Automatic View-Point Selection for Inter-Operative Endoscopic Surveillance

Anant S. Vemuri, Stéphane Nicolau, Jacques Marescaux, Luc Soler, Nicholas Ayache

▶ To cite this version:

HAL Id: hal-01203463
https://hal.inria.fr/hal-01203463

Submitted on 16 Oct 2015

Automatic View-Point Selection for Inter-Operative Endoscopic Surveillance

A.S. Vemuri[1,2], S.A. Nicolau[3], J. Marescaux[3], L. Soler[3], and N. Ayache[2]

[1] IHU, 1 Place de l'Hopital, 67091 Strasbourg Cedex, FRANCE
[2] INRIA Sophia Antipolis, 06902 Sophia Antipolis Cedex, FRANCE
[3] IRCAD, 1 Place de l'Hopital, 67091 Strasbourg Cedex, FRANCE
anant.vemuri@inria.fr

Abstract. Esophageal adenocarcinoma arises from Barrett's esophagus, which is the most serious complication of gastroesophageal reflux disease. Strategies for screening involve periodic surveillance and tissue biopsies. A major challenge in such regular examinations is to record and track the disease evolution and re-localization of biopsied sites to provide targeted treatments. In this paper, we extend our original inter-operative relocalization framework to provide a constrained image based search for obtaining the best view-point match to the live view. Within this context we investigate the effect of, (a) the choice of feature descriptors and color-space, (b) filtering of uninformative frames, (c) endoscopic modality, for view-point localization. Our experiments indicate an improvement in the best view-point retrieval rate to $[92\%, 87\%]$ from $[73\%, 76\%]$ (in our previous approach) for NBI and WL.

1 Introduction

The incidence of esophageal adenocarcinoma has risen dramatically over the past three decades in western countries. Adenocarcinoma of the esophagus appears to arise from the Barrett's muscosa through progressive degrees of dysplasia [1]. The guidelines [2] prescribe different levels of surveillance intervals depending on the degree of dysplasia and the cost-effectiveness. One of the key challenges to regular surveillance is re-localization of biopsy sites between two surveillance procedures.
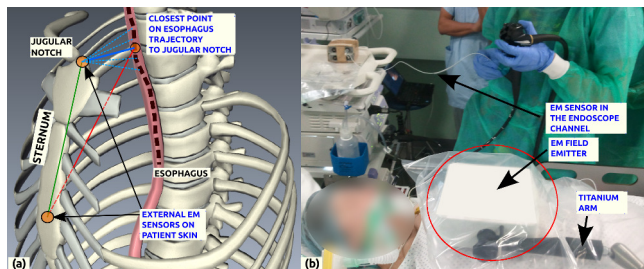

Fig. 1: System setup

In an earlier work [3] we provided a first approach to inter-operative video sychronization using an electromagnetic tracker (EMTS). Fig. (1) shows the system setup. In our framework we used an EM sensor inside the endoscope channel to track its position inside the esophagus and two external sensors for providing the anatomical landmarks on the patient. We performed simultaneous capture from EMTS and the endoscopic frame to generate a database where, each captured image has a corresponding 3D position associated with it. Firstly we performed inter-operative registration using external sensors on the patient. Then to provide inter-operative video synchronization, we used the 3D position obtained from the EM sensor to determine the nearest neighbour (EMNN). The corresponding image for the EMNN provided the localized view in the esophagus. We define this as gross-localization in the context of our problem. From the clinical point of view, this can be applicable to the gastroenterologist (GIS) in two ways. (i) *Differential Surveillance* (DS): By juxtaposing the synchronized view with the live view, it provides the GIS, the corresponding variations at the same location in the esophagus. (ii) *Biopsy Site Relocalization* (BSR): In an alternate scenario, when only the biopsy sites are stored in the database, the 3D position of the sensor in the endoscope would allow the GIS for relocalizing the biopsy sites.

