33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Real-Time Retrieval of Similar Videos with Application to Computer-Aided Retinal Surgery

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

***Abstract*— This paper introduces ongoing research on computer-aided ophthalmic surgery. In particular, a novel Content-Based Video Retrieval (CBVR) system is presented. Its purpose is the following: given a video stream captured by a digital camera monitoring the surgery, the system should retrieve, in real-time, similar video subsequences in video archives. In order to retrieve semantically-relevant videos, most existing CBVR systems rely on temporally flexible distance measures such as Dynamic Time Warping. These distance measures are slow and therefore do not allow real-time retrieval. In the proposed system, temporal flexibility is introduced in the way video subsequences are characterized, which allows the use of simple and fast distance measures. As a consequence, real-time retrieval of similar video subsequences, among hundreds of thousands of examples, is now possible. Besides, the proposed system is adaptive: a fast training procedure is presented. The system has been successfully applied to automated recognition of retinal surgery steps on a 69-video dataset: areas under the Receiver Operating Characteristic curves range from A_z=0.809 to A_z=0.989.**

I. INTRODUCTION

Automated analysis of video content, in the context of video-monitored surgery, is an increasingly active research field. Several methods have been proposed to identify key surgical events [1], categorize surgical stages [2], detect surgical tools (for augmented reality purposes) [3], or finely analyze regions of interest (through image mosaicing) [4]. In line with all these works, a study has been initiated at the LaTIM laboratory to setup an alarm/recommendation generation system for video-monitored surgery. The goal is to analyze the video stream in real-time and warn the surgeon whenever a risky situation is detected (*alarm generation*) or let the surgeon know what a more experienced fellow worker would do in a similar situation (*recommendation generation*).

To achieve this goal, we focused on the Content-Based Video Retrieval (CBVR) paradigm. The purpose of CBVR systems is to find, within digital archives, videos that resemble a query video. In CBVR, similarity between videos relies on motion, shape, texture, or color analysis. Initially popularized in video surveillance applications [5], CBVR recently started developing in other applications. For instance, its use for medical training is now considered [6]. CBVR systems typically accept a video *file* as input, and display similar video *files* on output [7]. In that sense, typical CBVR systems generalize Content-Based Image Retrieval systems [8]. A more ambitious scenario is considered in this paper: we propose to analyze, in real-time, the video stream captured by a digital camera and constantly search similar video subsequences in digital video archives. The search results can be used to generate alarms or recommendations whenever necessary.

When searching for similar video subsequences, and not simply video files as a whole, the number of items that should be compared to the query item explodes. In order to meet the real-time constraint, a very fast similarity measure must therefore be used to compare video subsequences. In particular, the use of temporally flexible, but slow, distance measures (such as *Dynamic Time Warping* [9], [7]) is prohibited. However, temporal flexibility is required to cope with speed differences among surgeons. An alternative solution is proposed in this paper: temporal flexibility is directly introduced in the way video subsequences are characterized.

II. VIDEO SUBSEQUENCE CHARACTERIZATION

A. Video Subsequences

Variable-length *video sequences* are considered in this paper: typically, each video sequence depicts a surgery step (§IV). From each video sequence V, several fixed-length *video subsequences* S_i were extracted. S_i consists of N consecutive images from V: $S_i = \left\{I_1^i, I_2^i, ..., I_N^i\right\}$. Note that two consecutive subsequences overlap: $I_n^i = I_{n-1}^{i+1}$.

In order to include temporal flexibility in the way subsequences are characterized, images in S_i were organized in M *equivalence groups* $G_m^i \subset S_i$, $m = 1..M$ (see Fig. 1); within an equivalence group, temporal order was ignored. Note that one image may belong to several equivalence groups (see Fig. 1 (c)).

To characterize a video subsequence S_i, each image $I_n^i \in S_i$ was first characterized individually (§II-B). Then, these characterizations were combined by equivalence groups and finally by video subsequence (§II-C).

B. Characterizing one Image in the Subsequence

In order to characterize each image $I_n^i \in S_i$, texture and color features were extracted from I_n^i and motion features were extracted from the optical flow between I_{n-1}^i and I_n^i.

