Fully automated classification of bone marrow infiltration in low-dose CT of patients with multiple myeloma based on probabilistic density model and supervised learning

Francisco Martínez-Martínez a,*, Jan Kybic a, Lukáš Lambert b, Zuzana Mecková b

a Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic
b Department of Radiology, First Faculty of Medicine, Charles University in Prague, Czech Republic

ARTICLE INFO

Article history:
Received 11 August 2015
Accepted 1 February 2016

Keywords:
Multiple myeloma
Classification
Low-dose CT
Bone marrow CT
Femur

ABSTRACT

This paper presents a fully automated method for the identification of bone marrow infiltration in femurs in low-dose CT of patients with multiple myeloma. We automatically find the femurs and the bone marrow within them. In the next step, we create a probabilistic, spatially dependent density model of normal tissue. At test time, we detect unexpectedly high density voxels which may be related to bone marrow infiltration, as outliers to this model. Based on a set of global, aggregated features representing all detections from one femur, we classify the subjects as being either healthy or not. This method was validated on a dataset of 127 subjects with ground truth created from a consensus of two expert radiologists, obtaining an AUC of 0.996 for the task of distinguishing healthy controls and patients with bone marrow infiltration. To the best of our knowledge, no other automatic image-based method for this task has been published before.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Multiple myeloma is a clonal plasma cell disorder that results in infiltration of bone marrow and osteolytic lesions of the skeleton. The diagnosis and staging of multiple myeloma are based on blood and urine tests, bone marrow biopsy and imaging (radiography, CT, PET-CT and MRI) [1,2]. The treatment is commonly initiated when the disease becomes symptomatic, usually by skeletal involvement, which can be detected by imaging.

Computer aided diagnostic tools exist for other types of cancer such as breast [3], lung [4] or large bowel cancer, and also for other manifestations of multiple myeloma, such as osteolytic lesions [5] and spine bone lesions [6].

There are non-imaging automatic methods to diagnose multiple myeloma using gene expression profiles of plasma cells [7,8] or biomarkers from blood serum samples [9]. This test is routinely used for screening but is by itself not sufficient for the diagnostics.

We are not aware of an automatic image based tool targeting multiple myeloma related to bone marrow infiltration detection in long bones. This task is currently performed manually by radiologists [10,11]. An automated image analysis method suggested in this study has the potential to accelerate the analysis, improve its reproducibility and reduce the radiologist’s workload.

We have chosen CT for the following reasons: Compared to 3D imaging methods, plain radiography is insensitive to the detection of bone marrow infiltration [12–14]. MRI performs best of all imaging methods, especially in the vertebral column [15–17]. However, compared to CT, MRI is slow, more burdensome, expensive and less available. The value of PET is incremental, consisting of the assessment of increased metabolic activity in the tumoral mass [18,19].

In this work, we use a multidetector CT which offers a good trade-off between the cost, acquisition speed and diagnostic performance [20]. We have used a low-dose CT with hybrid iterative reconstruction technique [21,22] in order to reduce the radiation dose to the patient. However, such images are noisier than in standard dose CT which makes the automatic analysis more challenging.

Bone marrow infiltration is difficult to assess in axial skeleton by CT due to the presence of calcium in cancellous bones unless spectral decomposition is performed [23]. Therefore, we chose to evaluate bone marrow infiltration in the medullary cavity of long bones that has low calcium content and in adults it should contain predominantly low density fat. We chose the femur (thigh bone) as it is the largest cortical bone in humans.
1.1. Proposed method overview

First, the bone marrow is segmented (Section 2.1) and femurs are identified (see Fig. 1 for a flow chart). Infiltrated bone marrow appears bright. However, the density itself cannot be used to reliably detect the lesions, as there are other features of the same density such as the red bone marrow, inhomogeneity of the yellow bone marrow, non-neoplastic lesions (bone infarction), and artifacts (beam hardening artifact and image noise). Multiple myeloma presents multiple foci of bone infiltration that may coalesce (Fig. 2).

