Texture classification of anatomical structures using a context–free machine learning approach

Oscar Alfonso Jiménez del Toro[ab], Antonio Foncubierta-Rodríguez[c], Adrien Depeursinge[ab] and Henning Müller[ab]

[a]University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland;
[b]University Hospitals and University of Geneva, Geneva, Switzerland;
[c]Swiss Federal Institute of Technology (ETH) Zurich, Zurich, Switzerland;

ABSTRACT

Medical images contain a large amount of visual information about structures and anomalies in the human body. To make sense of this information, human interpretation is often essential. On the other hand, computer–based approaches can exploit information contained in the images by numerically measuring and quantifying specific visual features. Annotation of organs and other anatomical regions is an important step before computing numerical features on medical images. In this paper, a texture–based organ classification algorithm is presented, which can be used to reduce the time required for annotating medical images. The texture of organs is analyzed using a combination of state–of–the–art techniques: the Riesz transform and a bag of meaningful visual words. The effect of a meaningfulness transformation in the visual word space yields two important advantages that can be seen in the results. The number of descriptors is enormously reduced down to 10% of the original size, whereas classification accuracy is improved by up to 25% with respect to the baseline approach.

Keywords: Organ texture, Riesz wavelets, computer–aided diagnosis.

1. INTRODUCTION

Analysis of visual information contained in medical images can benefit from computer–based approaches, especially when dealing with 3D and higher dimensional visual information, which is difficult to visualize and to interpret.[1] In Foncubierta–Rodríguez et al.[2] computer–based approaches have shown to leverage the information contained in high dimensional visual data.

In order to numerically analyze the information contained in medical images, annotation of regions of interest, organs and other anatomical structures is a prerequisite. Many segmentation and computer–aided diagnosis algorithms rely on a defined context to yield the desired results. For example, segmentation algorithms designed for the brain will probably fail when used to segment knee structures. DICOM headers often help to understand the context, since the metadata contai information from the patient record, imaging protocol and part of the body. However, this is not always the case and up to 15% of the DICOM headers contain inaccurate or wrong data[3] particularly in specific fields such as anatomic region.

Automatic detection of anatomical structures is a challenging problem due to the rich variability and often ambiguity of their visual appearance.[4] Computer–based analysis and detection of organs often rely on shape models and texture pattern analysis and in general focus on single organs. In,[5] shape–based discrimination[6,7] is discussed not to be appropriate and texture features based on Gray–Level Co–occurrence Matrices (GLCM) are proposed to discriminate organs and tissues. GLCM is one of the most frequently used techniques for texture analysis and although there are several implementations in 3D, the state–of–the–art for three dimensional *solid* texture are multi–resolution and multi–scale approaches[8,9] that often deliver better results.

Multi–scale features such as wavelets, curvelets, etc. can describe the information contained in medical images at least from a texture point of view. However, identifying organs requires a certain degree of artificial intelligence

Further author information: (Send correspondence to Oscar Alfonso Jiménez del Toro)
Oscar Alfonso Jiménez del Toro: E–mail: oscar.jimenez@hevs.ch, Telephone: +41 (0)27 606–9010

in order to make a decision based on the image description. In this paper we propose the use of a meaningfulness–transformed bag of visual words that learns which descriptors are essential in describing the visual appearance of the images. It weights and reduces the feature space to obtain a more compact representation of the organs, improving the results of a simple nearest–neighbor classifier by up to 25%.

The rest of the paper is organized as follows: Section 2 explains the methods used, specifically addressing the dataset (Section 2.1), the state of the art texture features used (Section 2.2), and the unsupervised machine learning methods that yield a meaningful bag of visual words for the classification task (Section 2.3). Section 3 explains the results of the experimental evaluation of the methods compared to the baseline. Results are discussed in Section 3 and conclusions and future work are outlined in Section 4.

2. METHODS

2.1. Dataset

A dataset obtained from the Visual Concept Extraction Challenge in Radiology (VISCERAL*,[10]) was used to perform the organ classification. VISCERAL organizes benchmarks on the processing of large–scale 3D radiology images.[11] Specifically it organizes two benchmarks on segmentation and identification of anatomical structures.

In this work, the 14 computed tomography volumes contained in the VISCERAL training dataset for the first benchmark were used. Seven of them are Computer Tomography (CT) volumes without contrast and 7 are contrast–enhanced CT images. The dataset is challenging because it consists of inter and intra modality volumes taken with a varying resolution (from $0.604 \times 0.604 \times 3$ mm to $0.793 \times 0.793 \times 3$ mm) classifying different anatomical structures. The VISCERAL test set was not publicly available, therefore a cross–validation scheme was used during the evaluation phase (see Section 2.4).

Five organs per volume, manually annotated by radiologists, were used as the ground truth for classification. The organs used in the paper are: liver, spleen, urinary bladder and both lungs. All participants in the VISCERAL benchmark submitted segmentations for these organs. The goal of using several organs is also to show that approaches that do not use any shape prior can work well based solely on texture informations. Optimized solutions for a single organ could obtain better results but this approach can in principle be used for any organ, once the system is properly trained. The same applies for other imaging modalities, as long as local texture plays an important role in organ characterization. This means that it is a data–driven approach instead of a model–driven approach. A change in protocol or image modality would mean a short retraining phases instead of a manual optimization as it is currently the case for most optimized algorithms.

2.2. 3D texture analysis based on directional Riesz wavelets

3D Riesz filterbanks are used to characterize the texture properties of the organs in the CT images. 3D Riesz wavelets are steerable and multiscale and yield overcomplete characterization of local scales and orientation properties.[12,13] They are therefore able to model subtle local 3D texture properties with high reproducibility compared to other methods. The latter require arbitrary selection of scales and directions. The N–th order Riesz transform $\boldsymbol{\mathcal{R}}^{(N)}$ of a three–dimensional signal $f(\boldsymbol{x})$ is defined in the Fourier domain as:

$$\widehat{\mathcal{R}^{(n_1,n_2,n_3)}} f(\boldsymbol{\omega}) = \sqrt{\frac{n_1+n_2+n_3}{n_1!n_2!n_3!}} \frac{(-j\omega_1)^{n_1}(-j\omega_2)^{n_2}(-j\omega_3)^{n_3}}{||\boldsymbol{\omega}||^{n_1+n_2+n_3}} \hat{f}(\boldsymbol{\omega}), \tag{1}$$

for all combinations of (n_1, n_2, n_3) with $n_1 + n_2 + n_3 = N$ and $n_{1,2,3} \in \mathbb{N}$. Eq. (1) yields $\binom{N+2}{2}$ templates $\mathcal{R}^{(n_1,n_2,n_3)}$ and forms multiscale filterbanks when coupled with a multi–resolution framework based on isotropic band–limited wavelets (e.g., Simoncelli).[12] It therefore allows continuous descriptions of three–dimensional scales and directions.

The images were resampled to have an isotropic voxel resolution using nearest–neighbor interpolation, each of the annotated ROIs was divided into $16 \times 16 \times 16$ overlapping blocks. Finally, 1000 of these blocks were randomly chosen as samples of each anatomical structure.

