Automated histological quantification of trabecular bone tissue in critical illness

Thomas Janssens[1]
thomas.janssens2@med.kuleuven.be
Ine Vanhees[1]
ineke.vanhees@med.kuleuven.be
Jan Gunst[1]
jan.gunst@med.kuleuven.be
Helen Owen[2]
h.owen@qmul.ac.uk
Greet Van den Berghe[1]
greet.vanden.berghe@kuleuven.be
Fabian Guiza Grandas[1]
fabian.guiza@med.kuleuven.be

[1] Laboratory of Intensive Care Medicine,
KU Leuven
UZ Herestraat 49, box 7003
3000 Leuven
Belgium

[2] Translational Medicine and
Therapeutics
William Harvey Research Institute
Barts and the London School of
Medicine and Dentistry
Charterhouse Square
London EC1M 6BQ

Abstract

Interpretation of histological images is often subjective and time-consuming. Automated approaches offer efficiency and objectivity along with increased precision in results. We focus on the automated quantification of trabecular bone histology images, which are used to study effects of critical illness in a rabbit model. The amount and ratio's of immature and mature bone tissue is of particular interest. We quantify these different tissues through a per-pixel classification model using HSV values. These results are then smoothed with a median filtering step and refined through spatial constraints encoding domain knowledge in order to arrive at a well-performing final segmentation. This allows precise staining areas to be calculated, and individual stain properties to be measured.

1 Introduction

Over the past decades, improvements in intensive care medicine has allowed a growing number of patients to survive previously lethal insults such as surgery, trauma or burn injury. However, these advances also result in a larger population of patients that are in a state of prolonged critical illness, remaining dependent on vital organ support for weeks or months [5, 7].

Prolonged critically ill patients also have distinct alterations in their bone metabolism, which can result in pronounced bone loss, impaired traumatic or surgical fracture healing, and osteoporosis [8]. Our group studied bone degradation in a previously developed rabbit model for critical illness. Hundreds of Masson Trichrome-stained images were obtained from 15 critically ill and control rabbits at 10x magnification. These images show sections containing both the immature bone (dark red) and demineralized, mature bone (blue); some

examples are shown in figure 1. The relative presence of these types of tissue needs to be accurately quantified to properly assess the effects of critical illness on bone metabolism [9].

However, manually evaluating these images poses known difficulties. Firstly, the amount generated is often too large for an in-depth evaluation, so in practice all images are often evaluated on a very rough scale (e.g. a discrete rating of 0-3 for presence of a tissue type). Even so, evaluating these images is very time consuming. Secondly, it is very hard for human evaluators to give an objective judgment due to the many different structures and color levels present. Often, there exists some variation in interpretation given by different evaluators, or by one evaluator over a period of time. One of the main problems is distinguishing immature bone from bone marrow. The former is dark red and adjacent to blue-stained mature bone, while the latter is present across the entire image and has a stained color ranging from dark to bright red.

There exist some semi-automated approaches, such as measuring the fraction of image pixels that crosses a threshold level on one of the images color channels (i.e. detecting a red staining by quantifying the amount of pixels above a certain threshold on the red channel). These methods might accelerate the process slightly, but still require a significant amount of time and contain a degree of subjectivity.

Automatically processing these images is non-trivial because of the large variation in morphology presented by the bone tissue. The structures are densely present in the image, unpredictably positioned, shaped and sized, with variable staining intensity and imperfect separation between the tissues' stain colors.

We have opted for a machine learning method as they are known to be well suited to detecting the underlying properties of large datasets (visual or otherwise). This knowledge can then be applied in classifying new, unseen examples. Such methods have been shown to have good results in the histological domain [4, 6].

We describe a methodology to segment these images in a fully automatic way by combining pixel-level classification through a machine learning model with larger scale morphological domain knowledge. We compare this to existing semi-automated methods for trabecular bone quantification such as described above and with manual scoring by domain experts.

