Bone Tumor Segmentation on Bone Scans Using Context Information and Random Forests

Gregory Chu[1], Pechin Lo[1], Bharath Ramakrishna[1], Hyun Kim[1], Darren Morris[2], Jonathan Goldin[1], and Matthew Brown[1]

[1] Center for Computer Vision and Imaging Biomarkers, UCLA Radiology, USA
[2] MedQIA Imaging CRO, USA

Abstract. Bone tumor segmentation on bone scans has recently been adopted as a basis for objective tumor assessment in several phase II and III clinical drug trials. Interpretation can be difficult due to the highly sensitive but non-specific nature of bone tumor appearance on bone scans. In this paper we present a machine learning approach to segmenting tumors on bone scans, using intensity and context features aimed at addressing areas prone to false positives. We computed the context features using landmark points, identified by a modified active shape model. We trained a random forest classifier on 100 and evaluated on 73 prostate cancer subjects from a multi-center clinical trial. A reference segmentation was provided by a board certified radiologist. We evaluated our learning based method using the Jaccard index and compared against the state of the art, rule based method. Results showed an improvement from 0.50 $\pm$0.31 to 0.57 $\pm$0.27. We found that the context features played a significant role in the random forest classifier, helping to correctly classify regions prone to false positives.

1 Introduction

Metastasis to bone occurs in nearly all patients with advanced forms of the most common human cancers – breast, lung, and prostate [1]. The extent of bone metastasis strongly associates with shorter survival times as well as a degradation in quality of life [2,3]. Whole-body bone scan is a highly sensitive nuclear medicine technique for visualizing bone tumors and is the accepted standard imaging modality for assessment. Recent studies have shown nontrivial differences in physician interpretation of bone scans [4]. This motivates the need for an automated bone tumor segmentation method aimed at reducing the significant time and variability of hand-annotated bone scan analysis. The aim is to provide a method for objective and reproducible bone scan tumor measurements, and a foundation for their potential correlation with other clinical measures.

Interpretation can be difficult due to the highly sensitive but non-specific nature of bone tumor appearance on bone scans [5]. There are several factors to this. Firstly, the image intensity is proportional to the osteoblastic activity in bones. This is a marker for bone tumor growth – however, it is also a marker for degenerative joint disease (DJD), bone fractures, and bone infections. DJD

P. Golland et al. (Eds.): MICCAI 2014, Part I, LNCS 8673, pp. 601–608, 2014.

is very common and can manifest as high intensity in the jaws, neck, shoulders, spine, pelvis, elbows, knees, and ankles. Secondly, the appearance and intensity of a given tumor is bone and location dependent [6,7]. For example, on a given patient scan, a tumor in the humerus may appear fainter than the intensity of that patient's tumor-free spine. Thirdly, a non-trivial proportion of the radioactive tracer often remains in the kidneys, bladder, sinus cavities, thyroid, and catheters. These three factors, combined with the lack of depth discrimination inherent to the two-dimensional nature of the imaging modality, make accurate tumor segmentation challenging. Visual bone scan diagnosis recommendations advise the physician to use contextual information when evaluating an area of high intensity. For example, for a given area of high intensity, if it exists near a joint or the thyroid, it is less likely to be a tumor. Additionally, if the joint has a symmetric counterpart (e.g. shoulders), and if the counterpart has a similar appearance, both areas are likely to be DJD [5,7].

Brown et al. published a region based thresholding method for segmenting tumors on bone scans, and used it to compute a tumor area measure that has since been adopted as a primary outcome measure in several phase II and III clinical drug trials [8,9]. Chu et al. extended this method by adding several rules to remove false positives and demonstrated state of the art performance [10].

In contrast to the rule based method in the literature, we propose a unified model that uses both intensity and context features in a random forest classifier to segment bone tumors on bone scans. The intensity features, computed at multiple Gaussian scales, describe local neighborhood information. The context features describe a pixel's location relative to the rest of the body, and relative differences in local intensity features between two points. The location information is relative to a set of landmark points on the bone scan. We identify the landmark points by a modified active shape model (ASM) algorithm with histogram of oriented gradient (HOG) features [11,12].

2 Methods

We built a Random Forest classifier to segment metastatic bone tumors using intensity and context features. The steps in the pipeline include pre-processing, landmark detection, feature computation using the landmarks, and classification.

Due to the variation in image sizes and resolutions, we pre-processed the scans by standardizing the resolution by segmenting the whole-body using histogram analysis as described in [10], and resampling the image within a bounding box of the segmentation to a standard space (800 by 200 pixels). We used this resampled image for all subsequent steps.