*`http://www.visceral.eu`, as of 10 December 2014.

Based on previous work,[13,14] we found that $N = 2$ and 4 scales (i.e., 24 subbands) provided a good trade–off between the dimensionality of the feature space and the wealth of the filterbanks. The energy of the Riesz coefficients of the 24 subbands were used as a feature vector for each block.

2.3. Feature modelling and learning with meaningful bags of visual words

The question addressed in this paper is how organ classification can be improved when using state of the art texture features and a bag of visual words. In this section we briefly describe the concept of visual words, and then we explain a feature modelling technique that learns the most meaningful visual descriptors within the visual words, producing a transformation of the visual word space.

2.3.1. The bag of visual words

The *bag of words* (BOW) approach consists of describing a document based on the words that it contains and the number of times they occur in it without taking into account the word order. By establishing a common *vocabulary* for all documents, the documents can be described using a histogram of words: a vector where the i–th component corresponds to the multiplicity of the i–th word from the vocabulary in the document.

This technique was extended to visual documents such as images and videos, and the term Bag of Visual Words (BOVW) was coined[15] based on several approaches that used text retrieval techniques for image retrieval.[16,17] It is frequently used[18] and it has become very popular in the multimedia retrieval area as it has obtained very good results in most visual retrieval and object recognition benchmarks.

In general, a visual vocabulary is obtained by clustering a continuous feature space into a fixed number of regions or *visual words*. In this case, the feature space is populated with the Riesz features of the sampled blocks from each organ and patient, and words are generated using k–means clustering for various values of k. Therefore, each anatomical structure can be described in terms of the histogram of the visual words corresponding to its 1000 blocks.

The bag of visual words approach presents the advantage of using descriptors that are derived from the content actually present in the data set, generating data–driven, potentially discriminative features. It considers not only the visual words individually but their relative presence in the data set and in single images. Figure 1 shows how this technique is used in comparison with an approach that uses the features directly and without modelling.

2.3.2. Meaningfulness of visual words

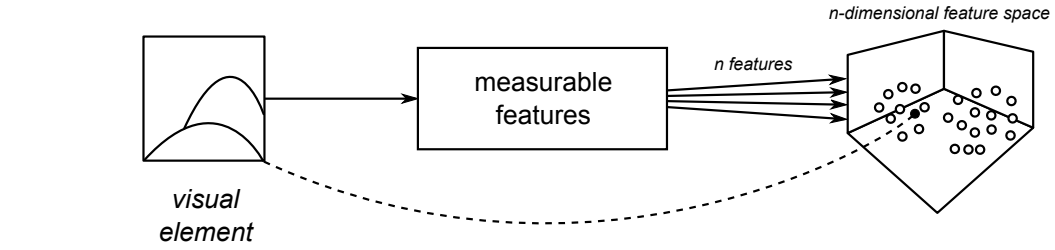
In spoken or written language, not all words contain the same amount of information. Similarly, in a vocabulary of N_W visual words generated by clustering a feature space populated with training data, not all words are useful to describe the appearance of the visual instances in the same way.

Classification or recognition of organs requires very specific words with high discriminative power. On the other hand, using very specific words alone does not always allow to establish and recognize similarities. This can be done by establishing a concept that generalizes very specific words that share similar meanings into a less specific *visual topic*. E.g. in order to recognize the similarities between the (specific) words *bird* and *fish* we need a less specific *topic* such as *animal*.

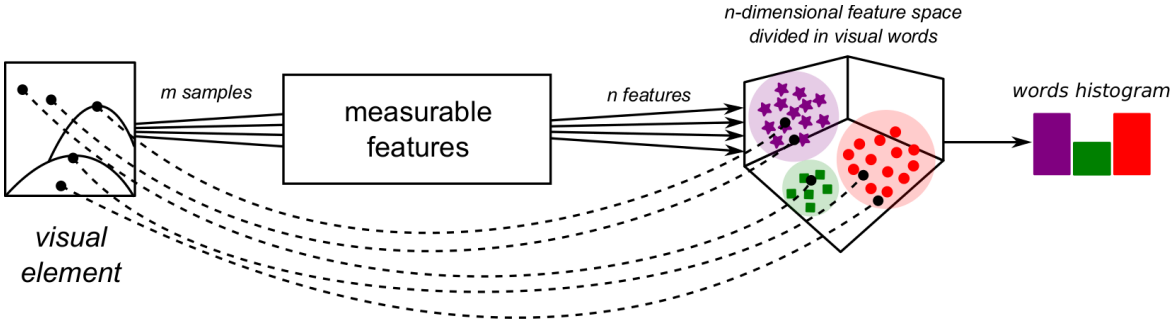
Definition 1 (Visual topic). *A visual topic z is the representation of a generalized version of the visual appearance of an anatomical structure modelled by various visual words. It corresponds to an intermediate level between visual words and the complete understanding of visual information contained in a 3D image. A set of visual topics $\mathcal{Z} = \{z_1, \ldots, z_{N_Z}\}$ can be defined in a way that every visual word can belong to none, one or several visual topics, therefore establishing and possibly quantifying the relationships among words (see Figure 2).*

In the definition of Probabilistic Latent Semantic Analysis (PLSA),[19] Hofmann defines a generative model that states that the observed probability of a word or term occurring in a given document is linked to a latent or unobserved set of topics (also called aspects) in the text. In addition, PLSA is able to identify these topics and establish their relations with each of the words.

PLSA in combination with visual words for classification and retrieval was also applied in.[20] In[21] PLSA is proposed to remove noisy visual words. This approach is further extended with the concept of *meaningfulness*



(a) Direct approach without modelling.



(b) Visual words approach.

Figure 1: Use of the bag of visual words in visual data description. In (a) an image corresponds to a single point in the feature space. In (b) an image is described by the histogram of visual words which contains information about the relations of various sampled points in the feature space. Words are learnt from the distribution of training data in the feature space.

in,[22] obtaining reductions of up to 92% of the vocabulary size without significant effect on image classification accuracy.

Definition 2 (PLSA–based visual topic). *Let I_i be the bag of words representation of an organ or anatomical structure of a given patient. A visual topic is an unobserved or latent variable $z \in \mathcal{Z} = \{z_1, \ldots, z_{N_Z}\}$ so that the probability of observing the word w_n in the anatomical structure I_i:*

$$P(w_n, I_i) = \sum_{j=1}^{N_Z} P(w_n|z_j)P(z_j|I_i).$$

This model assumes that organs and anatomical structures present a set of visual aspects or visual topics, that then generate a of bag of words. The model is fit via the EM (Expectation–Maximization) algorithm (see Figure 3). For the expectation step:

$$P(z_j|I_i, w_n) = \frac{P(w_n|z_j)P(z_j|I_i)}{\sum_{j=1}^{N_Z} P(w_n|z_j)P(z_j|I_i)}. \tag{2}$$

and for the maximization step:

$$P(w_n|z_j) = \frac{\sum_{i=1}^{N_I} n(I_i, w_n)P(z_j|I_i, w_n)}{\sum_{m=1}^{N_W}\sum_{i=1}^{N_I} n(I_i, w_m)P(z_j|I_i, w_m)}, \tag{3}$$

$$P(z_j, I_i) = \frac{\sum_{m=1}^{M_W} n(I_i, w_m)P(z_j|I_i, w_m)}{n(I_i)}. \tag{4}$$

where $n(I_i, w_n)$ denotes the number of times the visual word w_n occurred in the anatomical structure I_i; and $n(I_i) = \sum_k (I_i, w_n)$ refers to the total number of visual words in I_i.

