offset}$:

$$D^{R}_{TH} = E\{\| \overrightarrow{MV}^{int}_{R} - \overrightarrow{MV}^{Centroid}_{R} \|\} + D_{offset} \quad (3)$$

$$D_{offset} = min(\sigma_{MV}, T_{Max\_offset}) \quad (4)$$

$D_{offset}$ is the region growing step size used to control the speed of region growing [18]. It is related to the variance of the distances between MVs in the frame ($\sigma_{MV}$ in (4)) and to $T_{Max\_offset}$, a parameter obtained by a training procedure (3.10). A large step size will make motion segmentation converge quickly, but it might increase the risk of grouping the blocks into a wrong motion region. If equation (2) is satisfied, then $\overrightarrow{MV}_i$ is added to region $R$, otherwise $\overrightarrow{MV}_i$ is left ungrouped.

- Step 3: If $\overrightarrow{MV}_i$ was added to region $R$, then the centroid's MV is updated using the newly added $\overrightarrow{MV}_i$.
- Step 4: Steps 1 to 3 are repeated until no further seeds can be found.
- Step 5: Finally, to reduce over-segmentation, adjacent regions with similar prevalent motions are merged. Two adjacent regions are merged if the L2 distance between their centroid's MV is less than $D_{min}$. $D_{min}$ is given by the following equation:

$$D_{min} = min(\sigma_{MV}, T_{Mov\_region}) \tag{5}$$

where $T_{Mov\_region}$ is a parameter obtained by a training procedure ( §3.10).

### 3.5 Region Tracking

Region centroids estimated in consecutive I-frames (§3.3) are then associated if they have coherent motions. Regions are associated using the well-known Kalman filters (KF). KFs were chosen for tracking and estimation purposes, because they are efficient at estimating the state of a linear dynamic system [19].

KFs can match the target dynamics to give accurate estimations of the target states. The minimum mean squared error (MMSE) is used to refine these Kalman estimates. The location of the moving regions are predicted in accordance with the literature [19]. The following transition and measure equations were used to model the state system:

$$X(k+1) = FX(1) + w_k \tag{6}$$

$$Z(k+1) = HX(1) + v_k \tag{7}$$

where $X = \{x_k, y_k, \dot{x}, \dot{y}\}$ is the state of system (the estimated position of the regions) : $x_k$ and $y_k$ denote the position of the region center along the horizontal and vertical directions, $\dot{x}$ and $\dot{y}$ denote their speed. Z is the mesurement (the potential position of the regions) obtained by image segmentation at each time step.

The objective is to jointly estimate over time how many regions are present and where has it come. In order to easily calculate the similarity between videos, only the K largest regions in size are considered (Typically K = 5). The regions in one frame are predicted using KF and associated with the regions in the next frame by measuring the Euclidean distance between the predicted center and the center of any other region in the next frame. finally we associated region into nearest.

However, If region disappear during the tracking, KF associate the last appearance estimations centers with the closest region in the next time step, KF is updated to another region traking. $k$ is the time index. $F$ and $H$ are the transition and measure matrices, defined as:

$$F = \begin{pmatrix} 1 & 1 & T & 0 \\ 0 & 1 & 0 & T \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

$$H = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

where $T$ is the time interval between two consecutive frames. Finally, $v_k$ and $w_k$ are the state and measure noises, respectively. They are both assumed to be white noises.

We have seen how region trajectory information are extracted. The way these information are used to build a signature is explained in section 3.3. But first, let us see how residual information are extracted.

### 3.6 Motion-compensated residual data

Besides region trajectories, another useful kind of motion information can be extracted from the H.264 compressed video stream: motion-compensated residual data. Let (I, P) be an (I-frame, following P-frame) pair. The motion-compensated residual data is obtained by subtracting from each macroblock in P the best matching block in I. The resulting is encoded using one of the transforms specified in MPEG-4 AVC corresponding to the coding mode [13].

Several statistical models have been proposed to model motion-compensated residual data. Pao et al. modeled their distribution by zero-mean Laplacian distributions [20]. According to Wang et al., due to recent advances in motion estimation algorithms, motion-compensated residual data are very close to random Gaussian noise [21]. So, their distribution is best modeled by Gaussian distributions.