The problem of BSR has been addressed in our earlier work [4]. The goal of this paper is to provide a constrained approach to scene association for DS. The EMNN does not necessarily provide the best view-point from GIS's perspective. This could be due to a combination of three reasons: *a*) The matched view could vary from the live view in the esophagus fig. 2(b)-2(f), *b*) The matched view could be an uninformative (UI) frame as shown in fig. 2(g)-2(k). *c*) In [4] we have evaluated the influence of uncertainty of the placement the external markers. Essentially, it leads to an average depth estimation error along the esophagus of $\pm 10mm$ for a 95% confidence interval. It is thus important to provide an intelligent selection of the best matching view using the images in the neighbourhood of the gross-localized region.
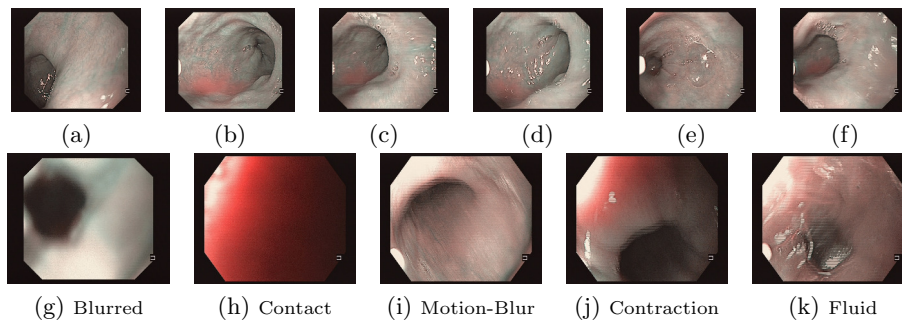

Fig. 2: *k*-EMNN matches. (a): Query, (b)-(f): matched frames according to EMTS with scores: $[0, 1, 0, 2, 1]$ as assinged by the expert section. 3. (g)-(k) present the sample UI frames.

   We outline the framework of our approach, by classifying the differential surveillance task into three stages:

1. *Gross-localization*: Computing the nearest neighbour from the 3D position obtained from EMTS, we obtain an approximate location of the endoscope in the esophagus.
2. *View-point localization*: Considering the k-EMNN matches within a chosen search radius as shown in fig. (3), we extract descriptors to represent the scene and obtain the best matching viewpoint to the live view. The value of k varies depending on the number of EMNNs found within the chosen search radius from the closest EMNN. It is important to note that for view-point localization a corresponding matching view that closely resembles the live view may not always be available. However it is important to be able to provide the "best available" matching view.
3. *Inter-frame mapping*: Finally after the best view-point image has been obtained, regions of interest (and tagged biopsy sites) can be mapped from the matched image to the live frame for re-targeting.
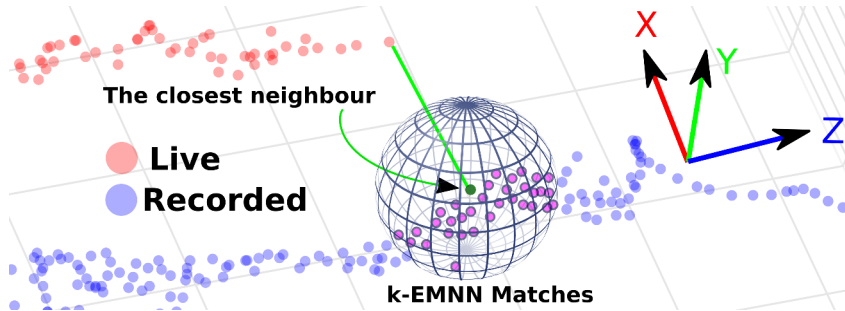

Fig. 3: k-Nearest neighbour electromagnetic tracker matches

   This paper focuses on the view-point localization problem which is closely related to scene classification and recognition, that has received considerable attention in computer vision [5]. Endoscopic scene classification has been applied for automatic detection of pathological conditions [6]. In the scope of our work, [7] had proposed training locality preserving projections for low-dimensional embedding of images from a single intervention for localization in subsequent interventions. Our contributions are; firstly, in establishing the need for gross-localization. Secondly, an approach to filtering of UI frames. And thirdly, an evaluation of various descriptor-color-space combinations for narrow band imaging (NBI) and white light (WL) endoscopic modalities (commonly used in clinical practice), to prove the validity of our approach. Section. 2.1 presents the various descriptors for endoscopic scene description. Section. 2.2 presents our approach to removal of UI frames. Section. 3 describes the data collection and assignment of ground-truth for both scene recognition and UI frame classification with their results. Section. 4 lists the important observations.