In addition, the lesion appearance is very variable and the number of positive cases available for training is limited. To address these issues, we do not attempt to model the appearance of the lesions. Instead, we create a probabilistic, spatially dependent model of bone-marrow density in healthy femurs. High density voxels, which may be related to the infiltrated bone marrow, are detected as outliers of this model. A set of global, aggregated features is calculated, representing all detections. In turn, these features are used to classify each subject as either healthy or diseased.

2. Methods

2.1. Bone marrow segmentation

We start by obtaining a binary mask $M_c$ of the cortical (compact) bone by thresholding at 250 HU. (See Fig. 3.) In each $xy$ 2D slice, we find the bone marrow mask $M_{bm}$ (Fig. 3 top-right, $xz$ slice shown) using connected component analysis [24], as a region enclosed by pixels in $M_c$. The two biggest 3D connected components of $M_{bm}$ correspond each to the bone marrow mask $M_f$ of a single femur, which are processed separately. This method is simple and fast, yet sufficiently robust – we could successfully segment all subjects in our dataset.

2.2. Femur alignment

We align the femurs vertically, by performing the principal component analysis (PCA) [25] on the 3D coordinates of the femur bone marrow voxels $M_f$. A bounding box enclosing $M_f$ is rotated...
such that the principal eigenvector of the covariance matrix is aligned with the \( z \)-axis. The rotation along the \( z \)-axis is not uniquely determined but this is not a problem, since subsequent processing (Section 2.5) is invariant in this aspect.

2.3. Region of interest

A ROI (Fig. 3 bottom-right) to limit further processing is obtained by fitting the smallest possible axis-parallel cuboid to the rotated bone marrow mask \( M_f \) and removing 15% in the \( z \)-direction at both ends. Removing femur extremities reduces the number of false positives, as they contain high density cancellous (trabecular) bone.

2.4. Probabilistic density model

The standard approach of building probabilistic models of both healthy and diseased femurs turned out to be unsuitable, because of the great variability of the lesions (infiltrations) and the relatively small amount of training data available. Instead, we model the density only in the more homogeneous population of healthy femurs, and assume that infiltrations are outliers of this model. The density generally increases outwards from the center in both axial and longitudinal directions. In order to obtain a generic, scale-independent model, the Euclidean coordinates \( r \) are converted to normalized radial and longitudinal coordinates \((l, t) = \phi(r)\), with \((l, t) \in [0, 1] \times [0, 1]\), see below. The probability density of observing a value \( f \) at a spatial location \( r \) is then modeled as \( p(f | r) \approx q(f | l, t) \).

2.5. Spatial normalization

Let us describe the mapping \((l, t) = \phi(r)\), \( r = (x, y, z) \). In the \( z \)-direction we simply normalize the position to the interval \([0, 1]\):

\[
l(z) = (z - z_{\text{min}})/(z_{\text{max}} - z_{\text{min}})
\]

where \( z_{\text{min}} \) and \( z_{\text{max}} \) are limits of the ROI in the longitudinal direction (Fig. 3 bottom-right).

The radial coordinate \( t \) is the relative distance of the point \( r \) to the boundary \( B \) of the bone marrow mask \( M_f \) for each \( xy \) slice:

\[
t(x, y) = \frac{b = B}{\max(d(r, B))}
\]

Fig. 4 shows an example of femur bone marrow densities in the Euclidean and the \((l, t)\) spaces.

2.6. Model representation and outlier detection

The probability \( q(f | l, t) \) is represented non-parametrically, as a set of histograms for \( l \) and \( t \) quantized into 10 \times 10 uniform bins (Fig. 5). In each histogram, the density \( f \) is uniformly quantized into 2048 bins in the interval \([-1024, 1024]\) HU. The histograms are summed over all training images and normalized.
The cumulative sum of each histogram leads to an estimate of the probability

\[ Q(f' | l, t) = P[f(r) > f'] \quad \text{with} \quad (l, t) = \phi(r) \]  

(3)

of observing a density higher than \( f' \) at a position \( r \).