More generally, in this paper, the motion-compensated residual data are approximated by Generalized Gaussian Distributions (GGD). A GGD is defined by the following parameters:

- $\alpha$: a scale factor corresponding to the standard deviation in the Gaussian distribution.
- $\beta$: a shape parameter.

Its probability density function is defined as:

$$P(x, \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(\frac{1}{\beta})} e^{-\left(\frac{|x|}{\alpha}\right)^{\beta}} \tag{8}$$

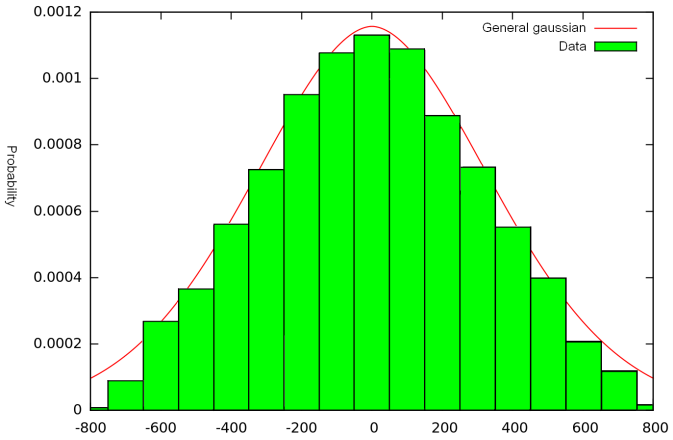


**Figure 4.** *Histogram of the motion-compensated residual data in one frame. The estimated generalized Gaussian distribution is superimposed (its parameters are* $\alpha = 486.385$ *and* $\beta = 1.95$*).*

where $\Gamma(.)$ is the gamma function: $\Gamma(z) = \int_0^\infty e^{-t} t^{z-t} dt$

When $\beta = 1$, the GGD becomes a Laplacian distribution. When $\beta = 2$, it becomes a Gaussian distribution. Within the surgical datasets used in this experiment (§4.), we found that the shape parameter $\beta$ usually lies in $[1.5, 2]$. An example is shown in Fig. 4. So, in this application, the residual data is actually better modeled with a GGD.

The GGD parameters, $\alpha$ and $\beta$, are found with a maximum likelihood estimator [23]. Let $x = (x_1, ..., x_L)$ be the residual data coefficients in one frame, where $L$ is the number of blocks per frame. Assuming $x$ has independent components, Varanasi and Aazhang [24] demonstrated that $(\alpha, \beta)$ is the unique solution of the following equations:

$$\hat{\alpha} = \left(\frac{\hat{\beta}}{L} \sum_{i=0}^{l} | x_i |^{\beta}\right)^{\frac{1}{\hat{\beta}}} \tag{9}$$

$$1 + \frac{\Psi(\frac{1}{\hat{\beta}})}{\hat{\beta}} - \frac{\sum_{i=1}^{l} | x_i |^{\beta} \, log | x_i |}{\sum_{i=1}^{l} | x_i |^{\beta}} + \frac{log(\frac{\hat{\beta}}{L} \sum_{i=1}^{l} | x_i |^{\beta})}{\hat{\beta}} = 0 \tag{10}$$

Where $\Psi(.)$ is the digamma function $\Psi(\frac{1}{\hat{\beta}}) = \frac{\Gamma(\frac{1}{\hat{\beta}})}{\Gamma(\frac{1}{\hat{\beta}})}$

### 3.7 Signatures

Region trajectory and residual data information extracted from each I-frame are now combined into an heterogeneous signature for the I-frame. The centroid of each region along the x-axis and the y-axis, its speed and the direction of the dominant displacement in the region [22] are used as region trajectory features. In order to easily calculate the similarity between videos, only the $K$ largest regions in size are considered. Due to limited motion-compensated residual data in regions, one value of motion-compensated residual data per bloc : that leads to non-convergence of the algorithm [23]. the parameters of the GGD modeling the motion-compensated residual data of the entire image, $\alpha$ and $\beta$, are affected in each regions and used as residual data regions features.

Let Q be the video sequence, a video is then represented by the following vector:

$$Q_i = < Centre_{i,k}, Speed_{i,k}, Direction_{ik}, \alpha_{ik}, \beta_{ik} >$$

where $i$ denotes the frame index and $k$ denotes the region index: $1 \leq k \leq K$.