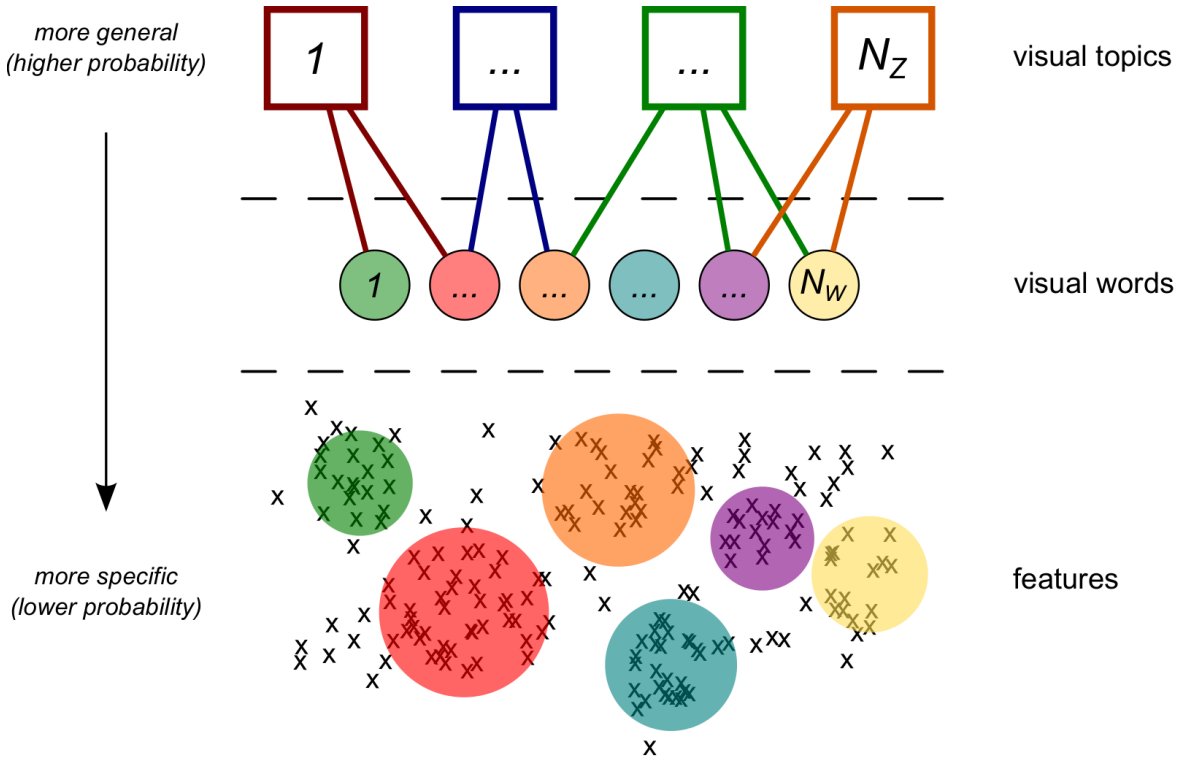


Figure 2: Conceptual model of visual topics, words and features. Whereas continuous features are the most informative descriptors from an information theoretical point of view, visual words generalize feature points that are close in the feature space. We propose visual topics as a higher generalization level, modelling partially shared meanings among words.

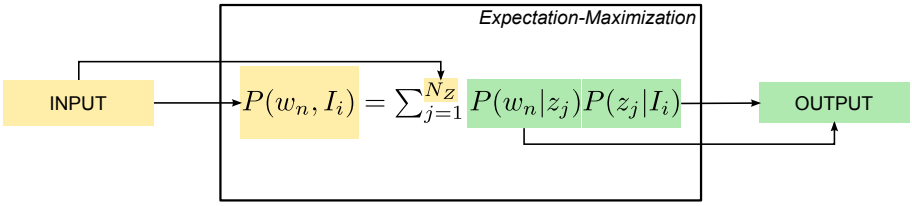


Figure 3: Input–output diagram of PLSA applied to visual words.

These steps are repeated until convergence or until a termination condition is met. As a result, two probability matrices are obtained: the word–concept probability matrix $W_{N_W \times N_Z} = (P(w_n|z_j))_{n,j}$ and the concept–organ probability matrix $D_{N_Z \times N_W} = (P(z_j|I_i))_{j,i}$.

As a first approximation to topic–based word weighting, visual significance for each visual word/topic pair can be quantified. Following the ideas from[21,22] we define the visual significance of a word for a given topic. This quantifies how much a word belongs to a given topic.

Definition 3 (Topic–based significance). *Given a visual topic $z_j \in \mathcal{Z}$ and the set of probabilities $\mathcal{P} = \{P(w_m|z_j)\}\ \forall m = 1, \ldots, N_W$, the significance of a word w_n for the visual topic z_j is defined as the ratio of elements in $\mathcal{P}$ with a lower conditional probability than $P(w_n|z_j)$:*

$$t_{n,j} = \frac{|\{p \in \mathcal{P} \mid p \leq P(w_n|z_j)\}|}{N_W}$$

Definition 4 (Visual meaningfulness). *The visual meaningfulness of a visual word w_n is its maximum topic–based significance level:*

$$m_n = \begin{cases} \max_j \{t_{n,j}\} \text{ if } \max_j \{t_{n,j}\} \geq T_{meaning} \\ 0 \text{ otherwise} \end{cases}$$

Given a meaningfulness threshold $T_{meaning}$, words that are not meaningful for any concept at this level can be removed from the visual word space, producing a *meaningfulness–truncated feature space*. This approach was tested in,[22] achieving reduction ratios of up to 92% of the feature space with a very limited cost in classification accuracy and retrieval precision.

Instead of using a hard decision based on a meaningfulness threshold, a transformation can be defined to weight visual words according to their meaningfulness.