2.1 Landmark Detection Using an Active Shape Model

We defined 31 landmark points for anterior (AP) and posterior (PA) bone scans that marked salient areas, e.g. the top of the head, the shoulders, the hip joints, the knees, etc. The landmarks are shown in Fig. 1a.

To train the point distribution model (PDM), we first applied a Procrustes analysis to align the shapes, followed by a principal component analysis (PCA) to the set of M aligned shapes. Each shape $\boldsymbol{x}$ was described by 31 points. The mean shape $\bar{\boldsymbol{x}}$, covariance matrix $\boldsymbol{S}$, the matrix of reduced eigenvectors $\boldsymbol{P}$, and the vector of parameters $\boldsymbol{b}$, were calculated as described in [11]. To reduce the noise in the model, we retained only 90% of the eigenvectors. The PDM was represented by

$$b = P^T(x - \bar{x}) \tag{1}$$

To train the template matching ASM, we modified the work in [11], using HOG descriptors [12] to capture the different set of gradient directions for each landmark. The HOG descriptor was calculated at 8 orientations on a 17 by 17 pixel patch around a landmark. The patch was split into 2 by 2 cells of size 8 by 8 pixels. We computed the descriptor as described in [12] using the VLFeat library [13]. We trained the ASM with HOG descriptor at 3 image resolutions: full, half, and quarter. Due to the sparse nature of the HOG descriptor, we reduced the dimensionality using PCA, retaining 90% of the eigenvectors.

When applying the ASM to a new image, the task at each iteration was to find the best suggested movement for each landmark point based on HOG descriptor matching. To do so, we searched horizontally, vertically, and diagonally in a 9 by 9 pixel patch centered on the point. The best new point was defined as the test point that minimized the Mahalanobis distance between this test point's HOG descriptor and the landmark point's trained HOG descriptor. Subsequently, we constrained the set of the suggested new landmark points, contained in $\boldsymbol{b}$, to within the limits of $\pm 1.5\sqrt{\lambda_i}$, where λ_i is the eigenvalue for the i^{th} principal component. We searched at 3-resolutions, with 15 iterations at each resolution, in the following order: quarter, half, full resolution,

2.2 Intensity and Context Features

Due to the relative nature of the image intensity units across scans, we normalized the intensity prior to feature computation. We defined a mask containing the legs and upper arms of the patient, and computed the 75^{th} centile of the intensity histogram in that mask, defined as the normalization intensity value n, similar to [10]. We created the mask by defining rectangular regions using the landmark points as vertices. The intersection of this mask and the whole-body segmentation is shown in Fig. 1b. We then linearly rescaled the entire image by $f(I_i) = (r/n) \times I_i$, where $i = 1, 2..., N$, N is the number of pixels, r is the reference intensity value, I_i is the original intensity at pixel i, and $f(I_i)$ is the normalized intensity. In our experiments we set r to 15.

For each pixel p, we computed a set of 12 common local intensity features including the output of Gaussian filters up to the second order derivatives $(L, L_x, L_y, L_{xx}, L_{yy}, L_{xy})$, the gradient magnitude, the Laplacian, eigenvalues of the Hessian matrix ($k1, k2$, where $k1 > k2$), $k1/k2$, and the L2-norm of $k1$ and $k2$. Radiologists observed that tumors appeared focal and benign uptake appeared diffuse and textured. To capture this, we computed the mean and

range of 4 of the most common gray level co-occurrence matrix (GLCM) properties, i.e. contrast, correlation, energy, and homogeneity, as described in [14]. We computed the GLCM on patches of size 11x11, with intensity values quantized into 8 bins between the minimum and maximum intensity of the patch, at 4 orientations, symmetrically, and with offset distances of a single pixel. To ensure representation of objects with varying size, we computed the previous features on 3 Gaussian smoothed images ($\sigma = 1, 2, 4$ pixels). The use of Gaussian derivatives at multiple scales to describe local image structure was motivated by scale-space theory [15].

The context features consisted of 1) an offset vector between a given pixel, p, and each of the landmark points, q_i, 2) a difference in intensity at p and each of the q_i, and 3) the difference in the set of 12 intensity features between p, and the reflected point about the midline p'. We computed the context features of type (2) and (3) at 3 Gaussian scales ($\sigma = 1, 2, 4$ pixels). We computed p' by 1) reflecting the pixel p across the midline, which we obtained by fitting a line to 8 landmark points associated with the superior-inferior axis, and 2) searching a 17 by 17 pixel patch for the closest match to p. We defined the closest match to be the minimum absolute difference in intensity on a smoothed image ($\sigma = 4$ pixels).