### 3.8 Similarity measurement

To compare videos, an extension of fast Dynamic Time Warping to multidimensional time series, namely EFDTW (III.H.1), was used. In addition to the residual data signature, each time point $i$ is characterized by the set of $K$ region signatures, $1 \leq k \leq K$. Regions are compared via Fast Dynamic Time Warping.

As for entire videos, they are compared via the Earth Movers Distance (EMD) [25] using the distance between regions as basic distance.

#### 3.8.1 Extended of Fast Dynamic Time Warping (EFDTW):

Dynamic Time Warping distance was first introduced in the speech and digital processing area. It can handle sequences of unequal lengths and it allows temporal distortion. Given a reference sequence, its goal is to find the optimal sequence of deletions, suppressions, and matches that minimizes the distance between the reference sequence and the distorted sequence. Finding the optimal distortion, and therefore the minimal distance, can be formulated as a path finding problem. Let $Q$ be the query sequence and let $S$ be a reference sequence in the database: $Q = \{Q_1, Q_2, ..., Q_N\}$ and $S = \{S_1, S_2, ..., S_M\}$, $N \neq M$. Let $d$ $(M \times N)$ denote the matrix of distances between sequence elements: $d_{i,j} = distance(Q_i, S_j)$. The optimal distorsion is given by the shortest path between $d_{1,1}$ and $d_{M,N}$ [26, 27]. In the Fast Dynamic Time Warping (FDTW) extension, the search for the optimal path is constrained to a subset of the $d$ matrix, delimited by an upper and a lower bounding envelope [28]. In the proposed variation on FDTW, sequences are represented by a list of signatures, $V_Q = \{V_{Q_i}\}, 1 \leq i \leq N$, and $V_S = \{V_{S_j}\}, 1 \leq j \leq M$, rather than a list of scalars. Indeed, $V_{Q_i}$ and $V_{S_j}$ are region signatures at different time instants: $V_{Q_i} = \{V_{Q_{i,R_k}}\}$ and $V_{S_j} = \{V_{S_{j,R_l}}\}$, $1 \leq k, l \leq K$.$\{V_{Q_{i,R_k}}\}$ and $\{V_{S_{j,R_l}}\}$ consist of five signature components (§3.7) :

$$V_{Q_{i,R_k}} = \{V_{Q_{i,R_{k,c}}}, 1 \leq c \leq 5\} \tag{11}$$

$$V_{S_{j,R_l}} = \{V_{S_{j,R_{l,c}}}, 1 \leq c \leq 5\} \tag{12}$$

- An upper and a lower bounding envelope is computed for each of the five signature components in the query sequence $Q$ [28]: $Up_i(V_{Q_{i,R_{k,c}}})$ and $Low_i(V_{Q_{i,R_{k,c}}})$, $1 \leq i \leq N$, $1 \leq c \leq 5$. $Up_i(V_{Q_{i,R_{k,c}}})$ (respectively $Low_i(V_{Q_{i,R_{k,c}}})$) is the maximum (respectively the minimum) value of $V_{Q_{i',R_{k,c}}}$ for $i'$ in the $[i-r, i+r]$ interval:

$$Up_i(V_{Q_{i,R_{k,c}}}) = max\{V_{Q_{i-r,R_{k,c}}} : V_{Q_{i+r,R_{k,c}}}\} \quad (13)$$

$$Low_i(V_{Q_{i,R_{k,c}}}) = min\{V_{Q_{i-r,R_{k,c}}} : V_{Q_{i+r,R_{k,c}}}\} \quad (14)$$

Parameter $r$ is known as the warping window width [28] obtained by a training procedure (§3.10) . All the stretches allowed for $V_{Q_{j,R_{l,c}}}$ (insertions and detetions) are bound to these envelopes.