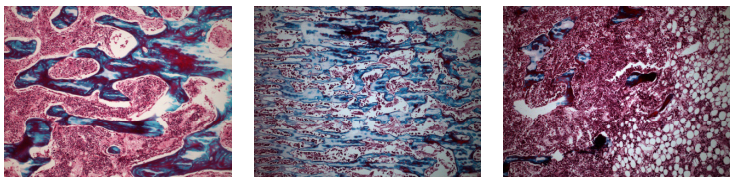


Figure 1: Some examples of histological bone images.

2 Methodology

Our framework consists of two parts: a pixel-wise classification step and morphological post-processing; we discuss each step in turn below. This approach can be contrasted with the semi-automatic one, where the user splits the image into its red, blue and green channels, and defines an image-specific threshold value for each channel, in order to extract the red and blue stained regions. In such an approach no steps are taken to avoid the inclusion of similarly-stained bone marrow in the immature bone area count, often resulting in an

overestimation of the latter's area. The two main advantages our method offers are the full automation, and the capacity to take the above mentioned tissue difference into account.

2.1 Pixel-wise Classification

A suitably representative image is selected from the dataset and used to train the classifier. We assign all its pixels a class label $C = \{Red, Blue, Other\}$. We choose these color-based categories over tissue type because it is not yet possible at the single pixel stage to make a full distinction between immature bone and marrow. A random forest classier [2] is then trained on a randomly selected subsample of $n = 10^5$ pixels, as larger subsets were not found to introduce any meaningful improvement in model performance.

In this model, each pixel is characterized by its *Hue*, *Saturation* and *Value* components. Though this is a very limited feature set, the high level of class information contained in a pixel's color values results in a classification accuracy of 94.3%, measured by 10-fold cross-validation. Though this evaluation was only performed on one image, the results extrapolate well to others in the dataset, as we discuss in sections 3 and 4.

2.2 Post-processing

After the initial pixel classification a number of post-processing steps are applied to improve the accuracy of the result. The pixel-wise nature of the classification leads to many isolated pixel classifications, where the majority of its neighbors have a different class. These single pixels can be assumed to be noise, since the desired cellular structures are presumed to have an area on a scale of hundreds of pixels. Therefore we apply a median filtering step, giving each pixel the class of the majority of its neighbors in an $(2f+1)$-by-$(2f+1)$ square around it, where f typically is 1 or 2. This way, each pixel is assigned the most common class value present it the set of itself and its 8-connected neighbors.

After this step, we impose a minimum and maximum size for identified bone tissue structures. On the advice of the domain expert, we disregard all segments identified as immature and mature bone under size 50px and 300px, respectively.

Finally we impose spatial constraints by only recognizing *Red* segments as being immature bone if they are directly adjacent to (i.e. have a shared border) with a remaining *Blue* segment. This is necessary as it eliminates the visually very similar bone marrow segments, that were also identified as belonging to the *Red* class.

After this post processing, we obtain our resulting segmented image. These steps are summarized in figure 2.

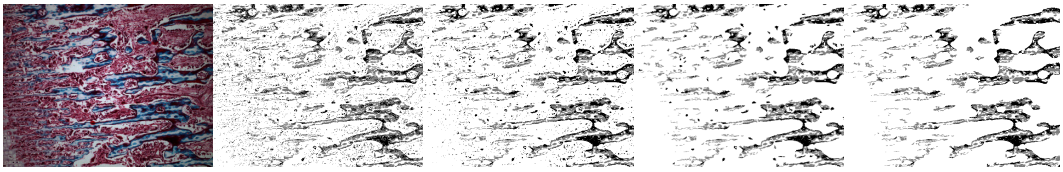
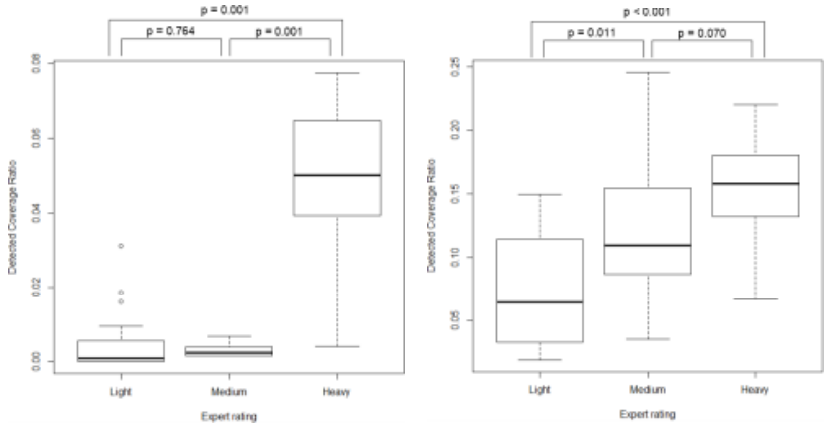