Texture and color features were extracted from the wavelet transform of each color channel of I_n^i: for each color channel, the distribution of the wavelet coefficients was characterized

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

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

All authors are with Inserm, U650, IFR 148 ScInBioS, Brest, F-29200 France

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

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

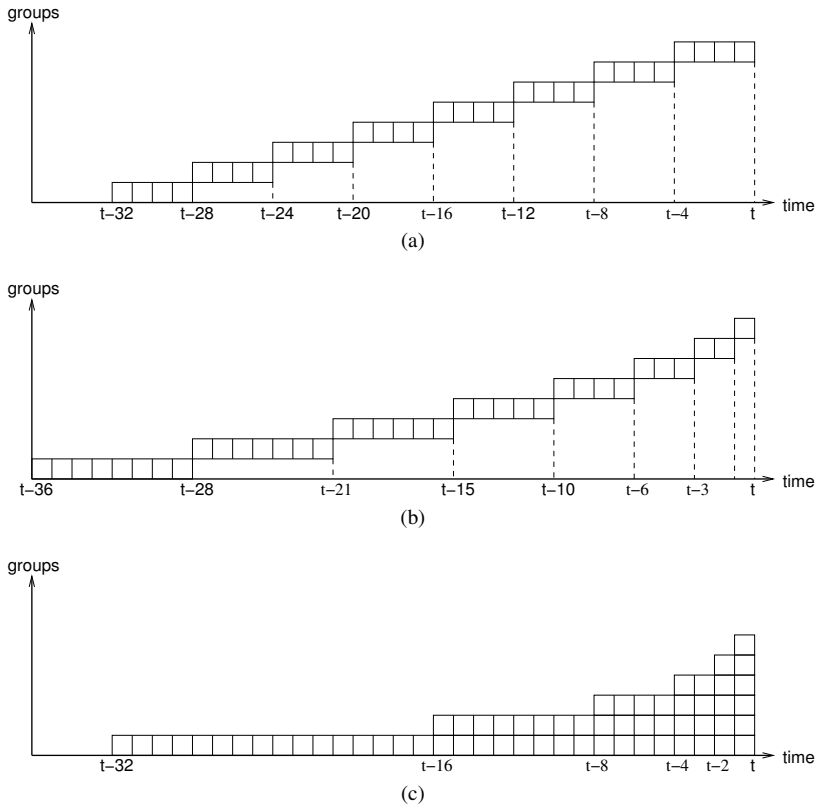


Fig. 1: Examples of temporal setups. In examples (a) and (c) (resp. (b)), each subsequence consists of $N = 32$ (resp. $N = 36$) images. In examples (a) and (b) (resp. (c)), each subsequence consists of $M = 8$ (resp. $M = 6$) equivalence groups.

by a parametric model, as described in a previous paper from our group [8] (that paper was about still image retrieval).

To extract motion features, strong corners were first detected in I^i_{n-1}. These corners were selected, among all image pixels p, with respect to the smallest eigen value of matrix M_p below:

$$\begin{cases} M_p = \begin{pmatrix} A_p & B_p \\ B_p & C_p \end{pmatrix} \\ A_p = \sum_{(x,y)\in\mathcal{N}_p} \left(\frac{dI^i_n}{dx}(x,y)\right)^2 \\ B_p = \sum_{(x,y)\in\mathcal{N}_p} \frac{dI^i_n}{dx}(x,y) \cdot \frac{dI^i_n}{dy}(x,y) \\ C_p = \sum_{(x,y)\in\mathcal{N}_p} \left(\frac{dI^i_n}{dy}(x,y)\right)^2 \end{cases} \quad (1)$$

where $\mathcal{N}_p$ is a neighborhood of pixel p. Then, the optical flow between I^i_{n-1} and I^i_n was computed at each strong corner by the Lucas-Kanade iterative method [10]. The OpenCV[1] library was used to select strong corners and compute the optical flow. Motion was characterized by an amplitude histogram, an amplitude-weighted spatial histogram and an amplitude-weighted directional histogram.

Let $\boldsymbol{f}^i_n$ be the feature vector containing all features extracted from I^i_n and the optical flow between I^i_{n-1} and I^i_n.

C. Characterizing the Subsequence

Each equivalence group $G^i_m \subset S_i$ was characterized by the mean of all feature vectors $\boldsymbol{f}^i_n$ such that $I^i_n \in G^i_m$: let $\breve{\boldsymbol{f}}^i_m$ be that average feature vector. As for subsequence S_i, it was initially characterized by the concatenation of all $\breve{\boldsymbol{f}}^i_m$ vectors, $m = 1..M$: let $\hat{\boldsymbol{f}}^i$ be that compound feature vector. Note that equivalence groups in a subsequence are likely to be correlated. It follows that feature vectors $\breve{\boldsymbol{f}}^i_m$ also are. In order to obtain more compact feature vectors, with less correlated components, a principal component analysis of all vectors $\hat{\boldsymbol{f}}^i$ in a training set was performed [11]. Then, $\hat{\boldsymbol{f}}^i$ was replaced by its projection $\bar{\boldsymbol{f}}^i$ on the C principal components[2].

Remember that two consecutive subsequences from the same sequence overlap (§II-A). Therefore, to characterize subsequence S_i, only the last image $I^i_N \in S_i$ actually needs to be characterized (§II-B): a FIFO queue was used to store the last $N-1$ image characterizations in memory.