- Five matrices $d_c$ $(M \times N)$, $1 \leq c \leq 5$, are calculated: $\forall (i,j) \in 1,...,M \times 1,...,N$

$$d_c(i,j) = ||V_{Q_{i,R_{k,c}}} - V_{S_{j,R_{l,c}}}||^2 \quad (15)$$

- In order to find the optimal path between two regions, using the bounding envelopes and the matrices calculated above, the Keogh distance $Keogh\_D_c$ between $Env(V_{Q_{i,R_{k,c}}})$ and $V_{S_{j,R_{l,c}}}$ is defined as:

$$Keogh\_D_c = \quad (16)$$
$$\sum_{i=1}^{n} \begin{cases} (B - Up_i(A)^2 & if \quad B > Up_i(A) \\ 0 & if \quad B \in [Low_i(A), Up_i(A)] \\ (Low_i(A) - B)^2 & if \quad Low_i(A) > V_B \end{cases}$$

where $A = V_{Q_{i,R_{k,c}}}$ and $B = V_{S_{j,R_{l,c}}}$ [28].

- Finally the distance between two regions $R_k$ and $R_l$ is defined as the sum of all Keogh distances:

$$Keogh\_D_R = \sum_{c=1}^{5} \lambda_c Keogh\_D_c \quad (17)$$

where $\lambda_c, 1 \leq c \leq 5$, are weights between signature components.

This variation on the FDTW distance can only be used to compare two regions. Therefore, to compare two videos, each consisting of $K$ regions, we propose to use a combination of FDTW and EMD (Earth Movers Distance). The proposed extension of FDTW is referred to as EFDTW.

The EMD is based on the well-known transportation problem [25]. It is used to compute a distance between two sets of elements, when the distance between all elements is known. This problem can be solved efficiently by linear optimization algorithms that take advantage of its special structure formulation. In the EMD algorithm, $W_{Q_i}$ and $W_{S_j}$ are the weight of signatures of Q (respectevely S), In the experiments, because weight used in the equation 17, the distributions of EMD distance have equal total weight : $W_{Q_i} = W_{S_j} = 1$, $1 \leq i, j \leq N$.

Let $D$ $(K \times K)$ denote the ground distance matrix [25]. $D$ contains the distance between pair of regions in the videos to compare. The distance between two videos is given by :

$$EFDTW(Q,S) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} f_{i,j} D}{\sum_{i=1}^{N}\sum_{j=1}^{M} f_{i,j}} \quad (18)$$

where $f_{i,j} \geq 0$, the flow between $Q$ and $S$, minimizes the numerator of (19, 20) subject to the following constraints:

$$\sum_{i=1}^{N} f_{i,j} \leq W_{Q_i}, \sum_{j=1}^{M} f_{i,j} \leq W_{S_j} \quad (19)$$

$$\sum_{i=1}^{N}\sum_{j=1}^{M} f_{i,j} = min(\sum_{j=1}^{M} W_{Q_i}, W_{S_j}) \quad (20)$$

This concludes the description of the proposed system. The following sections explain how it was trained and evaluated.

### 3.9 Performance evaluation

The efficiency of CBVR systems can be evaluated by a large number of criteria. These criteria can be grouped in several classes: relevance of the results, retrieval times and flexibility. The most widely used criteria in CBVR literature are precision and recall.

In our work, we adopted the Mean Average Precision (MAP) measure. The precision of a query is defined as the number of relevant videos retrieved for this query divided by the total number of retrieved videos. The average precision is given by the following formula:

$$AveP = \frac{\sum_{l=1}^{n}(P(l, rel(l)))}{\text{number of relevant videos in the database}} \quad (21)$$

Where $l$ is the rank in the sequence of retrieved videos, $P(l)$ is the precision at cut-off $l$ in the list, $n$ is the number of retrieved videos, and $rel(l)$ is an indicator function equaling 1 if the item at rank $l$ is a relevant document, zero otherwise. All the query average precisions are averaged to obtain the MAP of the system.

Each of the three datasets (IV) was randomly divided into two sets of approximately equal size: a training and a test set. Because of the small number of videos in the E.R.M dataset (IV.A), a 2-fold cross-validation strategy was adopted in this dataset: alternatively, one of these sets was used as test set, and the other one was used as training set. To compute the MAP,

- each video in a test subset played, in turn, the role of the query video,
- the algorithm found the most relevant videos in the training subset, i.e. the videos minimizing the distance to the query video, in the training subset,
- the average precision was computed for each query in the test subset,
- finally, the average precision was obtained by averaging all precision values.

The system parameters are tuned to maximize the MAP on the training set, as described below. Therefore, during training, the training subset also plays the role of the test subset.

