Local Self Similar Descriptors: Comparison and Application to Gastroenterology Images

Ricardo Sousa[2], Daniel C. Moura[1], Mario Dinis-Ribeiro[3] and Miguel T. Coimbra[1]

***Abstract*— Local descriptors coupled with robust methods for learning visual dictionaries have been a pivotal tool in computer vision. Although the identification of similar patterns is commonly conducted on some stage of the bag-of-words framework, a prior assessment of spatial local similarities can be indicative of specific objects, and thus improved recognition rates. In this work we delve a function of similarity for enhancing the discriminative power of local constrained SIFT descriptors. Motivated by gastrointestinal images where diagnosis through endoscopy plays a decisive role in cancer detection and resulting prognosis, visual cues in these early stages are slim and of difficult perception. In order to capture these patterns we propose a self-similarity approach (based on a neighbourhood analysis of SIFT descriptors) to assess local variances through a weight function. Based on extensive simulations our approach achieved a performance of 88%: 3% higher than the standard SIFT, 10% higher than Haar wavelet and 13% higher than LBPs.**

I. INTRODUCTION

Over the course of the years different Computer Vision methods have been devised to improve the understanding (and recognition) of scenes and objects [1]–[3], or, for instance, tissues and biological structures in biomedical domains [4], [5]. Biologically inspired vision methods (e.g., SIFT, HMAX or HOG [1], [6]), mid-level descriptors (e.g., Fisher vectors, VLAD [7]), signal processing (e.g, wavelets, curvelets [8]) or machine learning formulations are some of the recent works that have been proposed [9]. Above all, Scale-invariant feature transform (SIFT) is one of the most used descriptors and continues to prove its robustness.

Global descriptors such as wavelets and curvelets are commonly employed in gastroenterological images [8], [10]. However, local patterns and their variations cannot be captured by these methods. Another way is to explore micro-textures with local binary patterns (LBP) and their variations [11]. Images from the gastrointestinal track have spatially constrained features that are indicative of the degree of the deformation that cannot be captured as a whole. Although local descriptors such as SIFT have been explored to capture these pattern changes [12], their straightforward application is insufficient. Considering that similar patterns are only identified in posterior phases of the learning process in this work we devise an approach to measure SIFT descriptor resemblance. Motivated by gastroenterology images where visual cues can be feeble and tissue structures of difficult perception in relation to the tissue pathology, here we explore functions sufficiently robust to capture similar patterns and sensible enough to tolerate marginal differences. We extend and generalize the work proposed by Tamaki et. al. in [4] by exploring the SIFT descriptor for the analysis of local similarities. A neighborhood of densely sampled SIFT's are pairwise compared based on different functions with the purpose to evaluate similarities. By ascertaining local and neighborhood information, dissimilar bins of the descriptors will be put far apart from similar ones. This process eases posterior tasks of the quantization and recognition process as we will see in our experimental study.

II. RELATED WORK

Global descriptors have prevailed for the analysis of gastroenterological tissues [10], [13], [14]. In [13] it is studied the capability of texture descriptors such as Gray Level Difference Matrix (GLDM) to describe the different variations in magnified endoscopy images. In Riaz [10] Gabor filter banks are coupled with autocorrelation matrices for texture description to enforce shift and rotation invariances. In [5] a semantic analysis in the feature space applied to confocal laser endomicroscopy (CLE) is proposed. Analysis of local features has been increasingly researched in recent years [1], [2]. In [4] a local similarity function was explored to assess different patterns in spatially constrained regions of SIFT descriptors. In [12] it was explored how the different multiclass strategies for SVMs affect the recognition of cancer recognition in gastroenterological images.

The presence of given patterns in an image aids the recognition task of pathologies as it has been consequently proven in several of different studies in the gastroenterology medical domain [13]. However, much of the existing works either concentrate their efforts in extracting visual relevant information that can aid in describing the information under analysis, or in the development of learning algorithms that can mimic physicians reasoning by discriminating the different pathologies that there may be present. In fact, local image processing techniques have not been comprehensively explored in the context of gastrointestinal images. In the following Sections we will pave the way for a successful usage of these techniques.

*We acknowledge PEST-OE/EEI/LA0008/2013 (GEMINI), I-City for Future Health: NORTE-07-0124-FEDER-000068 and Future Cities FP7 Project (grant agreement 316296) for providing funding for this work.