2   Methods

2.1   Feature Descriptors and Matching

Table 1 presents a summary of various descriptors used in this study. These descriptors were computed in RGB, HSV, Gray-Scale (GS), normalized RGB (norm), log and opponent color-spaces (chosen appropriately for each descriptor). It should be noted that we are not bound by the need for rotation invariance of the descriptors since, in our earlier work [4], we had established a way to perform orientation correction between matched images using the 6-*dof* EM sensor information.

ID	Descriptor
mLBP[8]	Multi-scale Local Binary Patterns. At each level the pyramid, the image was divided into non-overlapping cells. A LBP feature vector was computed for each cell which was concatenated into a large feature vector representing the image.
mHOG[9]	Multi-scale Histogram of Oriented Gradients. Approach similar to mLBP but with a HOG descriptor for each cell.
sw-mLBP	Sliding window mLBP. Similar to mLBP, with each cell is a region within a sliding window over the image.
mLTP[10]	Multi-scale Local Ternary Patterns.
mLBP+mHOG	A combined mLBP and mHOG descriptor.
dSIFT[11]	Dense scale invariant feature transform. A fast variant computed on non-overlapping cells for each image.
mLIOP[12]	Multi-scale Local Intensity Order Pattern.

Table 1: Summary of feature descriptors evaluated in this study. For multi-scale approaches a scale space image pyramid was constructed.

2.2   Uninformative Frame Removal

A typical endoscopic exploration lasts for several minutes and may include many UI frames fig. 2(g)-2(k). Several approaches have been proposed for removal of UI frames [13]. However most have focused on images from capsule endoscopy (CE). Unlike in CE, an endoscopic procedure is shorter, there is insufflation during the procedure and the type of UI frames encountered are different. In our approach, for each image (in GS) we computed a mLBP descriptor. Using the strategy proposed in [14] for data selection; k-means was applied to obtain the representative cluster centers for the individual classes. Dimensionality reduction was performed for the descriptors from the selected samples using PCA and the resulting feature vectors were used for training. The basis vectors from PCA were then used to project all the image descriptors, shown in fig. (4). An RBF-kernel SVM classifier was trained using LIBSVM [15]. First we perfomed leave-one-out cross validation using all the descriptors for parameter selection, where data from k-1 interventions was used for training and the kth intervention for testing. These parameters were then used for multiple iterations of data selection (described earlier) to obtain the best model and the average classification rate.

3   Experiments and Results

We collected data from 7 human subjects, with two surveillance procedures per subject. Between each surveillance procedure the patient underwent gastric treatment, and biopsies were taken for analysis. An Olympus gastroscope was used with WL and NBI modalities. From each recorded trajectory 9 equally spaced query locations were selected to cover about 25cm along the esophagus length. For each selected query locations k-EMNN with increasing search radii (10mm-70mm), as shown in fig. (3), were obtained. Hence, for 7 pairs of surveillance endoscopies a total of 63 query locations (each for NBI and WL) were selected. The GIS reviewed the k-EMNN matches obtained for each of the search radii and scored the matched images as, 2 - best match, 1 - partial match and 0 - incorrect match. Although these are subjective scores, they help quantify the ideology and the approach to feature based matching in the choice of a good view-point. For each query frame and k-EMNN frames, descriptors were computed and matched using chi-squared distance metric. We compared 36 descriptor-color-space combinations. Table 2 presents results for the 9 best combinations for NBI and WL respectively. The feature descriptors used for this paper are variations of those provided in [16]. An important criteria in selection of these descriptors is also the computation time. Each of these can be applied in a real-time scenario for online classification.