### 3.10 System Training

In order to maximize the MAP on the training subset, we search for the weight vector $\lambda = (\lambda_c, c = 1, ..., 5)$, defined in equation 17 (3.8.1), that maximizes the MAP on the training subset.

$\lambda$ was computed using a genetic algorithm [29] followed by a Powell's direction set ascent [30]. The search starts with a genetic algorithm, the steady state algorithm[29], with the following parameters:

- Population size = 50.
- Maximum number of generations = 20.
- Selection methods: tournament selector.
- Crossover probability = 70%.
- Mutation probability = 60%.

This configuration was adopted for its fast convergence. Once approximations of the fitness function maxima are found by the genetic algorithm, these approximations are used as initial points by Powell's method to reach the actual local maxima.

$T_{Max\_offset}$, $T_{Mov\_region}$, the clustering parameters (§3.4), are chosen experimentally, on the training subset, to allow a good segmentation. The number $K$ of regions (§3.7) is chosen experimentally, on the training subset, to ensure that only moving regions are selected. Parameter $r$, warping window width, is chosen experimentally, on the training subset, to maximizes the classification accuracy.

## 4. Video Datasets and Results

The proposed framework was applied to three video datasets: an epiretinal membrane surgery database (§4.1), a cataract dataset (§4.2) and a large movie clip dataset (§4.3).

The first two datasets were selected to validate the semantic relevance of retrieved results in medical (ophthalmic) applications.

The movie clip dataset was selected to validate the framework's generality and to allow a comparison with other methods.

**Table 1.** *Performance evaluation (Average precision)*

| Dataset | Class | Average precision |
|---|---|---|
| **E.R.M (IV.A)** | Injection | 0.459 |
| | Coat | 0.435 |
| | Vitrectomy | 0.408 |
| | **Total Average** | **0.434** |
| **C.D (IV.B)** | Miscellaneous | 0.453 |
| | Epinucleus removal | 0.515 |
| | Stitching up | 0.386 |
| | Hydrodissection | 0.653 |
| | Incision | 0.453 |
| | Implant setting-up | 0.422 |
| | Phacoemulsification | 0.383 |
| | Viscous agent removal | 0.453 |
| | Rhexis | 0.332 |
| | Idle | 0.385 |
| | Viscous agent injection | 0.395 |
| | **Total Average** | **0.439** |

| | | Proposed approach | SIFT [33] | SIFT HOG HOF [33] |
|---|---|---|---|---|
| **M.C.D (IV.C)** | AnswerPhone | 0.178 | 0.105 | 0.107 |
| | DriveCar | 0.350 | 0.313 | 0.750 |
| | Eat | 0.225 | 0.082 | 0.225 |
| | FightPerson | 0.250 | 0.081 | 0.571 |
| | GetOutCar | 0.189 | 0.191 | 0.116 |
| | HandShake | 0.523 | 0.123 | 0.141 |
| | HugPerson | 0.230 | 0.129 | 0.138 |
| | Kiss | 0.278 | 0.348 | 0.556 |
| | SitDown | 0.252 | 0.161 | 0.278 |
| | SitUp | 0.125 | 0.142 | 0.078 |
| | StandUp | 0.310 | 0.262 | 0.325 |
| | **Total Average** | **0.260** | **0.200** | **0.326** |

**Table 2.** *Retrieval times*

| Dataset | E.R.M (IV.A) | C.D (IV.B) | M.C.D (IV.C) |
|---|---|---|---|
| $Time^{(1)}$ | 7 min 03s | 7 min 03s | 7 min 03s |
| $Time^{(2)}$ | 1 min 10s | 6 min 20s | 5 min 35s |
| **Total** | 8 min 13s | 13 min 23s | 12 min 38s |

(1) time required to compute the feature vectors in a 9 minute video.

(2) time required to compute the distance with each video in the training dataset.

