34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Motion-based Video Retrieval with Application to Computer-Assisted Retinal Surgery

Zakarya Droueche, Mathieu Lamard, Guy Cazuguel, Gwénolé Quellec, Christian Roux and Béatrice Cochener

Abstract— **In this paper, we address the problem of computer-aided ophthalmic surgery. In particular, a novel Content-Based Video Retrieval (CBVR) system is presented : given a video stream captured by a digital camera monitoring the current surgery, the system retrieves, within digital archives, videos that resemble the current surgery monitoring video. The search results may be used to guide surgeons' decisions, for example, let the surgeon know what a more experienced fellow worker would do in a similar situation. With this goal, we propose to use motion information contained in MPEG-4 AVC/H.264 video standard to extract features from videos. We propose two approaches, one of which is based on motion histogram created for every frame of a compressed video sequence to extract motion direction and intensity statistics. The other combine segmentation and tracking to extract region displacements between consecutive frames and therefore characterize region trajectories. To compare videos, an extension of the fast dynamic time warping to multidimensional time series was adopted. The system is applied to a dataset of 69 video-recorded retinal surgery steps. Results are promising: the retrieval efficiency is higher than 69%.**

I. Introduction

Content analysis of video tends to be an important tool in the context of video-monitored surgery. Several methods have been proposed, Giannarou and Yang introduce a new technique to detect surgical shots [1], Cao et al. proposes a 3-D localization technique for surgical instruments (laparoscopic tools) [2], a system to classify the overall surgical procedures is given in [3], and analyze regions of interest (through image mosaicing) in [4].

In line with all these works, a novel video-monitored surgery is presented. The goal is to analyze the video stream to improving surgical procedures, 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, 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. To achieve this goal, 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 proposed video characterization, which involves extracting video signatures and using these signatures to compare videos. Section III presents the video datasets used for evaluation. Result are given in section IV. We end with a discussion and conclusion in section V.

Z. Droueche, G. Quellec, G. Cazuguel, and C. Roux are with INSTITUT TELECOM; TELECOM Bretagne; UEB; Dpt ITI, Brest, F-29200 France mohammed.droueche@telecom-bretagne.eu

M. Lamard and B. Cochener are with Univ Bretagne Occidentale, Brest, F-29200 France

All authors are with Inserm, UMR 1101, IFR 148 ScInBioS, Brest, F-29200 France

B. Cochener is with CHU Brest, Service d'Ophtalmologie, Brest, F-29200 France

II. Description of the Proposed CBVR Framework

We propose to use motion information to characterize videos. Instead of using usual methods as optical flow to compute these features, we extract them directly from the compressed MPEG-4 AVC/H.264 stream. It is very simple and fast and does not require a full decompression of the stream. We used the Joint Model (JM) reference software [5], which has the MPEG-4 AVC/H.264 video codec implemented.

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. Motion is extracted from the coding of the two other 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.

Two appraoches was adopted to caraterize our video, let us see the fist one.

A. Video caracterization based on motion histogram

In this section, we represent approach based on motion histogram to extract videos signatures.

1) Motion classification: We represent motion in videos based on motion vector histograms (one per frame) using a motion vector classification scheme given in [6] [7].

Direction of each vector $V = (x, y)$ is calculated by the folowing equation:

$$W(V) = \begin{cases} arccos\frac{x}{|V|} & if \quad y \geq 0 \\ 2\Pi - arccos\frac{x}{|V|} & if \quad y < 0 \end{cases} \tag{1}$$

Where $W(V)$ represent direction of $V = (x, y) \neq (0, 0)$ with length $\mid V \mid$. The motion direction vectors (for each macroblock) of a particular frame (See Fig. 1) are affected into motion histogram with K bins by the folowing equation:

$$Hist(V) = \begin{cases} 0 & if \quad V = (0, 0) \\ 1 + ([W(V)\frac{K}{2\Pi} + \frac{1}{2}] \text{mod} K) & if \quad \text{otherwise} \end{cases} \tag{2}$$

K = 13, represent bins of motion histogram, is chosen to allow an equidistant motion direction intervals and a separed bin ($b = 0$) (See Fig. 1).

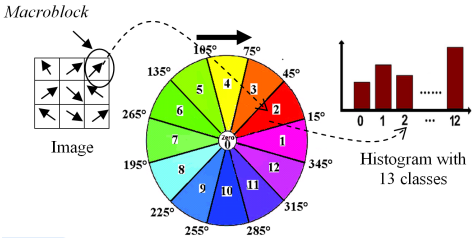

Fig. 1: Example of classification of motion vector of a macroblock

2) Signatures: Motion vectors of each frame quantified in (§II.A.1) are then used to construct signature. The number of vector of the dominant class : class that contains the highest number of vector (Direction), number of this class (Angle), and the median length of all motion vectors of the dominant class (Intensity) (see equation (3)), are extracted.

$$Median_C = \frac{1}{C}\sum_{i=1}^{C} \mid V \mid \tag{3}$$

Where C is the number of vector of the dominant class. A video is then represented by the following vector :

$$Sign_{Video} =< Direction, Angle, Intensity >$$

B. Video caracterization based region motion trajectories

The approach presented in (§II.A) caracterize motion based on motion vectors for each macroblock in the frame. This section present another approach : Motion information is segmented into regions 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. Besides, its made significant progress in reducing computation time : not all image sequences are used (§II.B.1)

1) Motion extraction: 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).