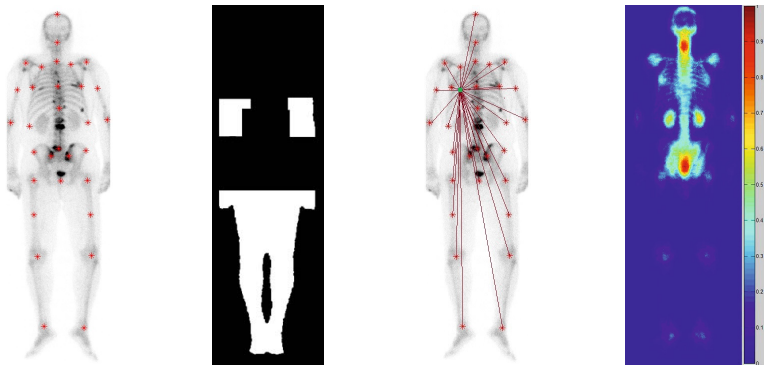
In total, we computed 60 intensity and 191 context features.

2.3 Sampling

Our task was to classify pixels into two classes: tumor (positive) and non-tumor (negative). We obtained the positive samples by randomly sampling pixels from the reference segmentation, while enforcing a minimum distance constraint of 5 pixels between samples in order to reduce the amount of correlation between samples.

An issue with our data was that the number of negative samples, which was everything in the image besides tumor, was very large compared to the number of positive samples. To prevent the classifier from biasing toward the negative cases, we needed to enforce equal numbers of negative and positive samples. Since the number of negative samples was small, we ideally preferred negative samples prone to being false positives, namely, samples corresponding to non-tumor pixels with high intensity.

We constructed a probability map of pixels prone to being false positives, shown in Fig. 1d. To construct the map, we created a mask containing areas excluding tumor but of high intensity, for each image in the training set. The probability map was the normalized sum of the masks across all images in the training set. We set the target number of negative samples per image, K, to be equal to the average number of positive samples per image in the training set. Then, we selected K negative samples per image where each sample was selected using the probability map as the sampling distribution. This resulted in a higher frequency of samples in regions prone to false positives.



(a) (b) (c) (d)

Fig. 1. (a) 31 landmark points on a PA scan. (b) Mask of upper arms and legs. (c) 31 offset vectors from point p (green) to all landmark points. (d) Probability map of regions prone to false positives.

2.4 Random Forest Classification and Segmentation

We used the random forest classifier from the sci-kit learn library to discriminate between classes of tumor and non-tumor [16]. To train the classifier, we first standardized the features by subtracting the mean and dividing by the standard deviation. The classifer used 100 trees, 20 random features at each split, and the Gini impurity measure as a splitting criteria, motivated by [17].

To obtain a segmentation, we thresholded the classifier's probability output. To train the threshold value, we split the training set into two subsets, set A and B. We trained a classifier on set A using the sampling method in Sect. 2.3, and then tested the classifier on all pixels in the images in set B. We then found the threshold that yielded the optimal Jaccard index (JI).

3 Experiments and Results

Our dataset consisted of 213 pairs of anterior (AP) and posterior (PA) bone scans from different subjects, collected from 56 different sites in a large multi-center metastatic prostate cancer drug trial. All scans were acquired 2-4 hours post injection of Tc-MDP. The pixel spacings in our data prior to resampling ranged from 2 to 3 mm. All 213 subjects had a reference tumor segmentation. An initial segmentation was provided by the method in [10], and subsequently, manually edited and reviewed by a board certified radiologist. We randomly split the dataset into a training set of 140 subjects and a testing set of 73 subjects. For 60 of the 140 training subjects, we manually annotated 31 landmark points.