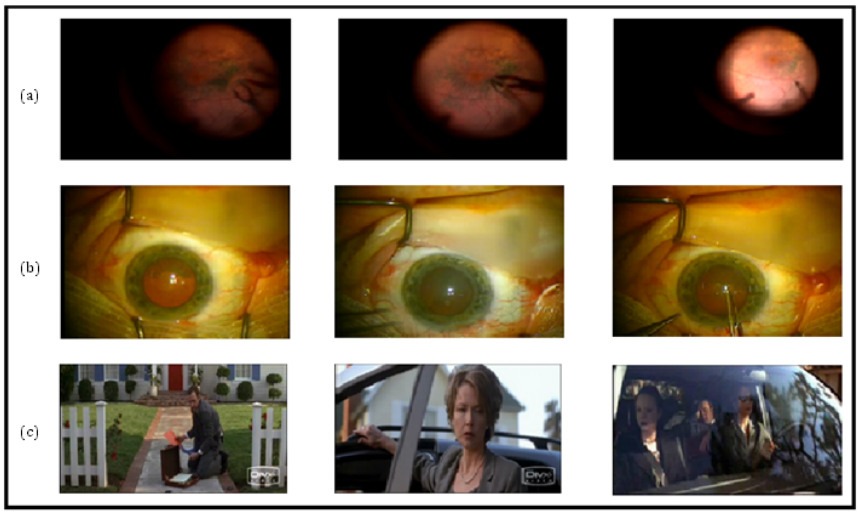


**Figure 5.** *Images from the three Dataset, (a) Epiretinal membrane surgery database, (b) Gastro-enterology dataset, (c) Movie Clip Dataset (MCD)*

### 4.1 Epiretinal Membrane Surgery Database (E.R.M)

Epiretinal membrane is a disease of the eye in response to changes in the vitreous humor or, sometimes, diabetes. It is a scar tissue-like membrane that forms over the macula, which may significantly affect the vision and create other diseases. The epiretinal membrane surgery database, collected at Brest University Hospital (France) for the purpose of this experiment contains 23 videos of epiretinal membrane surgeries. Videos have an average length of 621s (standard deviation: 299s) and images have a definition of 720x576 pixels.

Epiretinal surgery is the most commonly performed vitreo-retinal surgery, according to the Centers of Medicare and Medicaid Services [31]. An ophthalmic surgeon has divided each video into three new videos, each corresponding to one step of the membrane peeling procedure: Injection, Coat and Vitrectomy. As a result, 69 videos were obtained and each video is associated with one class (Injection, Coat or Vitrectomy).

### 4.2 Cataract Dataset (CD)

Cataract is a deterioration of the crystalline lens' optical quality. It is the leading cause of blindness worldwide and it remains an important cause of blindness and visual impairment in developed countries. There are several different types of cataracts: nuclear, cortical (spokelike), subcapsular (anterior and posterior), and mixed. Each type has its own anatomical location, pathology, and risk factors for development.

Cataract surgery is the removal of the natural lens of the eye that has developed an opacification. The cataract surgery database, collected at Brest University Hospital (France) for this experiment, contains 250 videos of cataract surgeries. Videos have an average length of 17m and 6s (standard deviation: 10m and 25s) and image definitions of 720x576 pixels. Each video was divided into eleven new videos, each corresponding to one step of the lens removal procedure. As a result, 1,638 videos were obtained and each video is associated with one class (incision, rhexis, etc.).

### 4.3 Movie Clip Dataset (MCD)

The Movie clip dataset consists of 1,707 video sequences extracted from 69 Hollywood movies: the HOLLYWOOD2 human action dataset [32]. The training set consists of 823 Video sequences; the test subset consists of 884 video sequences.

Videos have an average length of 20s. Typical image definitions include 640x352, 576x312 and 548x226. The presence of 12 types of human actions was indicated in each video sequence [32]. For each action type, a class (1= appears in, 0=does not appear in) was assigned to each sequence. Fig. 5 shows video samples from each database.

### 4.4 Results

Table I provides some evaluation results obtained on all three datasets.

High retrieval scores were measured, by the MAP, in both medical databases: 43.4% for a epiretinal membrane surgery database (§4.1), and 43.9% for a cataract dataset (§4.2).

In the Movie Clip dataset (§4.3), a MAP of 26.0% was achieved. These results were compared to those obtained in [33], using SIFT features or a the combination of three features (SIFT, HoGs and HoFs). Our results turned out to be better.

As an example, table II reports the computation time required to find similar videos in the training dataset when the query video lasts 9 minutes. All computations were performed on one core of an Intel Xeon E5520 processor running at 2.27GHz.