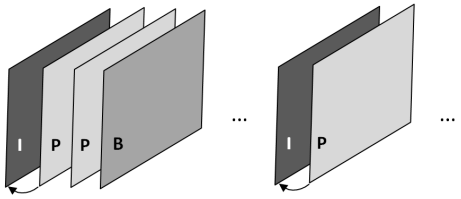

Fig. 2: Motion extraction from consecutive I-frames and P-frames

2) Motion segmentation: A region growing strategy is used to find consistent regions. Moring regions segmentation is based on a combination of k-means clustering and motion consistency verification using I-frames only. For each I-frame, the segmentation procedure can be found in [8].

3) Region Tracking: Region centroids estimated in consecutive I-frames (§B.2) 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 [9]. 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 [9]. The following transition and measure equations were used to model the state system:

$$X(k+1) = FX(1) + w_k \tag{4}$$

$$Z(k+1) = HX(1) + v_k \tag{5}$$

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. k is the time index. F and H are the transition and measure matrices, defined as:

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

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.

4) Signatures: The center of each centroid along the x-axis and the y-axis, its speed and the direction of the dominant displacement in the region (See §II.A) are used as region trajectory features. In order to easily calculate the similarity between videos, only the K largest regions in size are considered.

A video is then represented by the following vector:

$$Sign_{Video} =< Centre_{i,k}, Speed_{i,k}, Direction_{ik} >$$

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

C. Similarity measurement

To compare videos, a fast Dynamic Time Warping was used to compare signature resulting from the based motion histogram approach (See §II.A.2). Because the multidimensional time series signature presented in (§II.B.4) : where each dimension represents a region feature, an extension of fast Dynamic Time Warping namely EFDTW was adopted. First, let us see how fast Dynamic Time Warping are used.

1) Fast Dynamic Time Warping (FDTW): 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. Its goal is to find the optimal path which minimizes the distance between sequences. More details about the DTW can be found in [10] [11].

In the proposed variation on FDTW, sequences are represented by a list of signatures, $V_Q = \{V_{Q_{i,c}}\}, 1 \leq i \leq N$, and $V_S = \{V_{S_{j,c}}\}, 1 \leq j \leq M$, $1 \leq c \leq C$ = number of signature component.

- An upper and a lower bounding envelope is computed for each of the three signature components in the query sequence Q [12]: $Up_i(V_{Q_{i,c}})$ and $Low_i(V_{Q_{i,c}})$, 1 ≤ i ≤ N, 1 ≤ c ≤ C. $Up_i(V_{Q_{i,c}})$ (respectively $Low_i(V_{Q_{i,c}})$) is the maximum (respectively the minimum) value of $V_{Q_{i',c}}$ for i' in the $[i-r, i+r]$ interval:

$$Up_i(V_{Q_{i,c}}) = max\{V_{Q_{i-r,c}} : V_{Q_{i+r,c}}\} \quad (6)$$

$$Low_i(V_{Q_{i,c}}) = min\{V_{Q_{i-r,c}} : V_{Q_{i+r,c}}\} \quad (7)$$

  Parameter r is known as the warping window width [12]. All the stretches allowed for $V_{Q_{i,c}}$ (insertions and detetions) are bound to these envelopes.
- Three matrices d_c $(M \times N)$, $1 \leq c \leq C$, are calculated: $\forall (i,j) \in 1, ..., M \times 1, ..., N$

$$d_c(i,j) = ||V_{Q_{i,c}} - V_{S_{j,c}}||^2 \quad (8)$$

- In order to find the optimal path between two signatures, using the bounding envelopes and the matrices calculated above, the Keogh distance ($Keogh\_D_c$) [12] between $Env(V_{Q_{j,c}})$ and $V_{S_{j,c}}$ is defined as:

$$Keogh\_D_c = \sum_{i=1}^{n} \begin{cases} (B - Up_i(A))^2 & if \quad B > Up_i(A) \\ 0 & if \quad B \in [Low_i(A), Up_i(A)] \\ (Low_i(A) - B)^2 & if \quad Low_i(A) > V_B \end{cases} \quad (9)$$

  Where $A = V_{Q_{i,c}}$ and $B = V_{S_{j,c}}$.

2) Extended of Fast Dynamic Time Warping (EFDTW): To compare signature resulting from approach based on region motion trajectories (§II.B.4), the FDTW distance is used to compare two regions. However, 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 [13]. Let D $(K \times K)$ denote the ground distance matrix [13]. D contains the distance between pair of regions in the videos to compare. to obtain distance between videos, we used EMD equations given in [13].