Definition 5 (Meaningfulness–transformed visual word space). *Let* $\mathbf{h}$ *be a histogram vector where each component represents the multiplicity of a visual word, and* $\mathbf{M}$ *a meaningfulness transformation matrix:*

$$\mathbf{h} = (n(w_1), n(w_2), \ldots, n(w_{N_W}))^T \tag{5}$$

$$\mathbf{M} = \begin{pmatrix} m_1 & 0 & \cdots & 0 \\ 0 & m_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_{N_W} \end{pmatrix} \tag{6}$$

Then, the vector $\mathbf{h}^{\mathbf{M}} = (n(w_1^M), n(w_2^M), \ldots, n(w_{N_W}^M))^T$ *is the histogram vector of visual words in the meaningfulness–transformed space.*

$$\mathbf{h}^{\mathbf{M}} = \mathbf{M}\mathbf{h} \tag{7}$$

$$n(w_i^M) = m_i \cdot n(w_i) \tag{8}$$

2.4. Experimental setup

In order to evaluate the effect of the meaningfulness–transformed BOVW, an experimental evaluation was performed using the following processing pipeline (see Figure 4). Vocabularies, topics and the meaningfulness transformation were calculated based on 13 subjects. In all these steps, one test patient was left out of the model. These steps were repeated 14 times, achieving a Leave–One–Patient–Out Cross–Validation (LOPOCV):

1. For each organ, Riesz features were extracted from 1000 randomly selected blocks of $16 \times 16 \times 16$ voxels. Selection of blocks was based on their center, and therefore, some blocks would partly be outside the organ.
2. A different vocabulary was created for each of the organs, using 50, 100 and 150 visual words and a k–means clustering. Building a vocabulary per organ produced a complete codebook of 250, 500 and 750 visual words, in line with state of the art work based on visual vocabularies.
3. Each block was assigned to the closest cluster from any of the vocabularies, and a bag of visual words histogram was obtained for each of the organs of each patient.
4. For a varying number of topics from 25 to 350, and a varying meaningfulness threshold, a transformation was computed and applied to the bag of visual words histograms.

The dataset contains annotations for both the right and left lung, therefore the experiments were performed considering both instances as the same organ and also considering both instances as different organs. Despite the obvious clinical interest in distinguishing one lung from the other, it does not make sense to use only texture to perform this classification, since both lungs contain the same tissue types. Adding information on the place relative to the body could make the separation of left and right lung trivial.

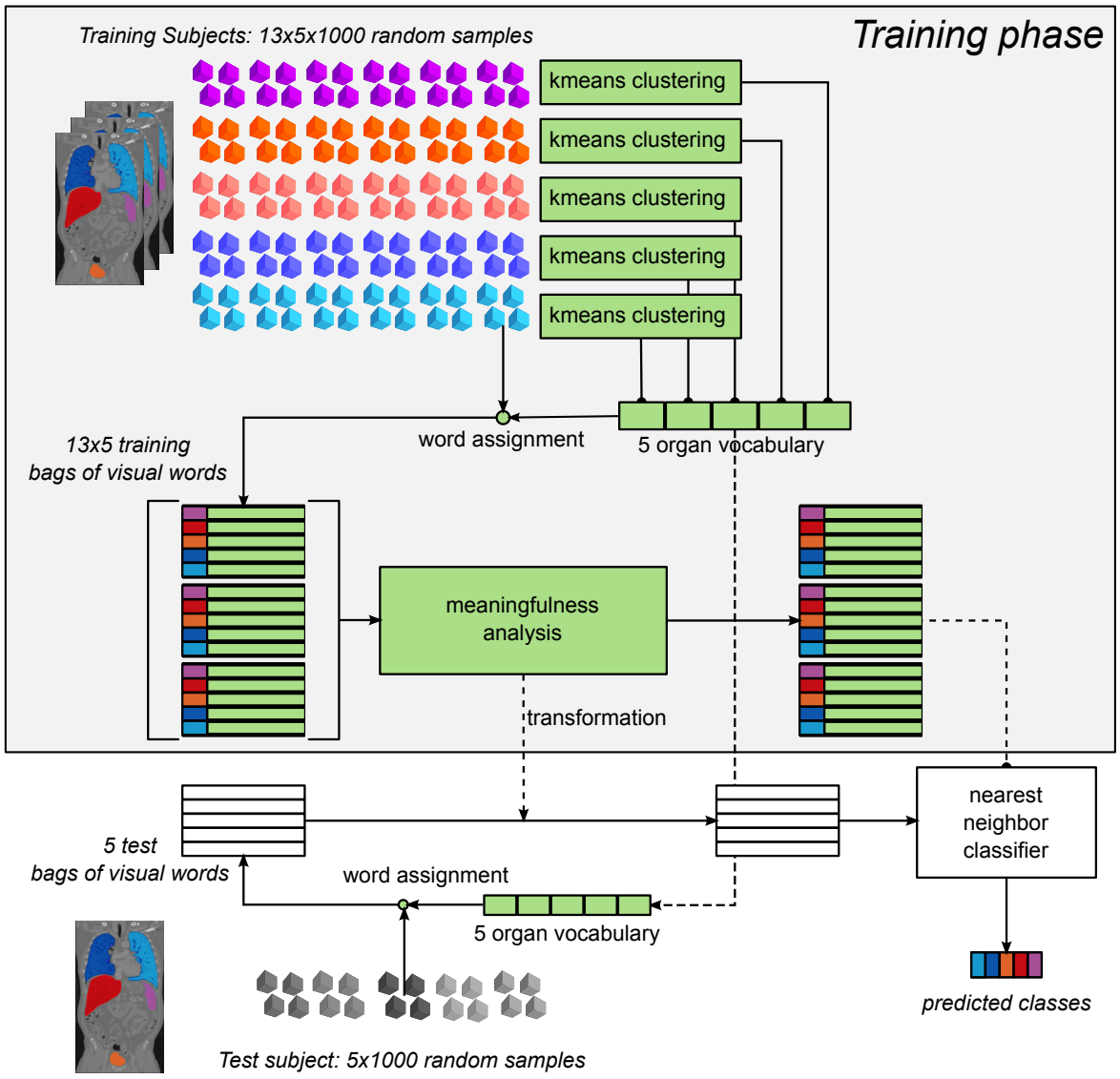


Figure 4: Experimental setup pipeline. In each LOPOCV iteration, 1000 randomly chosen samples from 5 organs in 13 subjects were used to build the model during the training phase. The 1000 samples from each organ of the remaining patient were classified using the model learnt in the training phase.

It is important to clarify that this experiment does not attempt segmentation of the organs, which would include shape priors, registration with an atlas or other information of the volume. Using a set of randomly sampled patches and using only local texture, organs can be identified. This opens a path for anatomical region identification based on poor segmentation or even bounding boxes.

3. RESULTS

In this section we summarize the results of each of the experiments performed.

3.1. Classification using five classes

Figure 5 shows the impact of the meaningfulness transformation on the accuracy and effective vocabulary size when varying the meaningfulness threshold compared to the baseline approaches for each of the three vocabulary sizes used for clustering.