For training the classifier of UI frames, the GIS reviewed the images from 10 NBI and 8 WL surveillance interventions. A score of 2 for informative frame, 1 for partially informative and 0 for UI frame were assigned. For the purposes of this paper, we decided to combine the first two classes together. A total of 4236 NBI frames and 2643 WL frames were tagged. The precision and recall, for NBI $= [98\%, 93\%]$ and WL $= [97\%, 88\%]$. The average scores for EM based match for NBI and WL improved from $[0.97, 0.82]$ to $[1.2, 1.2]$ after filtering the UI frames. Finally, fig. (7) shows the best matches from EM and imaged based approach.

4   Discussion and Conclusion

This paper extends our earlier work on inter-operative synchronization to include an important aspect of view-point localization. We compared various descriptors used for texture and scene classification discussed in the literature and also presented an alternative approach to UI frame removal. To our knowledge this is the first paper that provides such a comparison over different endoscopic modalities. Most importantly, the presented framework allows for quantitative evaluation of inter-operative view-point matching which is an important aspect of differential surveillance. We observed that, (i) the performance is better on NBI than on WL, which is expected because of the higher texture in NBI modality. (ii) Fig. (5) depicts the general trend of decreasing avg. score with increasing search radius, indicating the need for constraining the search space using gross-localization, whereas in [7] only temporal localization of frames was considered. (iii) Fig. (6) clearly shows that filtering the uninformative frames reduces the number of false

matches. This is however more observable in WL than in NBI. (iv) Using GS for NBI images is not meaningful as observed from Table 2, because it is not a true mapping from RGB space. (v) Table 2 also shows that texture based descriptors such as LBP, LIOP and LTP are much better suited in this scenario. Along with the choice of illumination invariant color-spaces such as hsv, norm and log; variations of these must be explored further.
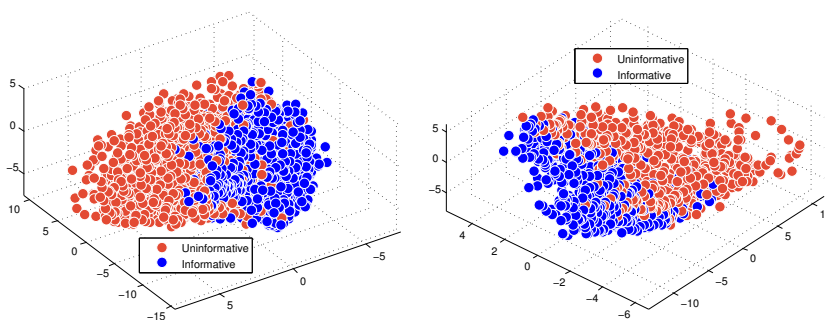

Fig. 4: The first three dimensions of the projected feature vectors, NBI (left) and WL (right), on to the trained basis vectors from PCA.
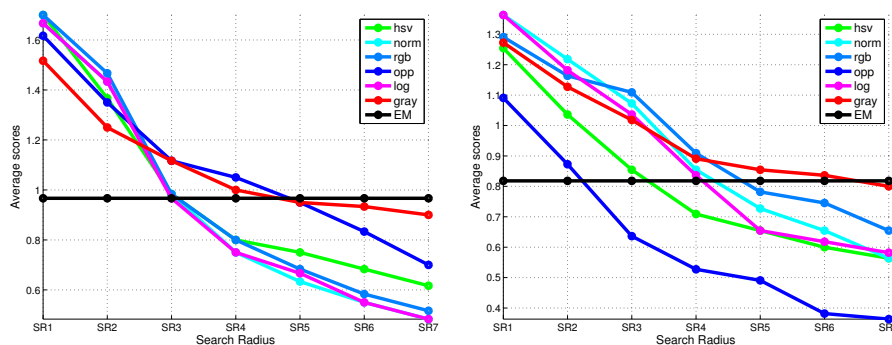

Fig. 5: Avg. scores of mLBP descriptor for six color-spaces over increasing search radii (10mm to 70mm). NBI (left) and WL (right).