D. Performance evaluation

The efficiency of CBVR systems can be evaluated by a large number of different criteria. These criteria can be grouped in several classes (relevance, retrieval time and flexibility). The more used criteria in CBVR literature are precision and recall.

The evaluation criteria we adopted is the mean precision at 5, which is the ratio between the number of pertinent videos recalled and the total videos recalled (here 5) having the same class as query video. This number (5) is the best compromise for physicians to perform a diagnosis. To compute this number,

- Each video in the database played, in turn, the role of the query video
- The algorithm found the five videos in the database minimizing the distance to the query video
- Precision was computed for each query
- Finally, the mean precision was obtained by averaging all precision values.

III. Video Datasets

The proposed framework has been applied to an epiretinal membrane surgery database.

- 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 in our laboratory 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 [14]. 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 have been obtained and each video is associated with one class (Injection, Coat or Vitrectomy).

IV. Result

TABLE I: Performance evaluation (Mean Precision at 5)

	Mean Precision at 5	
Class	**§II.A**	**§II.B**
Injection	0.712	0.625
Coat	0.653	0.588
Vitrectomy	0.709	0.672
Total Average	**0.691**	**0.628**

Table I provides some evaluation results obtained by the two approaches. High retrieval scores were measured, by the Mean Precion at 5 : 69.1% (3 to 4 videos from the first 5 videos are similar to the query video) for approach based on motion histogram (§II.A) and 62.8% for approach using regions motion trajectories (§II.B).

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

TABLE II: Retrieval times

Approaches	**§II.A**	**§II.B**
$Time^{(1)}$	17 min 03s	7 min 03s
$Time^{(2)}$	3 min 11s	1 min 10s
Total	**20 min 14s**	**8 min 13s**
$^{(1)}$ 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.		

V. Discussion and Conclusions

A novel Content-Based Video Retrieval (CBVR) system was presented in this paper. In the proposed framework, two approaches was presented, motion information in medical videos 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 optical flow, it does not require a full decompression of the stream. Furthermore, a novel extension of the fast dynamic time warping, which allows fast comparisons between videos. The best results are obtained using approach based on motion histogram created for every frame in the video (§II.A) (See Table I). This behavior was expected, because this method provides an analysis of the videos characteristics more complete than approach used I-frames only (§II.B).

We found that the answer to a 9 minute query video obtained by (§II.B) is three times faster than approach described in (§II.A) and less than 9 minutes. Furthermore, over the base is large, approach given in (§II.B) is more convenient. 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. This framework is currently being adapts to this context: image subsequences captured by the camera and compared to similar video subsequences in surgical video archives. Therefore, alarms and/or recommendations can be generated in real-time if the current surgery shares complications with already archived videos. Thats ideally suited to guide the surgery steps. To provide enhanced computational efficiency, optimization of data extraction from the standard MPEG-4 AVC/H.264, and distance computation has already began. For enhancing retrieval results, fusing video semantic content (clinical data) with video numerical contents was intended, thats allows us to enrich videos signatures.

References

[1] S. Giannarou and G.Z. Yang, Content-based surgical workflow representation using probabilistic motion modeling, in LNCS Medical Imaging and Augmented Reality, vol. 6326, 2010, pp. 314323

[2] 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 Trans Biomed Eng, vol. 54, no. 7, pp. 12681279, 2007

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

[4] S. Seshamani, W. Lau, and G. Hager, Real-time endoscopic mosaicking,in MICCAI, no. 9, 2006, pp. 355363

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

[6] Klaus Schoeffmann, Mathias Lux, Mario Taschwer, and Laszlo Boeszoermenyi, Visualization of Video Motion in Context of Video Browsing, in Proceedings of the IEEE International Conference on Multimedia and Expo, New York, USA, 2009

[7] Alvy Ray Smith, Color gamut transform pairs, SIG-XGRAPH Comput. Graph., vol. 12, no. 3, pp. 1219, 1978

[8] Yue-Meng Chen, Ivan V. Bajic (2007), Predictive Decoding for Delay Reduction in Video Communications. GLOBECOM 2053-2057

[9] Li Zhao, Quan-li Chen (2007), Implementation of vehicle detection and tracking based on Kalman filter, Electronic Measurement Technology, 30 (2), p. 165-168I.-M

[10] Park S, Chu W, Yoon J, Hsu C, (2000), Fast Retrieval of Similar Subsequences of Different lengths in Sequence Databases, 16th IEEE Int. Conf. on Data Engineering (ICDE), San Diego, San Diego, USA, p. 2332

[11] Hitchcock F. L (1941), The distribution of a product from several sources to numerous localities J. Math. Phys., 20:224230

[12] Chotirat (Ann) Ratanamahatana, Eamonn J. Keogh, Anthony J. Bagnall, Stefano Lonardi. A Novel Bit Level Time Series Representation

[13] Rubner Y (1999), Perceptual Metrics for Image Database Navigation, Ph.D. Thesis, Stanford University with Implication of Similarity Search and Clustering. In Proceedings of PAKDD'2005. pp.771 777

[14] Epiretinal Membrane at http://eyewiki.aao.org/EpiretinalMembrane