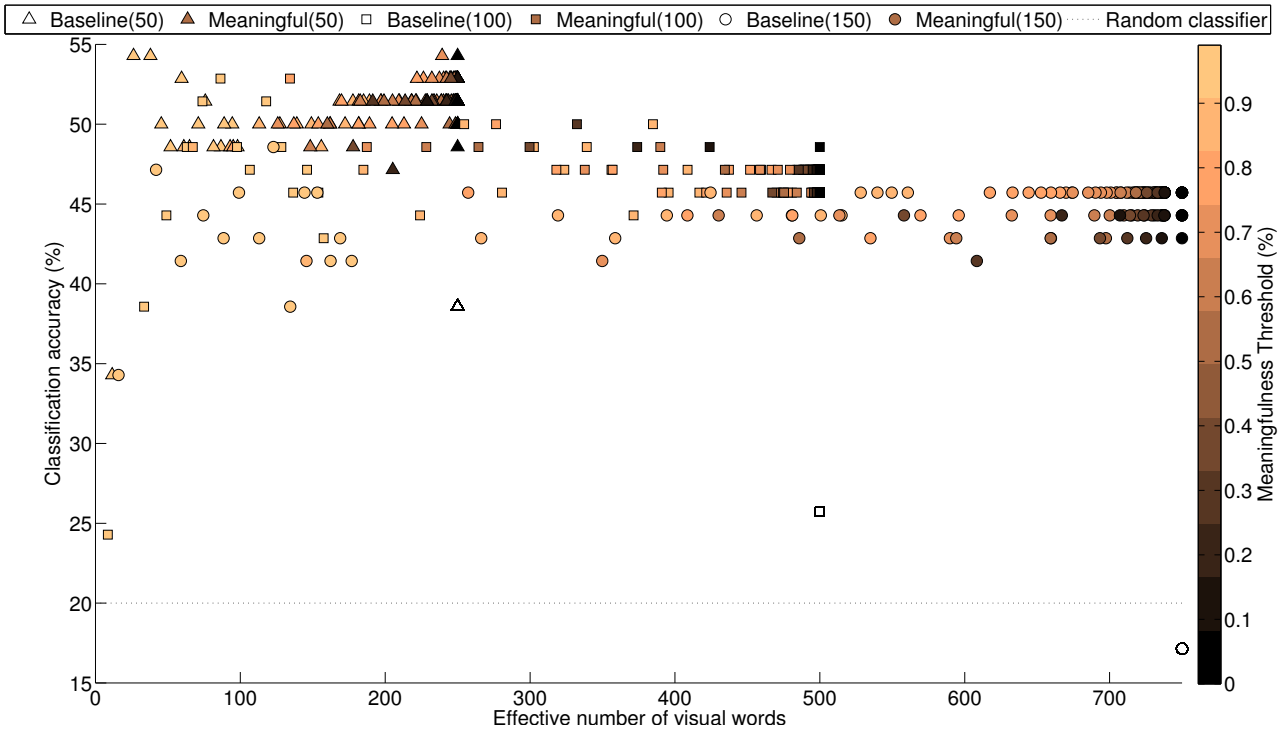


Figure 5: Classification accuracy and effective vocabulary size with varying meaningfulness threshold (color–coded). Number of visual words generated via clustering for each data series is shown in brackets.

Table 1: Optimal results for classification considering five organs. The optimal parameters are computed for the minimum vocabulary size that provides the maximum accuracy.

Clusters generated per organ	50	100	150
Full vocabulary size	250	500	750
Optimal vocabulary size	26.29	86.29	122.86
Baseline accuracy	38.57%	25.71%	17.14%
Optimal accuracy	54.29%	52.86%	48.57%
Meaningfulness threshold	99%	99%	99%
Number of topics	50	100	200

The optimal results obtained for each of the vocabulary sizes used for clustering are shown in Table 1.

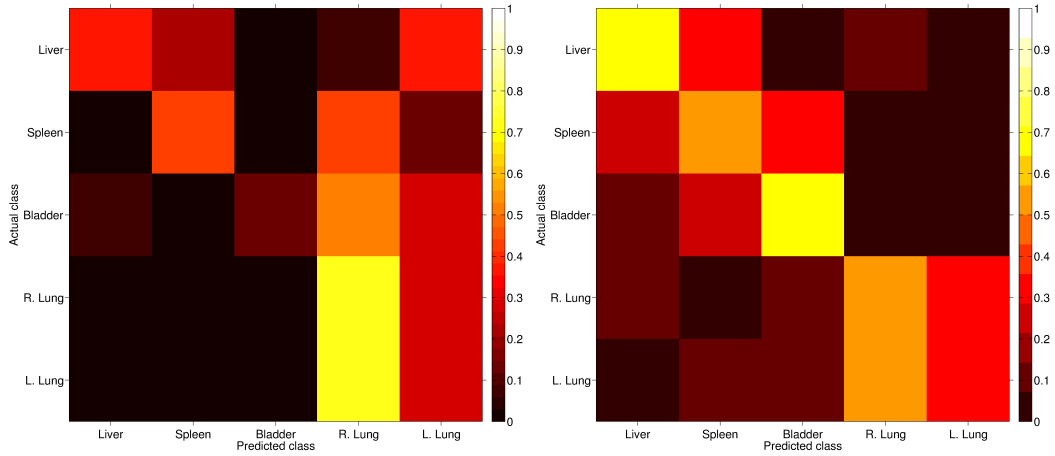
When considering both lungs as independent organs, all five classes are equally distributed, and therefore a random classifier would have an accuracy of 20%. In a multiple class situation, confusion matrices are useful to understand the cases where a non–random classification algorithm fails or succeeds. Figure 8 contains the confusion matrices for the baseline approach and the meaningfulness–transformed bag of visual words approach for each of the vocabulary sizes.

3.2. Classification using four classes

Figure 9 shows the impact of the meaningfulness transformation on the accuracy and effective vocabulary size when varying the meaningfulness threshold compared to the baseline approaches for each of the three vocabulary sizes used for clustering.

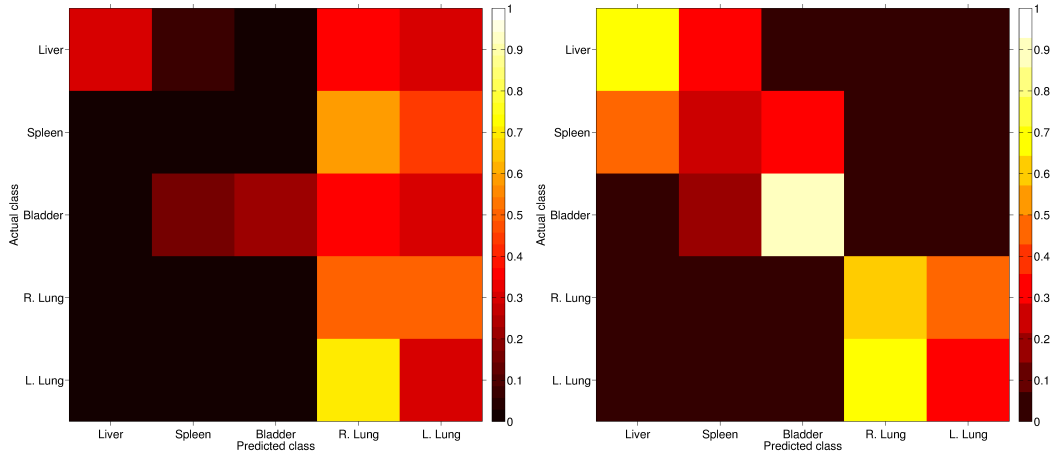
The optimal results obtained for each of the vocabulary sizes used for clustering are shown in Table 2.