At test time, if the observed value \( f' \) is too high to be explained by our model

\[ Q(f' | l, t) < \zeta_0 \]  

(4)

then we consider the voxel to be a detection, where \( \zeta_0 \) is a parameter defining the significance level (Fig. 6).

### 2.7. Subject classification

We classify subjects instead of individual voxels as only subject-level labels are available to us. This also helps to average out pixel-level false positives from the outlier detection step, Eq. (4). We use a simple set of global features aggregating all detections for a subject. The features are the total volume of detections,
and the total number of connected components of the detections, for both femurs.

These two features are then used in a classifier (a $k$-NN [27] or a soft-margin SVM [28]) to determine whether a subject is healthy or not. The features are normalized to zero mean and unit variance.

To accommodate the different a priori probabilities of the classes, we consider the discrimination function threshold $\xi$ as an additional parameter. In $k$-NN, the discrimination function is the relative number of nearest neighbors of the foreground class in a neighborhood, with a default threshold $\xi_{kNN} = 0.5$. The other parameter is the number of nearest neighbors $k$ to be considered. For SVM, the default threshold is $\xi_{SVM} = 0$. Other parameters are the standard deviation of the Gaussian kernel $\sigma$ and the soft margin coefficient $C$ [28]. The classifier parameters are learned from the training data by random permutation cross validation, as explained below.

## 3. Experiments

### 3.1. Experimental setup

CT images of 116 patients (67 ± 10 years, Table 1) with monoclonal plasma cell disorders as detected by a blood test were evaluated by two radiologists with experience in skeletal imaging. They independently classified these subjects into three categories: “With bone marrow infiltration”, “without bone marrow infiltration” and “not clear”. The interobserver agreement for bone marrow infiltration when assessed by the two readers was $k = 0.77$ (good agreement). The two experts then agreed on a consensus classification for all subjects. Subjects classified as “not clear” were discarded, as well as subjects with prosthesis (which produces artifacts in the images). We added to the dataset 53 healthy controls (68 ± 12 years) that consisted of randomly selected subjects with suspected peripheral arterial disease who underwent CT of the aorta and lower limbs, without monoclonal plasma cell disorder. The difference in age between the study group and healthy controls were not statistically significant according to the $t$-test ($p = 0.18$).

This way, we obtained a dataset that consisted of 127 subjects with 254 femurs (Table 2). We divided it into three groups: 51 healthy controls formed the group A, 39 subjects with monoclonal plasma cell disorder (monoclonal gammapathy of uncertain significance – MGUS or multiple myeloma – MM) but without observed bone marrow infiltration formed the group B, and 37 subjects with monoclonal plasma cell disorder and with observed bone marrow infiltration formed the group C.

Apart from the patients with MGUS, there were also patients who underwent treatment previously, some with osteolytic lesions elsewhere in the skeleton, or symptomatic by renal insufficiency, anemia, or hypercalcemia. As shown by the latest International Myeloma Working Group (IMWG) Criteria for the Diagnosis of Multiple Myeloma, bone marrow in patients with multiple myeloma may have initially even less than 10% of clonal plasma cells. Moreover, biopsy from the pelvic bone may not show bone marrow infiltration if it is present somewhere else than in the pelvic bone. These subjects are also included in group B.

The image acquisition was performed on a 256-slice scanner (Brilliance iCT 256; Philips Healthcare, Best, The Netherlands). The

![Flow chart that represents the training and testing methodology. The diagram represents one random split. Blue and yellow arrows stand for T1 and T1' datasets, respectively. Dashed and solid boxes stand for input/output data and methods, respectively. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)(fig:flow_chart)

### Table 1

Characteristics of the study group (control group not included). ISS stands for international staging system, MM for multiple myeloma, and MGUS for monoclonal gammopathy of uncertain significance.