Figure 2: Illustration of the different parts of our method: Unprocessed, after pixel classification, after median filtering, after eliminating small red and blue segments, and finally after elimination based on spatial constraints.

3 Application

Our framework has been applied to the quantification of 341 trabecular bone images taken from 15 critically ill rabbits. These processed images were evaluated and approved by domain experts. Sections from the right proximal tibia from 15 healthy controls and prolonged critically ill rabbits were stained by Masson's Trichrome stain in order to analyze the amount of demineralized, mature bone vs. unmineralized osteoid or immature bone. These slides were then imaged using fluorescence microscopy.

Of these images, 36 were also scored visually by three experts from our lab for presence of immature and mature bone tissue, rated in discrete categories {*light*, *medium*, *heavy*}. Figure 3 plots the results of our analysis against these scores to give an indication of the agreement between these discrete expert evaluations and the continuous values derived with our method. As the figure shows, classes are differentiated well (with a significance level of $p = 0.05$).



(a) Immature bone (b) Mature bone

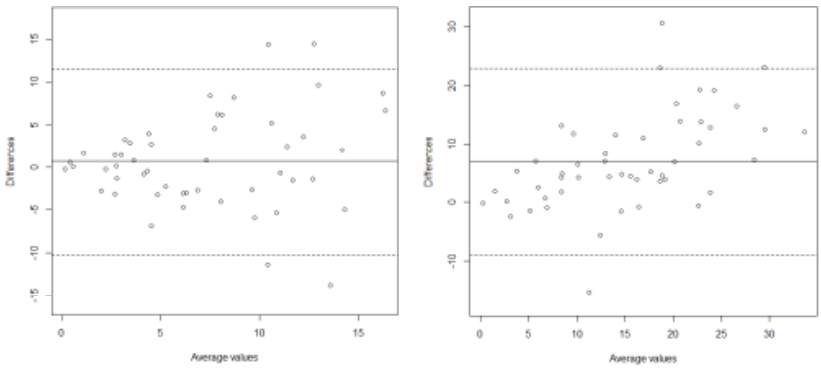
Figure 3: Human classification (median of three) of 36 trabecular bone images vs. relative image coverage area as quantified by our method.

We also need to note that there exists a non-negligible variability in opinion between these experts. We measure the inter-observer agreement by using Fleiss' Kappa statistic [3]. A value of $\kappa = 0$ indicates no better agreement that would happen by random chance, while $\kappa = 1$ represents perfect agreement. For our three raters the quantification of the presence of mature (blue) and immature(red) bone, rated with the three categories mentioned above, is $\kappa_{Mature} = 0.795$ (95 % CI = [0.760,0.831]) and $\kappa_{Immature} = 0.289$ (95 % CI = [0.251, 0.314]).

To compare our method to the more time-intensive manual threshold-based approach, where the user selects appropriate thresholds on the image color channels, we use Bland-Altman plots [1], shown in figure 4 for 50 images.

4 Discussion

We analyzed 341 images for the presence of mature and immature bone and compared this to the evaluation given by experts with both a qualitative and semi-automated quantitative method.



(a) Immature bone (b) Mature bone

Figure 4: Bland-Altman plots for (a) immature (b) and mature bone. The horizontal axis represents percentage of image area covered (mean of both measurements), while the vertical axis shows the difference in measurement values (manual - automatic).