[1]Instituto de Telecomunicações da Universidade do Porto `mcoimbra at dcc.fc.up.pt`

[2]Instituto de Engenharia Biomédica, Universidade do Porto `ricardo.sousa at ieee.org`

[3]CINTESIS/Faculdade de Medicina da Universidade do Porto and the Instituto Portugues de Oncologia Porto, Portugal `mario at med.up`

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

III. Local Image Descriptor Similarity for Cancer Recognition

It is known that SIFT [1] continues to prove its robustness in the detection and description in very different set of problems. Conventionally, to obtain the SIFT descriptor we start by detecting the interest points determined by an invariant feature detector (LoG or DoG) [1]. Then, following a BoW approach, the quantization of the descriptor (provided by an unsupervised method such as K-Means) is conducted. In the end, we obtain a dictionary (or texton) that will be (generally) representative of the dataset in order to build the final representation that will serve as input for a learning algorithm, usually SVM [12]. In the following sections we will motivate the reasoning of the similarity analysis prior to the vocabulary construction and generalize our approach.

A. *Capturing Local Patterns with SIFT*

SIFT although being very popular is often criticized and its usage is usually opted out by simpler approaches such as LBP [11]. LBP is a simple method which attains very impressive results. We start by comparing the standard SIFT with the LBP. In a nutshell, LBP compares, in a neighborhood with a predefined shape and size (e.g., square, circle or other defined by the user), the gray values of the neighbor pixels with the value of the pixel centered in that region. The comparison results in a binary code of the neighborhood [11]. Triggs in [3] presented a generalization for LBP, but many other variations exist [11], [15]. For simplicity of this study, we will keep to the standard implementation [11]. Regarding SIFT, ideally it suffices to obtain samples over different interest points (e.g., salient points). However, due to high variability on the mucosa and optical deformations, this will not accurately represent the existing pattern. As it was shown in [16] in more generic databases (revisited afterwards by other authors in gastroenterology related works [4], [12], [17]) a dense sampling provides more robust results since it can represent more comprehensively the data. A direct application is however limitative due to the low visual cues that these images have. One recent tactic either uses dense SIFT features over the whole image or analyzes adjacent regions [4], [17]. In [4], a low-level local analysis was conducted by assessing the differences of two adjacent SIFT descriptors (*d*SIFT). Indeed, a similar idea was explored in a more generic way by LeCun [9] in what they have called as "SIFT macrofeatures". The rationale is to capture low-level descriptors and encode them jointly. Whereas *d*SIFT may be sensible to high variances thus generating very different descriptors for similar patterns, the second one may prove too computational expensive or render descriptors highly dependent of their spatial coordinates. Albeit, the major drawbacks are that differences are linearly assessed and image resemblances are delegated to posterior (learning) stages. Postponing to a BoW approach and classification methodologies can be ineffective due to 1) the variability of patterns within the same image; 2) valuable information which can be lost during the quantization process [18]–[20].

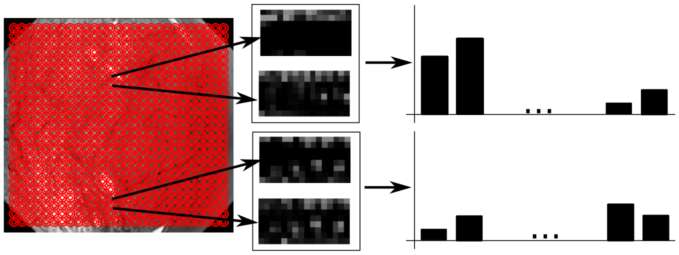

Fig. 1: Resulting histograms when calculating the differences between two adjacent SIFT descriptors.

In order to remove such ambiguities we will extend the difference similarity by a more general function. We will basis our descriptor analysis by using an exponential function where descriptor bins similarity will decay according an exponential function. Here we can benefit of the local variances to assess similarities of the SIFT descriptors. Intuitively, repetitive patterns along these regions would produce similar SIFT descriptors being analogous for irregular regions. Our strategy encompasses measuring histogram magnitude through a bin-by-bin analysis around a neighborhood with predefined size. This should maximize the discriminative power of the descriptor by separating points with higher differences even farther in the feature space. The weight function thus result in:

$$g(x) = \alpha \exp(\beta(255 - |s_i - s_j|)) + \gamma \qquad (1)$$

s corresponds a bin of a SIFT descriptor. Parameter values of Equation (1) were analytically determined by setting the maximum possible bin difference (e.g., for the standard normalized SIFT this value will be 255) to be equal to 0 and equal bins values to 0. Based on straightforward algebraic operations, we thus achieved the following values for $\alpha = 1$, $\beta = 0.0217$ and $\gamma = -1$. We obtain the new descriptor as follows:

```
Input: $\boldsymbol{p}$ as list of keypoints and $\boldsymbol{s}$ as the SIFT descriptors corresponding to the keypoints $\boldsymbol{p}$, respectively ($|\boldsymbol{p}| = |\boldsymbol{s}|$).
$w$SIFT ← {}
for each $p_i \in \boldsymbol{p}$ do
    $T \leftarrow \{p_j | d(p_i, p_j) < threshold, \forall p_j \in \boldsymbol{p}, j \neq i\}$
    for each $p_j \in T$ do
        $w' \leftarrow$ Apply Equation (1) for $s_i$ and $s_j$
        $w$SIFT ← $w$SIFT ∪ L2-norm($w'$)
    end for
end for
Return $w$SIFT
```

We have defined the threshold as 3/2 of the step size of the grid sampling. A simple experiment with larger thresholds did not report improved performances and the additional computational cost is negligible.

Normalization: Descriptor normalization is one issue (among many others—see [18] for more details) that can provide feeble recognition performance ratios if improperly conducted. Some works report better performances when features are normalized with a L2-norm [18]. However, other normalization schemes exist such as the sign squared root (SSR). To obtain a descriptor x SSR normalized one needs to transform it to $sign(x)\sqrt{|x|}$, performing after a L2-normalization. The squared root has the advantage to discount bins with high energy [7].

IV. Experimental Study

Dataset: Our dataset consists of 176 images manually annotated by a clinical expert with clinical relevant regions identified. This dataset is encompassed by 56, 96 and 24 cases for normal, metaplasia and dysplasia cases, respectively (see [21] for more information).

Results and Discussion: In order to assess the advantages on using our descriptor, in our experimental study we compared it against LBP, Wavelet with a Haar filter[1].We followed a rigorous simulation by splitting our data randomly in two sub-sets: 20 instances per class for training set and the remainder was used to generate the testing set. A 3-fold cross-validation was performed over the training set in order to find the best parameterization of our models. Performance was assessed on the testing data using the mean average performance (mAP). Simulations were repeated 10 times to obtain more stable results by averaging the mAP.

Image preprocessing and Feature Acquisition: Each image was rescaled to 259×240 pixels to smooth interlace and other image artifacts. Afterwards, a non-linear median filter of size 3×3 was also applied for improved noise removal and to avoid blurred regions. Local descriptors were densely sampled on a grid spaced by 10×10 pixels [16] acquired all over the annotated regions of our images.

Dictionary Construction: To build our dictionary, we used an unsupervised K-Means with K centroids.[2] The visual vocabulary (centroids of the K-Means) was trained on the training data and the dictionaries were obtained through average pooling [9]. The tradeoff between small and large dictionaries renders more generic or more specific representations of our datasets, respectively[3] [12].

Learning the model: For the LBP and SIFT based methods, SVM was modeled according an intersection kernel [12]. Regarding the wavelet descriptor we set a polynomial with degree 3 as kernel. The C parameter was comprised between $C = 10^{-3}$ and $C = 10^{2}$ with a step size of 10 (C is a penalty factor for each misclassified point) [23] and γ ranging the same interval for the polynomial kernel.

Based on the aforementioned configuration for our experimental study, we centred the analysis of these descriptors on the performance over the binary problem: recognizing pre-neoplastic and neoplastic tissues.

Our first experiment consisted on assessing the impact on using different scales for SIFT. Table I shows the expected increase of the performance with the usage of multiple scales. Another analysis is that using the standard SIFT with a single scale outperforms the standard LBP (a multi-scale LBP expressed feeble differences). This small test elucidates us for the fact that SIFT can capture the relevant information for recognizing pathologies in more than 3 out of 4 patients. We can also depict that LBP, although being designed to capture micro-textures, it performance was 4% inferior to the baseline SIFT with a single scale. In the following experiments we will use SIFT with three scales due to the low variance illustrated in these preliminary results.

	mAP		
LBP_8	75.5 ± 5.4		
	multiscale		
	8x8	8x8, 4x4	8x8, 4x4, 2x2
SIFT	79.0 ± 1.9	81.8 ± 3.6	81.6 ± 2.4

TABLE I: Preliminary results comparing LBP with SIFT. We can see that using multi-scales improves the results as expected and that SIFT with single scale provides better results than the LBP.