<table>
<thead>
<tr>
<th>ISS</th>
<th>1: 43%</th>
<th>2: 19%</th>
<th>3: 38%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Durie-Salmon stage</td>
<td>I: 30%</td>
<td>II: 24%</td>
<td>III: 46%</td>
</tr>
<tr>
<td>Type</td>
<td>MGUS: 3.64%</td>
<td>MM: 96.36%</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>After treatment: 68% without treatment: 32%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2

Details about the groups within the dataset. MPCD stands for monoclonal plasma cell disorder and BMI for bone marrow infiltration.

<table>
<thead>
<tr>
<th>Group</th>
<th>MPCD</th>
<th>BMI</th>
<th># subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>No</td>
<td>No</td>
<td>51</td>
</tr>
<tr>
<td>B</td>
<td>Yes</td>
<td>No</td>
<td>39</td>
</tr>
<tr>
<td>C</td>
<td>Yes</td>
<td>yes</td>
<td>37</td>
</tr>
</tbody>
</table>

### Table 3

The positive and negative classes for the three binary classification experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Negative class</th>
<th>Positive class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Group A</td>
<td>Group C</td>
</tr>
<tr>
<td>2</td>
<td>Group A</td>
<td>Groups B + C</td>
</tr>
<tr>
<td>3</td>
<td>Groups A + B</td>
<td>Group C</td>
</tr>
</tbody>
</table>
Three experiments were carried out (Table 3). Experiment 1 aimed to distinguish the healthy controls (group A) from subjects with bone lesions (group C). It did not attempt to classify the uncertain group B, where the bone marrow may not be entirely normal, but the changes are not yet pronounced enough to be classified as infiltrations in the clinical review. In Experiment 2, we attempted to distinguish group A from groups B and C, testing the relationship between the CT image analysis and the presence of monoclonal antibodies. In Experiment 3, we tried to distinguish group C from groups A and B, the task of differentiating hematlogy patients with and without bone marrow infiltration.

The matrix size was 512 × 512 with a voxel size of 0.976 × 0.976 × 0.450 mm.

Three experiments were carried out (Table 3). Experiment 1 aimed to distinguish the healthy controls (group A) from subjects with bone lesions (group C). It did not attempt to classify the uncertain group B, where the bone marrow may not be entirely normal, but the changes are not yet pronounced enough to be classified as infiltrations in the clinical review. In Experiment 2, we attempted to distinguish group A from groups B and C, testing the relationship between the CT image analysis and the presence of monoclonal antibodies. In Experiment 3, we tried to distinguish group C from groups A and B, the task of differentiating hematlogy patients with and without bone marrow infiltration.
3.3. Fixed parameters

We fixed the classifier parameters to the mean values from Tables 4 and 5 and evaluated the classification using the random permutation cross validation method described above. Fig. 10 shows the mean ROC curves to illustrate the performance drop with fixed parameters instead of using the best parameters per iteration. This demonstrates the benefit of learning the classifier on training data similar to the test data. Nevertheless, as shown in Table 6, the performance with fixed parameters remains acceptable. In fact, the difference between the fixed and best parameters as measured by the differences in the AUC values are not statistically significant according to the t-test ($p > 0.05$), except in Experiment 1 (see $p$-values in Table 6). Similarly, the difference between the $k$-NN and SVM classifiers using the best parameters is also not statistically significant except for Experiment 2. The difference between both classifiers using the fixed parameters is statistically significant except for Experiment 2.

4. Discussion

The presented method can be readily extended in several ways provided that suitable annotated data is available. First, we assume that this approach can be extended to other long bones besides the femur, such as the humerus and the tibia. The extension to the forearm bones should also be possible but more difficult because of their narrower medullary cavity.