The results of this method were approved by domain experts and used for quantification of bone tissue from a rabbit model [9], as they were determined to be more accurate than both other methods.

This is confirmed by the plots in figure 4. They show that the manual threshold-based method generally gives slight overestimation of immature bone tissue area are also slightly higher (0.68% more, figure 4(a)) and they diverge in proportion to the image area covered. It also gives a higher estimate for the mature bone tissue area (6.95% more, figure 4(b)). However, this method was designed as an ad-hoc solution, and shouldn't be considered a gold standard. As simple thresholding only takes into account pixel color and no structural information, in some cases it can be impossible not to over- or underestimate the areas involved. Since our results were visually confirmed to be more accurate by the domain experts, these plots can be seen as evidence of the limited accuracy of even this more quantitative manual method.

If we compare our method to the more granular categorical division made, we notice a fair correspondence with the given ordinal labels (see figure 3). Though this distinction is worse for the light and medium density immature bone, the low Fleiss' kappa value there indicates disagreement among the different raters as well.

Some user interaction is required in setting up the pipeline, as the post-processing steps have to conform to the domain requirements, and the user has to be aware of their operation and possibilities. However, once this pipeline is set up the algorithm can robustly classify large batches of images. By using relatively simple features as we did in our experiment, hundreds of images can be processed in a matter of minutes.

Despite only being trained and evaluated on one image, the robustness of the model is demonstrated by its good performance on a wide variety of images with unequal staining intensity and variable lighting. Though perhaps a wider variety in training images might increase the performance even further, the fact that an expert only need label one image to get usable results speaks to the user-friendliness of the algorithm.

To refine the results of our method, a logical next step would be to introduce some new features and perhaps redefine the classes to map onto the tissue types instead of color. Textural or spatial information would also be beneficial.

5 Conclusions and Future Work

We have developed a extensible framework capable of using information at the pixel scale and morphological domain knowledge for detecting specific regions of interest in Masson Trichrome-stained trabecular bone histological images. The tissue areas detected matched expert opinion and were used to evaluate a rabbit model experiment [9].

We envision the possibility of this processing pipeline being of use in quantifying other images with large variation in shape and size of regions of interest. The pixel-wise classification is independent of overall region shape of size, while the morphological rules after this step do allow selections of regions based on their spatial properties and relation to other regions.

References

[1] Martin J Bland and Douglas G Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476):307–310, 1986.

[2] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[4] Lei He, L Rodney Long, Sameer Antani, and G Thoma. Computer assisted diagnosis in histopathology. *Sequence and Genome Analysis: Methods and Applications*, pages 271–287, 2010.

[5] Greet Hermans, Ilse Vanhorebeek, Sarah Derde, and Greet Van den Berghe. Metabolic aspects of critical illness polyneuromyopathy. *Critical care medicine*, 37(10):S391–S397, 2009.

[6] Nikita Orlov, Josiah Johnston, Tomasz Macura, Lior Shamir, and Ilya Goldberg. Computer vision for microscopy applications. *Vision Systems: Segmentation and Pattern Recognition*, pages 221–242, 2007.

[7] Greet Van den Berghe, Francis de Zegher, and Roger Bouillon. Acute and prolonged critical illness as different neuroendocrine paradigms 1. *Journal of Clinical Endocrinology & Metabolism*, 83(6):1827–1834, 1998.

[8] Greet Van den Berghe, David Van Roosbroeck, Philippe Vanhove, Pieter J Wouters, Lutgart De Pourcq, and Roger Bouillon. Bone turnover in prolonged critical illness: effect of vitamin d. *Journal of Clinical Endocrinology & Metabolism*, 88(10):4623–4632, 2003.

[9] Ine Vanhees, Jan Gunst, Thomas Janssens, Andy Wauters, Erik Van Herck, Sophie Van Cromphaut, Greet Van den Berghe, and Helen C Owen. Enhanced immunoreceptor tyrosine-based activation motif signaling is related to pathological bone resorption during critical illness. *Hormone and Metabolic Research*, 63(12):862–869, 2013.