References

1. V. Conteduca, D. Sansonno, G. Ingravallo, S. Marangi, S. Russi, G. Lauletta, and F. Dammacco. Barrett's esophagus and esophageal cancer: an overview. *International journal of oncology*, 41(2):414–424, 2012.
2. K. K Wang and R. E Sampliner. Updated guidelines 2008 for the diagnosis, surveillance and therapy of Barrett's esophagus. *The American journal of gastroenterology*, 103(3):788–97, March 2008.
3. A. S. Vemuri, S. A Nicolau, N. Ayache, J. Marescaux, and L. Soler. Interoperative trajectory registration for endoluminal video synchronization: Application to biopsy site re-localization. In *MICCAI*, pages 372–379. Springer, 2013.
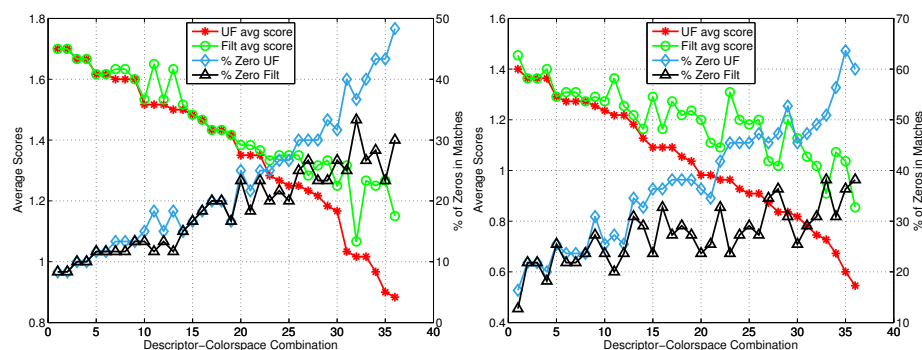
Fig. 6: Avg. scores on the left axis and % of images matched with score zero on the right axis for all 36 of the descriptor-color-space combinations. The x-axis is sorted by the best avg. score. NBI (left) and WL (right).
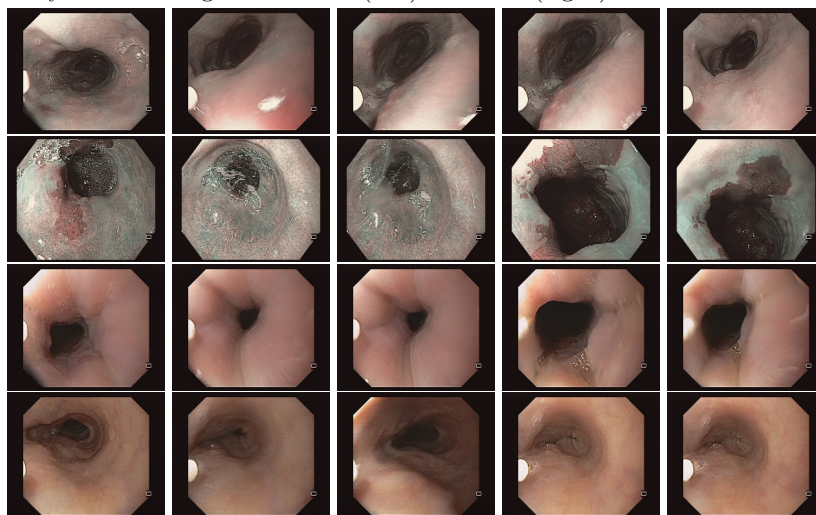

Fig. 7: Column 1 : query frames. Columns 2 & 3: the closest two EMNN matches. Columns 4 & 5: the best view-point localized images. Scores for matches for each row ordered by their columns [2,3,4,5]. Row 1: [0,1,2,1], Row 2: [0,0,2,2], Row 3: [0,0,2,2], Row 4: [0,1,2,2]

4. A. S. Vemuri, S. A Nicolau, J. Marescaux, L. Soler, and N. Ayache. Video synchronization: An approach to inter-operative biopsy site relocalization in endoluminal surgery. *IEEE Tran. Biomedical Engineering (Under Review)*.