If we assume that the two lungs are two items of the same class, we can perform a 4 organ classification with unbalanced classes. In this case, the simplest classifier would assign all instances to the over–represented class, with an accuracy of 40%. Figure 12 contains the confusion matrices for the baseline approach and the meaningfulness–transformed bag of visual words approach for each of the vocabulary sizes.



(a) Baseline results for vocabularies of 50 visual words per organ.

(b) Optimal results for vocabularies of 50 visual words per organ.



(a) Baseline results for vocabularies of 100 visual words per organ.

(b) Optimal results for vocabularies of 100 visual words per organ.

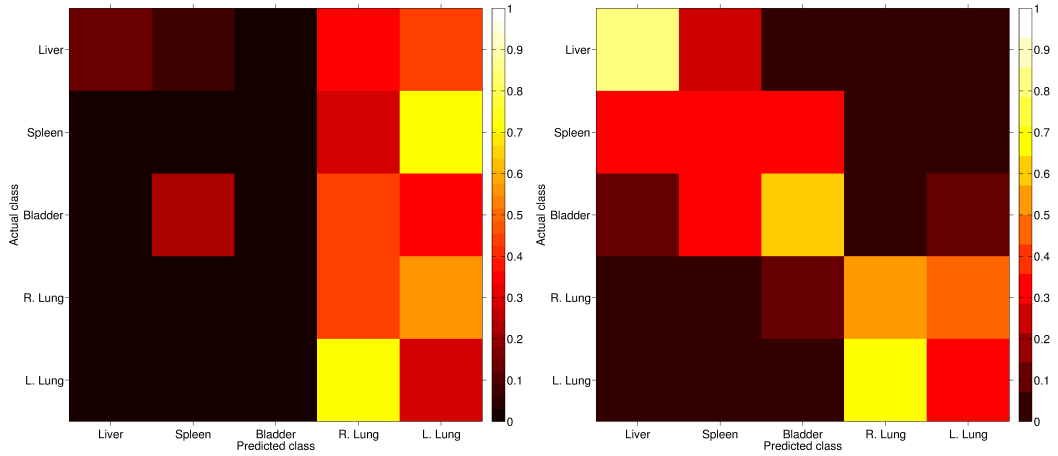
The results clearly show the impact of the meaningfulness transformation on the visual words. On the one hand, when there is no truncation of the number of visual words (meaningfulness threshold equals 0), the classification accuracy is largely improved with respect to the baseline. On the other hand, large reductions of the descriptor dimensionality can be achieved by removing meaningless words (meaningfulness threshold greater than 0).

Classification accuracy is consistently higher than the baseline for any combination of parameters (meaningfulness threshold and number of topics) and only in very extreme cases (large meaningfulness threshold combined with small number of topics) results are worse than the baseline.

Visual inspection of the confusion matrices (Figures 8 and 12) shows that accuracy improvement is not related

Table 2: Optimal results for classification considering four organs. The optimal parameters are computed for the minimum vocabulary size that provides the maximum accuracy.

Clusters generated per organ	50	100	150
Full vocabulary size	250	500	750
Optimal vocabulary size	44.57	85.07	526
Baseline accuracy	58.57%	50%	42.85%
Optimal accuracy	72.86%	75.71%	71.43%
Meaningfulness threshold	99%	99%	90%
Number of topics	100	150	275



(a) Baseline results for vocabularies of 150 visual words per organ.

(b) Optimal results for vocabularies of 150 visual words per organ.

Figure 8: Comparison of the confusion matrices for the optimal meaningfully–transformed bag of visual words and the baseline approach with five organs.

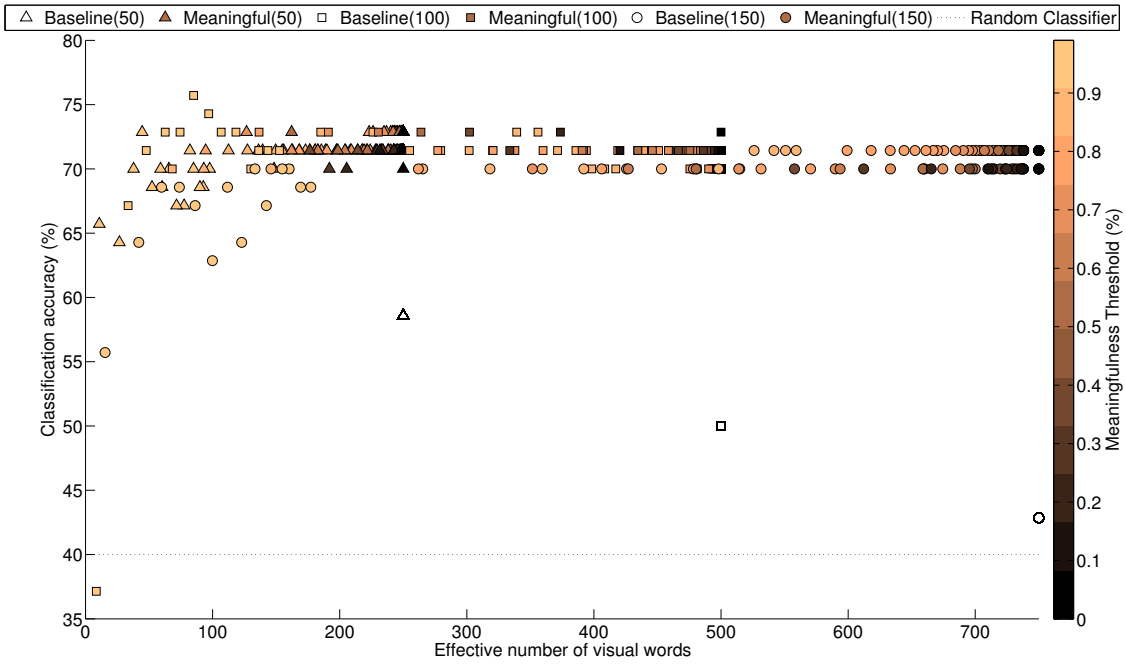
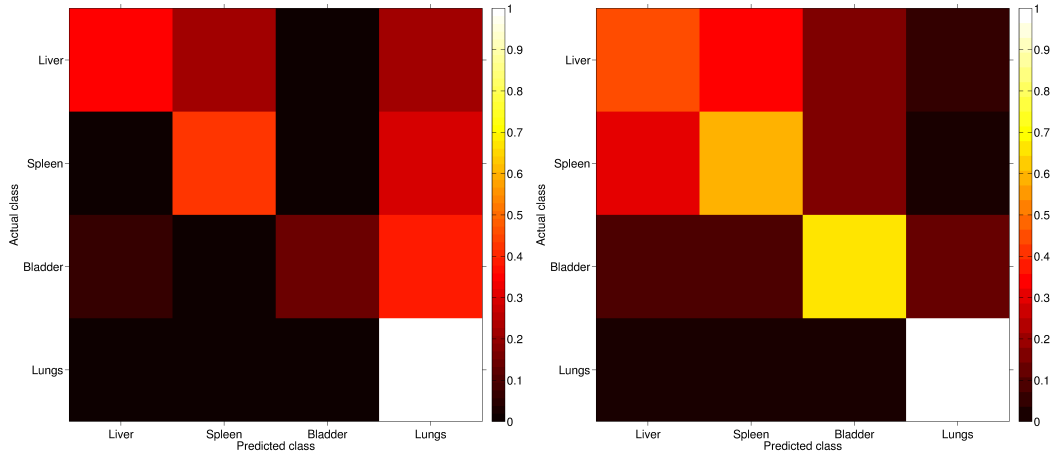


Figure 9: Classification accuracy and effective vocabulary size varying the meaningfulness threshold (color–coded). Number of visual words generated via clustering for each data series is shown in brackets.