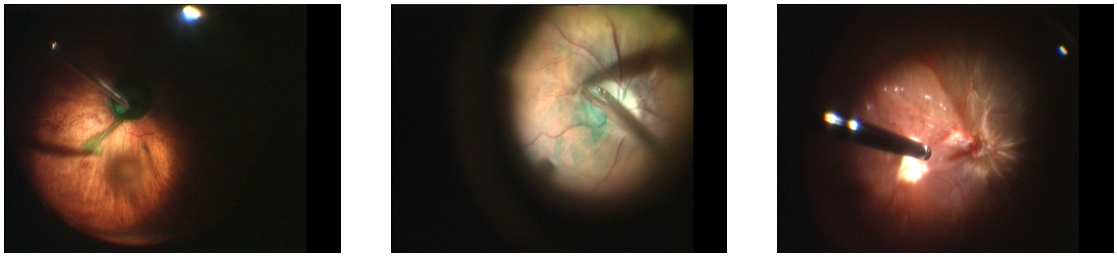
III. Similar Video Subsequence Retrieval

A. Real-time Comparison of Subsequence Characterizations

Working with fixed-length characterizations ($\bar{\boldsymbol{f}}^i$) has one major advantage: these characterizations can be sought with

[1] http://opencv.willowgarage.com/wiki/

[2] C was chosen such that 90% of the energy is preserved.



(a) Injection (b) Coat (c) Vitrectomy

Fig. 2: Images from Brest Retinal Surgery Dataset

very fast search engines, such as k-d trees [12] or Locality-Sensitive Hashing [13]. ANN[3], a fast variation on k-d trees, was used in this study. In ANN, feature vectors are compared with the squared Euclidean distance.

In order to fill the semantic gap between low-level characterizations and the high-level concept of *semantic distance*, each component $\bar{f}_c^i$ of feature vector $\bar{\boldsymbol{f}}^i$ was weighted by $\lambda_c \geq 0$, $c = 1..C$.[4] Because training video sequences (i.e. video sequences in the training set) have only been interpreted by experts as a whole, weight adjustment couldn't be supervised at the *video subsequence* level. This problem was addressed by adjusting weights at the *video sequence* level (§III-C). The semantic distance between two training video sequences U and V was defined as follows: $DS(U,V)$=0 if U and V belong to the same class, $DS(U,V)$=1 otherwise. A low-level distance $DN(U,V)$ was defined: the computation of $DN(U,V)$ derives from $\{\lambda_c, c = 1..C\}$ and all feature vectors $\bar{\boldsymbol{f}}^i$ extracted from U and V (§III-B).

B. Comparing Video Sequence Characterizations

For each feature vector component $c = 1..C$, a partial low-level distance $DN_c(U,V)$ was defined to compare two video sequences U and V. $DN_c(U,V)$ was defined as the maximal deviation between the Cumulative Distribution Function (CDF) of $\left\{\bar{f}_c^i / S_i \in U\right\}$ and the CDF of $\left\{\bar{f}_c^j / S_j \in V\right\}$. In other words, $DN_c(U,V)$ is the Kolmogorov-Smirnov statistic of the equality test of $\left\{\bar{f}_c^i / S_i \in U\right\}$ and $\left\{\bar{f}_c^j / S_j \in V\right\}$ [14].

C. Feature weighting

Let T be the number of video sequences in the training set. Let $T' = \frac{1}{2}T(T-1)$ be the number of pairs of video sequences. For each video sequence pair (U,V), one semantic distance $DS(U,V)$ and C partial low-level distances $DN_c(U,V)$ (§III-B) were computed. Semantic distances were grouped together in a vector $\boldsymbol{ds}$ of size T'. Low-level distances were grouped together in a matrix $\mathcal{DN}$ of size $(T' \times C)$. The weight vector $\boldsymbol{\lambda} = \{\lambda_c, c = 1..C\}$ minimizing the sum of the squared errors between $\boldsymbol{ds}$ and $\mathcal{DN} \cdot \boldsymbol{\lambda}$ was computed with the multi-parameter linear fitting function implemented in the *GNU Scientific Library*[5]. Whenever a negative weight λ_c was obtained, it was replaced by its absolute value; a more mathematically acceptable solution would consist in adding a positivity constraint in the fitting process, at the cost of increased computation times. Once computed, these weights were used for real-time retrieval of similar video subsequences (§III-A).

IV. Brest Retinal Surgery Dataset

The proposed framework has been applied to a retinal surgery dataset collected at Brest University Hospital (France). This dataset consists of 23 videos, each depicting one Epiretinal Membrane Surgery (EMS) performed by a retina surgeon. EMS is the most common vitreoretinal surgery[6]. It involves a pars plana vitrectomy procedure with membrane peeling (see Fig. 2). Videos have an average length of 621s (standard deviation: 299s) and images have a definition of 720x576 pixels. About 350,000 video subsequences were extracted from the entire dataset, as explained in previous sections.