Second, the intermediate results of our method include tentative voxel-wise detections (Fig. 6). Therefore, it should be possible to count the number of lesions, determine their extent and distinguish between focal and diffuse lesions (Fig. 11). This can be used for the assessment of the therapeutic response.

Third, the classification might be improved by including other modalities, namely MRI and PET. It is true that MRI has superior soft-tissue contrast. However, neither CT nor MRI does perfectly discriminate between the replacement of fat by hematopoietic cells and bone marrow infiltration. In PET, the increased metabolic activity in bone marrow occurs both in the bone marrow reconversion and infiltration. In fact, the three modalities often give very similar information and the small additional gain might not be worth the additional cost, radiation dose, and discomfort to the patient (Fig. 12).

Another important aspect of the problem of detection of multiple myeloma is the lack of reliable ground truth. The experts themselves were unable to classify 40 images out of 116 patients and for the remaining ones their first assessments differed in 8 cases (just from the patients classified as “with bone marrow infiltration” and “without bone marrow infiltration” by consensus). Especially problematic is the group B (positive blood test but no clearly visible bone marrow infiltration) which is half way between the healthy controls and patients with bone marrow infiltration visible on CT. In fact, some subjects from group B were diagnosed with multiple myeloma by means of positive bone marrow histology from the pelvic bone. On the other hand, the bone marrow infiltration may partially or fully disappear in patients successfully treated for multiple myeloma. Therefore, it is not surprising that for Experiments 2 and 3, which contain the group B, the performance is weaker. Although changes in bone

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Experiment 1</th>
<th>Experiment 2</th>
<th>Experiment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ξ_{NN}</td>
<td>0.659 ± 0.241</td>
<td>0.700 ± 0.244</td>
<td>0.510 ± 0.190</td>
</tr>
<tr>
<td>ξ_{SVM}</td>
<td>0.067 ± 0.261</td>
<td>0.049 ± 0.371</td>
<td>0.210 ± 0.287</td>
</tr>
</tbody>
</table>

Fig. 10. Comparison of mean ROC curves from $k$-NN and SVM using the best and fixed parameters for Experiments 1–3.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Experiment 1</th>
<th>Experiment 2</th>
<th>Experiment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k-NN</td>
<td>SVM</td>
<td>k-NN</td>
</tr>
<tr>
<td>Best</td>
<td>AUC</td>
<td></td>
<td>AUC</td>
</tr>
<tr>
<td></td>
<td>0.995 ± 0.017</td>
<td>0.996 ± 0.009</td>
<td>0.907 ± 0.073</td>
</tr>
<tr>
<td></td>
<td>Sensitivity</td>
<td>0.983 ± 0.047</td>
<td>0.977 ± 0.060</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>0.988 ± 0.033</td>
<td>0.998 ± 0.014</td>
</tr>
<tr>
<td>Fixed</td>
<td>AUC</td>
<td>0.987 ± 0.017</td>
<td>0.967 ± 0.050</td>
</tr>
<tr>
<td></td>
<td>Sensitivity</td>
<td>0.823 ± 0.157</td>
<td>0.900 ± 0.120</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>0.958 ± 0.061</td>
<td>0.912 ± 0.087</td>
</tr>
<tr>
<td>p value (best-fixed)</td>
<td>AUC</td>
<td>0.021</td>
<td></td>
</tr>
</tbody>
</table>
marrow are rarely encountered in patients with MGUS, it is important to know whether the algorithm can detect the disease from the CT images alone. This in turn could reduce the need for the painful bone marrow biopsy. It is known that some changes in the bone marrow can be detected by MRI even in patients with MGUS [30]. In order to evaluate the influence of the patients with MGUS, we have carried out experiments 2 and 3 again, excluding patients with MGUS. In fact, all but one of the MGUS patients have been already excluded from the analysis, either because of a lack of consensus between experts, or because of the presence of artifacts due to hip implants. The differences as measured by AUC were not significant, for example the AUC in Experiment 2 with SVM changed from 0.941 ± 0.072 to 0.943 ± 0.056, in Experiment 3 from 0.865 ± 0.079 to 0.852 ± 0.103.