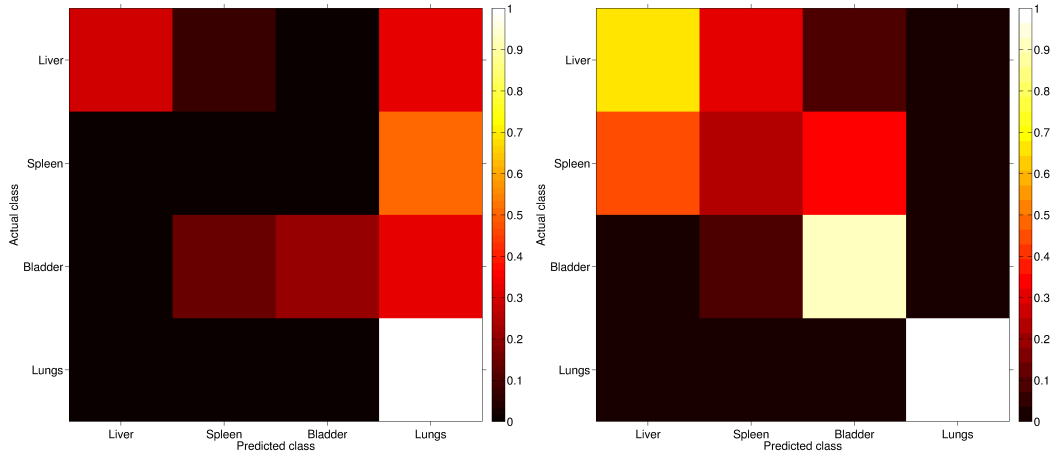
to a single class but to a consistently better classification across classes. These results can likely be applied to any other domain of image retrieval or image classification that use visual words or similar feature spaces.

A trend observed among the optimal parameters is that the number of visual words produced via clustering strongly affects the optimal number of visual topics to be discovered in the data.



(a) Baseline results for vocabularies of 50 visual words per organ.

(b) Optimal results for vocabularies of 50 visual words per organ.



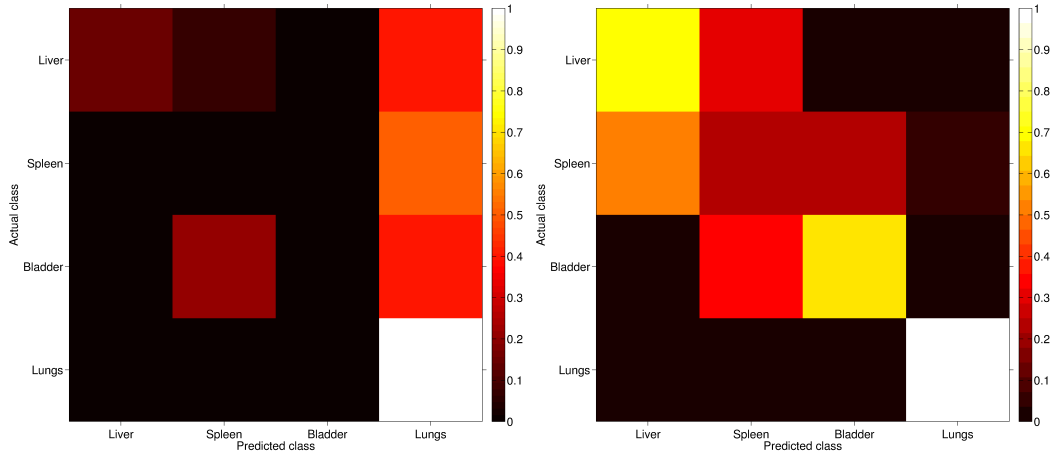
(a) Baseline results for vocabularies of 100 visual words per organ.

(b) Optimal results for vocabularies of 100 visual words per organ.

4. DISCUSSION AND CONCLUSIONS

An anatomical structure classification method is presented combining texture features obtained from the 3D Riesz wavelet energy coefficients with a machine learning approach using a bag of visual words. We evaluate the method with CT and ceCT volumes testing random 3D blocks created inside ten anatomical structures in a fully unsupervised manner. The use of randomly located blocks proves that inaccurate segmentations or dense grid sampling can still allow accurate classification of the anatomical structure using a much smaller neighbourhood without any initialization. The results support the robustness of the proposed method with classification accuracies up to 100% for the lungs and similarly high for other organs with specific textures. Some structures like the liver are harder to classify since the organ can contain different regions with varying texture that resemble in some regions texture patterns of the other structures. The classification accuracy shown for recently studied hard to segment organs like the pancreas (64% in Figure 8 vs. 64% in [23]) and urinary bladder (74% in Figure 8 vs. 0.81% in [24]) using a method which is at the same time unsupervised, context–free and multiple class, the results of the method presented in this paper are at least comparable to the state of the art organ specific segmentation methods.

The proposed approach in this paper can in principle be applied to any structure, once the system is properly trained. Although, only CT volumes were included in the experiments, the method successfully classified multiple anatomical structures in both unenhanced and contrast enhanced scans from different body regions. In addition, it managed to determine the most relevant texture signatures for each of these structures reducing considerably



(a) Baseline results for vocabularies of 150 visual words per organ.

(b) Optimal results for vocabularies of 150 visual words per organ.

Figure 12: Comparison of the confusion matrices for the optimal meaningfully–transformed bag of visual words and the baseline approach with four organs.

the vocabulary needed for the classification. The same pipeline could be applied for other imaging modalities, like magnetic resonance imaging in its various sequences.

Multi–scale texture features are modelled into visual words that are transformed based on the notions of meaningfulness and visual topics. Visual words that are meaningful are weighted higher than those that are identified as meaningless, which can be used to truncate meaningless words to produce a smaller vocabulary. Results show an improvement over the baseline approach (i.e., visual words without transformation) in terms of classification accuracy as well as descriptor size. The benefits of using a set of random samples to build a bag of words approach coupled with a meaningfulness transformation are on the one hand, that a system able to work with randomly selected samples is more robust against local changes in texture (e.g.: in case of localized disease texture). On the other hand using a completely unsupervised method for learning meaningful visual words supports the possibility of multiple visual appearances for the same organ or anatomical structure, since the class does not play any role in the learning. This can be useful in cases of diseases that change the overall appearance of organs, as long as there is enough training data to identify the similarities in visual appearance.