Our next experiment consists on validating the benefit on using the *w*SIFT. To assess the quality of our results we will refer as "significantly different" based on a paired two sided t-test if the different between two results being compared are statistically significant with 5% of statistical confidence [24]. We have marked our results with a ▼ symbol to represent that a method shows a statistically degradation than the result of the SIFT; and with a • to mean that a method performed statistically better than the result of the SIFT.

Method	mAP	
SIFT	81.7 ± 2.4	-
*d*SIFT (difference)	80.2 ± 2.4	▼
*d*SIFT (difference, SSR)	83.5 ± 3.3	▼
*w*SIFT (weigth function, SSR)	84.9 ± 2.6	•

TABLE II: Results for different similarity functions with vocabulary size $K = 100$.

By analyzing Table II we can see that using the difference as a function of similarity gave statistically worse results than SIFT. At the same time, our function performed statistically better than the baseline. The SSR normalized weight function also statistically outperformed the difference approach. These results render the following knowledge: the normalization process of the new feature descriptors is crucial; and the function to pairwise compare the SIFT descriptors provide very distinct results. Based on these results, we will follow our experimental study with our weight function with SSR.

Our last experiment was with different dictionary sizes. Table III summarizes this analysis. Results presented in Table III show very distinct performances improvements by using SIFT with *w*SIFT outperforming all other methods.

	mAP		
Wavelet (haar)	78.4 ± 4.3		
	Dictionary Size		
	$K = 100$	$K = 400$	
SIFT (baseline)	81.7 ± 2.4	85.4 ± 3.0	
*w*SIFT	84.9 ± 2.6	88.7 ± 3.3	•

TABLE III: shows the Mean Average Performance for SIFT and *w*SIFT in comparison to LBP and wavelets.

Confusion Matrix (CM): To conclude our analysis remains to see how these results reflect in terms of CM. These CMs represented in terms of Sensitivity and Specificity give 82%

[1] We have used as features the entropy and energy to represent the global statistics of images. In future works we will study other derivations.

[2] VLFeat toolbox was used to assist our study [22].

[3] We have experimented larger dictionaries with feeble differences.

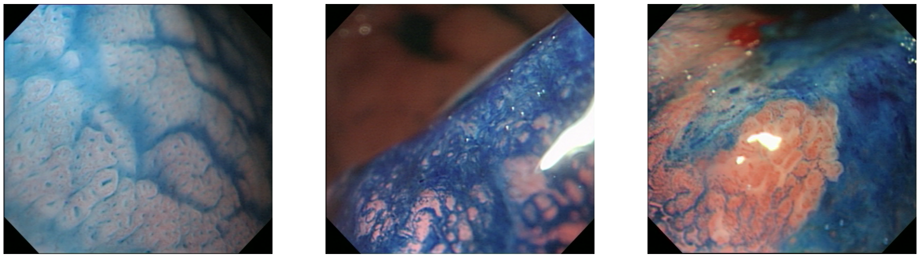

Fig. 2: Images describing different types of patterns from the gastrointestinal tract and that are illustrative of our dataset: (left) normal, (center) dysplasia and (right) metaplasia.

$$CM(\text{SIFT}) = \begin{bmatrix} 0.89 & 0.11 \\ 0.18 & 0.82 \end{bmatrix} CM(w\text{SIFT}) = \begin{bmatrix} 0.87 & 0.13 \\ 0.10 & 0.90 \end{bmatrix}$$

Fig. 3: Confusion Matrix for SIFT and *w*SIFT.

and 89% respectively for SIFT, and 90% and 87% for *w*SIFT. As expected, *w*SIFT improves the Sensitivity although it affects to some extent the Specificity. In a computer aided diagnosis system these results should be read as follows: a given patient suffering cancer will be positively detected more likely by *w*SIFT than SIFT descriptors; whereas, a patient with no health condition will be erroneous identified as a pathological patient more likely by *w*SIFT.

V. Conclusion

Here we have explored the usage of similarity functions to generate low-level descriptors capable for improving discriminability between pre-neoplastic and neoplastic tissues. With this work it was not our objective to show which computer vision methodologies are preferable to analyze pathological tissues in gastroenterological images. Indeed, we have shown that despite the advancements made, there is much space for improvements. In future works it would be very important to assess how clinical experts perform for the same set of images. This would clearly elucidate the boundaries of automatic computer vision methods.