For Experiment 1, where the positive and negative groups are defined more unambiguously (whole group B is excluded, including patients with MGUS), the classification is almost perfect. Both hematological diseases and their therapy may result in reconversion of the bone marrow. This however occurs in a predictable direction from the proximal metaphysis towards the diaphysis and typically does not form heterogeneous appearance with islands of bone marrow infiltration.

We have refrained from directly comparing the performance of human experts and the computer classifier. This was done on purpose as the tasks are not really comparable. The human experts have several advantages – they do not classify the healthy controls, they can refrain from giving a decision, and they determine the

---

**Fig. 11.** Diffuse (top) and focal (bottom) detections obtained for two different subjects.

**Fig. 12.** T2 fat saturated (STIR) MRI image of the distal half of the right femur shows patchy areas of increased signal intensity consistent with infiltration of bone marrow (left). CT (middle) and PET-CT (right) images acquired 10 days later show increased density and increased accumulation of 18F-FDG in the bone marrow of a 56-year-old patient with relapse of multiple myeloma (IgG kappa, stage IIA).

**Fig. 13.** Comparison between our previously published results [31] and the present work in Experiment 3.
ground truth. We believe that both issues will be mitigated as the amount of available training data increases, including the number of independent readers. In the future, we might also want to allow the algorithm to give a neutral answer, instead of forcing a binary decision.

The results showed that the classification using k-NaN of features obtained from detections provided by a probabilistic density model allowed differentiation of healthy and diseased subjects with good sensitivity and specificity: Around 0.98 for Experiment 1 (differentiation between healthy controls and subjects with bone marrow infiltration). This compares very favorably with the value of 0.867 for a method based on blood serum testing [9].

We are supported by PRVOUK-P27/LF1. The clinical part was supported by PRVOUK-P27/LF1. We are

References


Francisco Martínez-Martínez holds a Ph.D. in technologies for health and well-being from the Polytechnic University of Valencia, 2014. He is currently a postdoc research fellow at the Biomedical Imaging Algorithms Group, from the Center for Machine Perception, Department of Cybernetics, Czech Technical University in Prague. He was research assistant at LabHuman from the Polytechnic University of Valencia, 2009–2014. He was a visiting research assistant at the Simulation and Modelling in Medicine and Surgery Research Group from the Department of Surgery and Cancer, Imperial College of London, 2013 and at the Biomechanics Group from the Institute for Mechanical Systems, ETH Zurich, 2012. He has tens of contributions in international conferences and papers in journals with JCR.

Jan Kybic obtained a Ph.D. degree in biomedical image processing from EPFL, Switzerland, in 2001. In 2002–2003, he held a post-doc research position in INRIA, Sophia-Antipolis, France. Since 2003 he is with Faculty of Electrical Engineering, CTU in Prague. In 2010–2011 he was as a visiting scientist at EPFL, Switzerland. He passed his habilitation (associate professor) in 2011. In 2011–2013 he was a vice-dean for Information Technology of the Faculty of Electrical Engineering. Since 2013 he is a head of the Department of Cybernetics. Jan Kybic has authored or co-authored 24 articles in peer-reviewed international scientific journals, one book, and over 75 conference publications. He has 877 citations and an H-index 13 (WoS). He served until 2014 as an associated editor of the IEEE Transactions on Medical Imaging and is a general co-chair of the International Symposium on Biomedical Imaging 2016.

Lukáš Lambert graduated from the Third Faculty of Medicine, Charles University in Prague in 2004 and from the Faculty of Electrical Engineering, Czech Technical University in Prague in 2007. In 2014 he defended Ph.D. Currently he is the head of CT Department at the Department of Radiology, First Faculty of Medicine in Prague.