The method presented in this paper enables future work on identification of anatomical regions in images from the medical literature or images with inconsistent metadata and headers, as well as enabling anatomical filtering, for example of content–based retrieval results. As a data–driven approach, it has the advantages of short retraining phases instead of a manual optimization that model–based approaches would require, which is a limitation for most of the optimized algorithms currently available. The use of semantic analysis to enhance the understanding of medical images described with visual words shows also a clear path for future work, which includes the analysis of word–to–word relationships for visual words with shared meanings and exploration of the limits of vocabulary reduction when the classification tasks become more complex.

5. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 318068 VISCERAL project and from the SNF MANY 2 project: 205320_141300/1.

REFERENCES

1. A. Foncubierta-Rodríguez, H. Müller, and A. Depeursinge, "Retrieval of high–dimensional visual data: current state, trends and challenges ahead," *Multimedia Tools and Applications* , pp. 1–29, 2013.

2. A. Foncubierta-Rodríguez, P.-A. Poletti, A. Platon, A. Vargas, H. Müller, and A. Depeursinge, "Texture quantification in 4D dual energy CT for pulmonary embolism diagnosis," in *MICCAI workshop MCBR–CDS 2012*, *Springer LNCS* **7723**, pp. 45–56, 2013.
3. M. O. Güld, M. Kohnen, D. Keysers, H. Schubert, B. B. Wein, J. Bredno, and T. M. Lehmann, "Quality of DICOM header information for image categorization," in *International Symposium on Medical Imaging*, *SPIEProc* **4685**, pp. 280–287, (San Diego, CA, USA), Feb. 2002.
4. O. A. Jiménez del Toro and H. Müller, "Multi–structure atlas–based segmentation using anatomical regions of interest," in *MICCAI workshop on Medical Computer Vision*, *Lecture Notes in Computer Science*, Springer, 2013.
5. R. Susomboon, D. S. Raicu, and J. Furst, "Pixel-based texture classification of tissues in computed tomography," in *CTI Research Symposium*, 2006.
6. E. Persoon and K.-S. Fu, "Shape discrimination using fourier descriptors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (3), pp. 388–397, 1986.
7. R. N. Dave and T. Fu, "Robust shape detection using fuzzy clustering: practical applications," *Fuzzy Sets and Systems* **65**(2), pp. 161–185, 1994.
8. L. Dettori and L. Semler, "A comparison of wavelet, ridgelet, and curvelet–based texture classification algorithms in computed tomography," *Computers in Biology and Medicine* **37**, pp. 486–498, Apr. 2007. Wavelet–based Algorithms for Medical Problems.
9. A. Depeursinge, A. Foncubierta-Rodríguez, D. Van De Ville, and H. Müller, "Three–dimensional solid texture analysis and retrieval in biomedical imaging: review and opportunities," *Medical Image Analysis* **18**(1), pp. 176–196, 2014.
10. A. Hanbury, H. Müller, G. Langs, and B. H. Menze, "Cloud–based evaluation framework for big data," in *Future Internet Assembly (FIA) book 2013*, A. Galis and A. Gavras, eds., *Springer LNCS*, pp. 104–114, Springer Berlin Heidelberg, 2014.
11. O. A. Jiménez del Toro, O. Goksel, B. Menze, H. Müller, G. Langs, M.-A. Weber, I. Eggel, K. Gruenberg, M. Holzer, G. Kotsios-Kontokotsios, M. Krenn, R. Schaer, A. A. Taha, M. Winterstein, and A. Hanbury, "VISCERAL – VISual Concept Extraction challenge in RAdioLogy: ISBI 2014 challenge organization," in *Proceedings of the VISCERAL Challenge at ISBI*, O. Goksel, ed., *CEUR Workshop Proceedings*, pp. 6–15, (Beijing, China), May 2014.
12. N. Chenouard and M. Unser, "3D steerable wavelets and monogenic analysis for bioimaging," in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 2132–2135, April 2011.
13. A. Depeursinge, A. Foncubierta-Rodríguez, A. Vargas, D. Van De Ville, A. Platon, P.-A. Poletti, and H. Müller, "Rotation–covariant texture analysis of 4D dual–energy CT as an indicator of local pulmonary perfusion," in *IEEE 10th International Symposium on Biomedical Imaging*, *ISBI 2013*, pp. 149–152, IEEE, Apr. 2013.
14. O. A. Jiménez del Toro, A. Foncubierta-Rodríguez, M.-I. Vargas Gomez, H. Müller, and A. Depeursinge, "Epileptogenic lesion quantification in MRI using contralateral 3D texture comparisons," in *Medical Image Computing and Computer–Assisted Intervention —- MICCAI 2013*, *Springer Lecture Notes in Computer Science* **8150**, pp. 353–360, 2013.
15. J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, *ICCV '03*, pp. 1470–1477, IEEE Computer Society, (Washington, DC, USA), 2003.
16. D. M. Squire, W. Müller, H. Müller, and T. Pun, "Content–based query of image databases: inspirations from text retrieval," *Pattern Recognition Letters (Selected Papers from The 11th Scandinavian Conference on Image Analysis SCIA '99)* **21**(13–14), pp. 1193–1198, 2000. B.K. Ersboll, P. Johansen, Eds.
17. G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 1–22, 2004.
18. U. Avni, H. Greenspan, E. Konen, M. Sharon, and J. Goldberger, "X–ray categorization and retrieval on the organ and pathology level, using patch–based visual words.," *IEEE Transactions on Medical Imaging* **30**(3), pp. 733–746, 2011.

19. T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine learning* **42**(1-2), pp. 177–196, 2001.
20. I. El sayad, J. Martinet, T. Urruty, and C. Djeraba, "Toward a higher–level visual representation for content–based image retrieval," *Multimedia Tools and Applications* **60**(2), pp. 455–482, 2012.
21. P. Tirilly, V. Claveau, and P. Gros, "Language modeling for bag-of-visual words image categorization," in *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pp. 249–258, ACM, 2008.
22. A. Foncubierta-Rodríguez, A. García Seco de Herrera, and H. Müller, "Medical image retrieval using bag of meaningful visual words: Unsupervised visual vocabulary pruning with PLSA," in *Workshop on Multimedia Information Indexing and Retrieval for Healthcare, ACM Multimedia*, 2013.
23. M. Oda, T. Nakaoka, T. Kitasaka, K. Furukawa, K. Misawa, M. Fujiwara, and K. Mori, "Organ segmentation from 3D abdominal CT images based on atlas selection and graph cut," in *Abdominal Imaging. Computational and Clinical Applications*, H. Yoshida, G. Sakas, and M. G. Linguraru, eds., *Lecture Notes in Computer Science* **7029**, pp. 181–188, Springer Berlin Heidelberg, 2012.
24. M. J. Costa, H. Delingette, S. Novellas, and N. Ayache, "Automatic segmentation of bladder and prostate using coupled 3d deformable models," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007*, N. Ayache, S. Ourselin, and A. Maeder, eds., *Lecture Notes in Computer Science* **4791**, pp. 252–260, Springer Berlin Heidelberg, 2007.