Retrospectively, the surgeon has divided each video into three new video sequences, each corresponding to one step of the EMS: Injection, Coat and Vitrectomy. As a result, 69 video sequences have been obtained. For each EMS step, a class (1=*"corresponds to"*, 0=*"does not correspond to"*) has been assigned to each of these 69 sequences.

V. Experiment

For each surgery step, and for each temporal setup shown in Fig. 1, performance has been assessed in terms of A_z, the area under the Receiver Operating Characteristic (ROC) curve. A 2-fold cross-validation strategy was adopted: the 23 surgeries were divided into two sets of approximately equal size. Alternatively, one of these sets was used as test set, and the other one as training set.

First, a weight vector has been adjusted on the training set (§III-C). Second, for each video subsequence in the test set, the five most similar subsequences in the training set have been retrieved, as described in section III-A. Third, for each video sequence V in the test set, the probability p_V that V belongs to class 1 has been computed. p_V was defined as

[3]http://www.cs.umd.edu/ mount/ANN/

[4]It amounts to weighting the c^{th} term in the squared Euclidean distance by λ_c^2.

[5]http://www.gnu.org/software/gsl/

[6]http://eyewiki.aao.org/Epiretinal_Membrane

TABLE I: Performance evaluation (A_z) on the test set

temporal setup (see Fig. 1)	(a)	(b)	(c)
Injection	0.870	0.842	**0.901**
Coat	0.977	0.977	**0.989**
Vitrectomy	**0.809**	0.778	0.793

the percentage, across all subsequences of V, of retrieved video subsequences coming from a training video sequence in class 1. Finally, the A_z has been estimated using all p_V values. Results are presented in table I. In all experiments, search times (subsequence characterization + search itself) were less than $\frac{1}{25}$ seconds.

VI. Conclusion

A novel Content-Based Video Retrieval (CBVR) framework, allowing retrieval of video *subsequences*, has been presented in this paper. By introducing temporal flexibility in the way video subsequences are characterized, the use of flexible distance measures, such as *Dynamic Time Warping* [9], has been avoided. Transferring flexibility from the distance measure to the characterization allowed real-time retrieval ($< \frac{1}{25}$ seconds) of similar video subsequences among hundreds of thousands of video subsequences. Because the proposed distance measure is adaptive, retrieval was not only fast but also semantically relevant (see table I). It can be seen that the optimal temporal setup depends on the surgery step considered. After adapting the time scale and the temporal setup, the proposed framework could be applied to computer-aided video-monitored ophthalmic surgery. This framework is indeed ideally suited to this context: image subsequences captured by the camera can be constantly compared to similar video subsequences in surgical video archives. Therefore, alarms and/or recommendations can be generated whenever needed. A larger surgical video dataset is currently being collected and interpreted at Brest University Hospital; it will allow such an experiment in future works.

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. 314–323.

[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. 1268–1279, 2007.

[3] A. M. Cano, F. Gayá, P. Lamata, P. Sánchez-González, and E. J. Gómez, "Laparoscopic tool tracking method for augmented reality surgical applications," in *LNCS*, vol. 5104, 2008, pp. 191–196.

[4] S. Seshamani, W. Lau, and G. Hager, "Real-time endoscopic mosaicking," in *MICCAI*, no. 9, 2006, pp. 355–363.

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

[6] B. André, T. Vercauteren, A. M. Buchner, M. W. Shahid, M. B. Wallace, and N. Ayache, "An image retrieval approach to setup difficulty levels in training systems for endomicroscopy diagnosis," in *MICCAI*, vol. 13, 2010, pp. 480–487.

[7] D. Xu and S. F. Chang, "Video event recognition using kernel methods with multilevel temporal alignment," *IEEE Trans Pattern Anal Mach Intell*, vol. 30, no. 11, pp. 1985–1997, 2008.

[8] 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. 227–241, 2010.

[9] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans Acoust Speech Signal Process*, vol. 26, no. 1, pp. 43–49, 1978.

[10] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc Imaging Understanding Workshop*, 1981, pp. 121–130.

[11] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philos Mag*, vol. 2, no. 6, pp. 559–572, 1901.

[12] S. Arya and D. M. Mount, "Approximate nearest neighbor queries in fixed dimensions," in *Proc ACM-SIAM Symposium on Discrete Algorithms*, 1993, pp. 271–280.

[13] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *Proc 25th Very Large Database (VLDB) Conference*, 1999, pp. 518–529.

[14] R. von Mises, *Mathematical Theory of Probability and Statistics*, H. Geiringer, Ed. Academic Press, New York, 1964.