## 5. Discussion and Conclusions

A novel Content-Based Video Retrieval (CBVR) system was presented in this paper. In the proposed framework, motion information in medical videos (motion vectors and motion-compensated

residual data) are extracted from MPEG-4 AVC/H.264 video streams to build a video signature. This approach is advantageous in terms of computation times, compared to competing methods based on the optical flow [33], that needs full decompression of the stream. Once videos are characterized with a signature, they are compared using a novel extension of the fast dynamic time warping, which allows fast comparisons.

Experimental results on a public access database (§4.3) show that this new method significantly improves retrieval performance over existing methods [33]. High retrieval rates are also observed in two surgical datasets (see Table I): this opens new investigation trails for medical CBVR systems.

We found that the answer to a 9 minute query video is obtained in less than 9 minutes in the E.R.M dataset (§4.1). It means, for instance, that similar videos can be immediately available at the end of each surgical step. The retrieved video might therefore be useful to generate recommendations at the beginning of the next step. The system took more than 9 minutes to answer in the two other datasets. This is because computation times increases linearly with the number of comparisons (see Table II). However, the code can easily be paralelled, so retrieval times can be drastically reduced.

This framework is currently being adapted specifically to the context of surgical videos, in conjunction with a sub-sequence detector: image subsequences captured by the camera are compared to similar video sub-sequences in surgical video archives. Therefore, alarms and/or recommendations can be generated in real-time anytime during the surgery if retrieved videos shares complications with the query. In order to enhance computational efficiency, optimization of data extraction from the MPEG-4 AVC/H.264 standard and optimization of distance computation has already began. Besides, to enhance retrieval results, fusing video semantic content (clinical data) with video numerical contents is also considered.

## References


[1] E. Stringa, and C. S. Regazzoni, "Real-time video-shot detection for scene surveillance applications," *IEEE Transactions on Image Processing*, vol. 9, no. 1, pp. 69-79, Jan 2000.

[2] S. Dagtas, W. Al-Khatib, A. Ghafoor, and R. L. Kashyap, "Models for motion-based video indexing and retrieval," *IEEE Transactions on Image Processing*, vol. 9, no. 1, pp. 88-101, Jan 2000.

[3] H. Greenspan, J. Goldberger, and A. Mayer, "Probabilistic space-time video modeling via piecewise GMM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 3, pp. 384-396, Mar 2004.

[4] S. Giannarou, and G. Z. Yang, "Content-based surgical workflow representation using probabilistic motion modeling," in *LNCS Medical Imaging and Augmented Reality*, vol. 6326, pp. 314-323, 2010.

[5] Y. Cao, D. Liu, W. Tavanapong, J. Wong, J. Oh, and P. de Groen, "Computer-aided detection of diagnostic and therapeutic operations in colonoscopy videos," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 7, pp. 1268-1279, 2007.

[6] A. M. Cano, F. Gaya, P. Lamata, P. Sanchez-Gonzalez, and E. J. Gomez, "Laparoscopic tool tracking method for augmented reality surgical applications," in *LNCS*, vol. 5104, pp. 191-196, 2008.

[7] S. Seshamani, W. Lau, and G. Hager, "Real-time endoscopic mosaicking," *Medical Image Computing and Computer*, no. 9, pp. 355-363, 2006.

[8] Cao Y , Liu D, Tavanapong W, Wong J, Oh J and de Groen P. C, Automatic classification of image with appendiceal orifice in colonoscopy videos, *in Proceedings IEEE EMBC*, New York, USA, 2006, pp. 2349-2352.

[9] A. Noce, J. Triboulet, P. Poignet, Efficient tracking of the heart using texture. *In IEEE International Conference of the Engineering in Medicine and Biology Society*, Lyon, France, 2007, pp. 4480-4483.

[10] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank, "Semantic-based surveillance video retrieval," *IEEE Transactions on Image Process*, vol. 16, no. 4, pp. 1168-1181, 2007.

[11] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A Survey on Visual Content-Based Video Indexing and Retrieval," *IEEE Transactions on Systems, Man, and Cybernetics*, Part C, vol. 41, no. 6, pp. 797-819, 2011.

[12] K. Schoeffmann, F. Hopfgartner, O. Marques, L. Boeszoermenyi, and J. M. Jose, "Video browsing interfaces and applications: a review," *SPIE Reviews*, vol. 1, no. 1, pp. 018004, 2010.