References

[1] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2, 1999, pp. 1150 –1157 vol.2.

[2] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Foundations and Trends® in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2008.

[3] S. ul Hussain and B. Triggs, "Visual recognition using local quantized patterns," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 716–729.

[4] T. Tamaki, J. Yoshimuta, M. Kawakami, B. Raytchev, K. Kaneda, S. Yoshida, Y. Takemura, K. Onji, R. Miyaki, and S. Tanaka, "Computer-aided colorectal tumor classification in nbi endoscopy using local features," *Medical Image Analysis*, vol. 17, no. 1, pp. 78 – 100, 2013.

[5] R. Kwitt, N. Vasconcelos, N. Rasiwasia, A. Uhl, B. Davis, M. Häfner, and F. Wrba, "Endoscopic image analysis in semantic space," *Medical Image Analysis*, vol. 16, no. 7, pp. 1415 – 1422, 2012.

[6] C. Theriault, N. Thome, and M. Cord, "Extended coding and pooling in the hmax model," *Image Processing, IEEE Transactions on*, vol. 22, no. 2, pp. 764–777, 2013.

[7] R. Arandjelovic and A. Zisserman, "All about vlad," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 1578–1585.

[8] D. Barbosa, J. Ramos, and C. S. Lima, "A multi-scale comparison of texture descriptors extracted from the wavelet and curvelet domains for small bowel tumor detection in capsule endoscopy exams," in *World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany*. Springer, 2010, pp. 1546–1549.

[9] Y.-L. Boureau, N. Le Roux, F. Bach, J. Ponce, and Y. LeCun, "Ask the locals: multi-way local pooling for image recognition," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2651–2658.

[10] F. Riaz, F. Silva, M. Ribeiro, and M. Coimbra, "Invariant gabor texture descriptors for classification of gastro enterology images," *Biomedical Engineering, IEEE Transactions on*, vol. PP, no. 99, p. 1, 2012.

[11] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.

[12] R. Sousa, M. T. Coimbra, M.-D. Ribeiro, and P. Pimentel-Nunes, "Impact of SVM Multiclass Decomposition Rules for Recognition of Cancer in Gastroenterology Images," in *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*, June 2013, pp. 405–408.

[13] K. Onji, S. Yoshida, S. Tanaka, R. Kawase, Y. Takemura, S. Oka, T. Tamaki, B. Raytchev, K. Kaneda, M. Yoshihara, and K. Chayama, "Quantitative analysis of colorectal lesions observed on magnified endoscopy images," *Journal of Gastroenterology*, vol. 46, pp. 1382–1390, 2011, 10.1007/s00535-011-0459-x.

[14] M. Liedlgruber and A. Uhl, "Computer-aided decision support systems for endoscopy in the gastrointestinal tract: A review," *Biomedical Engineering, IEEE Reviews in*, vol. 4, pp. 73 –88, 2011.

[15] Z. Guo and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *Image Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1657–1663, 2010.

[16] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 524–531.

[17] Y. Takemura, S. Yoshida, S. Tanaka, R. Kawase, K. Onji, S. Oka, T. Tamaki, B. Raytchev, K. Kaneda, M. Yoshihara, and K. Chayama, "Computer-aided system for predicting the histology of colorectal tumors by using narrow-band imaging magnifying colonoscopy (with video)," *Gastrointestinal Endoscopy*, vol. 75, no. 1, pp. 179 – 185, 2012.

[18] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2011.

[19] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Descriptor learning for efficient retrieval," in *Proceedings of the 11th European conference on computer vision conference on Computer vision: Part III*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 677–691.

[20] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[21] M. Dinis-Ribeiro, A. da Costa-Pereira, C. Lopes, L. Lara-Santos, M. Guilherme, L. Moreira-Dias, H. Lomba-Viana, A. Ribeiro, C. Santos, J. Soares *et al.*, "Magnification chromoendoscopy for the diagnosis of gastric intestinal metaplasia and dysplasia," *Gastrointestinal endoscopy*, vol. 57, no. 4, pp. 498–504, 2003.

[22] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[23] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[24] T. M. Mitchell, *Machine learning. 1997*, 1997, vol. 45.