5. Koen EA Van DS, T. Gevers, and C. GM Snoek. Evaluating color descriptors for object and scene recognition. *IEEE PAMI*, 32(9):1582–1596, 2010.

6. R. Kwitt, N. Vasconcelos, N. Rasiwasia, A. Uhl, B. Davis, M. Häfner, and F. Wrba. Endoscopic image analysis in semantic space. *Medical image analysis*, 16(7):1415–1422, 2012.

7. S. Atasoy, D. Mateus, A. Meining, G-Z Yang, and N. Navab. Endoscopic video manifolds for targeted optical biopsy. *IEEE, TMI*, 31(3):637–653, 2012.

Modality	Descriptor	color-space	Avg. Score	Std. dev.	% Zeros	% Ones	% Twos
NBI	mLBP	hsv	1.7	0.619	8.33	13.33	78.33
	mLBP	rgb	1.7	0.619	8.33	13.33	78.33
	mLBP	norm	1.667	0.655	10	13.33	76.67
	mLBP	log	1.667	0.655	10	13.33	76.67
	mLBP	OPP	1.617	0.691	11.67	15	73.33
	mLBPHOG	hsv	1.65	0.685	11.67	11.67	76.67
	mLBPHOG	norm	1.633	0.688	11.67	13.33	75
	mLBPHOG	rgb	1.633	0.688	11.67	13.33	75
	mLBPHOG	log	1.633	0.688	11.67	13.33	75
	EM-Based	n.a.	1.15	0.82	26.67	31.67	41.67
WL	LIOP	gs	1.455	0.715	12.73	29.09	58.18
	swmLBP	gs	1.4	0.784	18.18	23.64	58.18
	mLTP	gs	1.364	0.802	20	23.64	56.36
	mLBP	norm	1.364	0.825	21.82	20	58.18
	mLBP	log	1.364	0.825	21.82	20	58.18
	mLBP	gs	1.309	0.814	21.82	25.45	52.73
	mLBPHOG	gs	1.309	0.814	21.82	25.45	52.73
	mLBP	OPP	1.291	0.832	23.64	23.64	52.73
	mHOG	gs	1.2	0.803	23.64	32.73	43.64
	EM-Based	n.a.	1.182	0.796	23.64	34.55	41.82

Table 2: Avg. scores of descriptors-color-space combinations for NBI and WL (filtered of UI frames), sorted in the increasing order of column 6 (truncated to 9 best values). Columns 6,7 and 8 are the % of cases where the matched image has a score of zero, one and two respectively.

8. T. Mäenpää and M. Pietikäinen. Multi-scale binary patterns for texture analysis. In *Image Analysis*, pages 885–892. Springer, 2003.
9. A. J. Newell and L. D. G. Multiscale histogram of oriented gradient descriptors for robust character recognition. In *ICDAR*, pages 1085–1089. IEEE, 2011.
10. X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE, TMI*, 19(6):1635–1650, 2010.
11. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2, pages 2169–2178. IEEE, 2006.
12. Z. Wang, B. Fan, and F. Wu. Local intensity order pattern for feature description. In *ICCV*, pages 603–610. IEEE, 2011.
13. Md K. Bashar, T. Kitasaka, Y. Suenaga, Y. Mekada, and K. Mori. Automatic detection of informative frames from wireless capsule endoscopy images. *Medical Image Analysis*, 14(3):449–470, 2010.
14. D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1434–1453. SIAM, 2013.
15. C-C Chang and C-J Lin. LIBSVM: A library for support vector machines. *ACM Trans. on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.
16. A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.