[13] P. Yuan, B. Zhang, J. Li: Multi-modal Information Retrieval for Content-based Medical Image and Video Data Mining. *in Proceedings of IMAGAPP*, Roma, Italy, 2009, pp. 83-86.

[14] Z. Droueche, M. Lamard, G. Cazuguel, G. Quellec, C. Roux and B. Cochener, "Content-Based Medical Video Retrieval Based on Region Motion Trajectories," *in Proceedings of IFMBE*, 2012, vol. 37, Part 1, Part 6, pp. 622-625.

[15] A. James, D. Vieira, B. Lo, A. Darzi, G.Z. Yang, "Eye-gaze driven surgicalworkflow segmentation," In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI, Part II. LNCS, vol. 4792, pp. 110117, 2007.

[16] H.264/AVC Reference Software JM 15.1 at http://bs.hhi.de/vsuehring/tml/

[17] Video Coding Experts Group, Advanced video coding for generic audiovisual services, ITU-T Recommendation H.264, International Telecommunication Union, 2003.

[18] Yue-Meng Chen, Ivan V. Bajic, "Predictive Decoding for Delay Reduction in Video Communications," *in Proceedings of GLOBECOM*, Whashington, USA, 2007, pp. 2053-2057.

[19] Li Zhao, Quan-li Chen, "Implementation of vehicle detection and tracking based on Kalman filter," *Electronic Measurement Technology*, vol. 30, no 2, pp. 165-168, 2007.

[20] I.-M. Pao and M.-T. Sun, "Modeling DCT coefficients for fast video encoding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 4, pp. 608-616, June 1999.

[21] H. Wang, and S. Kwong, "Hybrid model to detect zero quantized DCT coefficients in H.264," *IEEE Transactions on Multimedia*, vol. 9, no. 4, pp. 728-735, June 2007.

[22] Z. Droueche, M. Lamard, G. Cazuguel, C. Roux, and B. Cochener, "Lutilisation de linformation de mouvement pour la recherche des vidos mdicales par leur contenu," *Journée de Recherche en Imagerie et Technologies de la Santé*, rennes, France, 2011.

[23] M. Lamard, G. Cazuguel, G. Quellec, L. Bekri, C.Roux, B. Cochener, "Content Based Image Retrieval based on Wavelet Transform coefficients distribution," *in Proceedings of the 29th Annual International Conference of the IEEE EMBS*, Lyon, France, August, 2007, pp. 23-26.

[24] M. Varanasi, and B. Aazhang, "Parametricgeneralizedgaussian density estimation," *Journal of the Acoustical Society of America*, vol. 86, no. 4, pp. 1404-1415, 1989.

[25] Rubner Y (1999), "Perceptual Metrics for Image Database Navigation," Ph.D. Thesis, Stanford University, USA, May 1999.

[26] S. Park, W. Chu, J. Yoon, C. Hsu, "Fast Retrieval of Similar Subsequences of Different lengths in Sequence Databases," *in IEEE International Conference on Data Engineering (ICDE)*, San Diego, San Diego, USA, 2000, pp. 2332.

[27] F. L. Hitchcock, "The distribution of a product from several sources to numerous localities," *Journal of Mathematics and Physics*, vol. 20, no. 2, pp. 224-230, 1941.

[28] C. Ratanamahatana, E. Keogh, A. J. Bagnall, S. Lonardi, "A Novel Bit Level Time Series Representation with Implication of Similarity Search and Clustering," *In Proceedings of PAKDD*, Hanoi, Vietnam, May 2005, pp.771-777.

[29] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Kluwer Academic Publishers, Boston, 1989. MA.

[30] W. H. Press, S. Teukolsky, W. Vetterling, B. Flannery, 1992b. *Numerical Recipes in C : The Art of Scientific C omputing*, Cambridge University Press, chapter 10, 1992. URL http://www.library.cornell.edu/nr/bookcpdf.html

[31] Epiretinal Membrane. Available: http://eyewiki.aao.org/EpiretinalMembrane

[32] Available: http://www.irisa.fr/vista/actions/hollywood2/

[33] M. Marszaek, I. Laptev, and C. Schmid, "Actions in context," *in Proceedings of IEEE Conference Computer Vision Pattern Recognition*, 2009, pp. 2929-2936.