To verify our landmark detection method, we ran a 3-fold cross validation experiment. The mean distance between the points after applying the ASM and the annotated landmarks was 6.3 pixels across all landmarks points and all 60

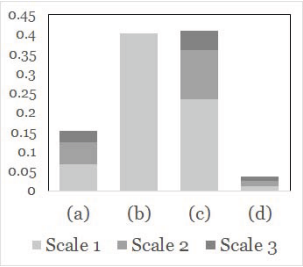


Fig. 2. Gini feature importance by type: (a) intensity, (b) offset, (c) landmark, and (d) symmetry, and further subdivided by Gaussian scale

Table 1. Mean±SD of JI and A_z across all subjects for the rule based state of the art [10], a classifier w/o context features $\text{CLF}_{w/o}$, and the proposed classifier CLF

	JI	A_z
[10]	0.50±0.31	-
$\text{CLF}_{w/o}$	0.49±0.26	0.91±0.05
CLF	0.57±0.27	0.96±0.03

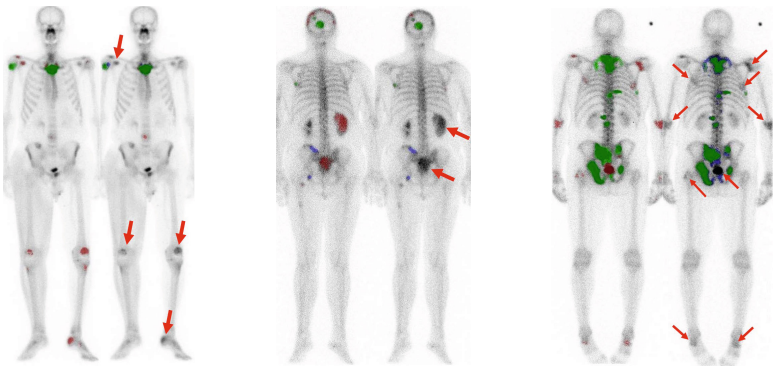


Fig. 3. Segmentations of 3 subjects, showing the rule based method [10] on the left, and the proposed, machine learned method on the right. True positives are shown in green, false positives in red, and false negatives in blue. Areas incorrectly classified by [10] and correctly classified by the proposed method as non-tumor are highlighted with red arrows. Note the ability for the proposed classifier to perform well in areas with tumor and high intensity non-tumor in close proximity, e.g. in the right shoulder of case 1, and the bladder in case 3.

scans. As a point of reference, we annotated the 31 landmarks a second time on 15 patients, and the mean distance between the first and second annotation was 3.1 pixels.

We trained two classifiers for AP and PA images. To train the threshold used to segment the probability output of the classifier, we used 100 subjects for set A and 40 subjects for set B. For the AP and PA classifier, the threshold was 0.87 and 0.81, respectively.

To test the classifiers, we classified all pixels in the image on 73 AP/PA image pairs, and evaluated the segmentation by computing the JI per subject.

We excluded from evaluation a 1-pixel rim on the inside and outside of the reference segmentation. This was to reduce the effect of partial voluming and a bias towards the segmentation output from [10], which we used to initialize the reference segmentation. We also performed a supplemental classification evaluation on a subset of pixels (limited by the number of positives) from each subject, sampled using the method described in Sect. 2.3. This method ensured an equal number of positive and negative samples. By varying the threshold on the probability, we computed the area under the ROC curve, A_z. To account for the randomness in the sampling, we repeated the evaluation 10 times.

We compared the mean performance across the AP and PA scans, and across the 73 test subjects, using the method in [10], a trained classifier without context features ($\text{CLF}_{w/o}$), and our proposed classifier with context features (CLF). Results are summarized in Table 1. We built the $\text{CLF}_{w/o}$ classifier in order to evaluate the influence of context features on the classification.

4 Discussion and Conclusion

In Fig. 2, we see that the random forest classifier found the context features (offset and landmark type features) to be highly important, equaling 81% of the total Gini feature importance. Furthermore, in Table 1, when we compare the JI and A_z for $\text{CLF}_{w/o}$ and the proposed CLF, we observe an increase in JI from 0.49 to 0.57, and a significant increase in A_z from 0.91 to 0.96. This indicates that the context features played an important role in bone tumor classification. We also see that the state of the art performed similarly to $\text{CLF}_{w/o}$, with a JI of 0.50 and 0.49, respectively, suggesting that the context features may have been a significant factor in the overall improvement over the rule based method.

Fig. 3 shows the segmentations of 3 subjects, displaying the state of the art on the left and the proposed method on the right. We see that regions prone to false positives like the shoulders, scapula, elbows, kidneys, bladder, knees and heels are properly classified as non-tumor. Note, in particular, the ability for the proposed classifier to perform well in areas with tumor and high intensity non-tumor in close proximity. This improvement may be due to the restrictiveness of a rule based approach compared to a learning based approach.

Bone tumor segmentation on bone scans is a new area of research with significant clinical impact. In this work, we developed a random forest classifier to segment tumors on bone scans using intensity and context features aimed at addressing areas prone to false positives. We found that context features played a critical role in the classification process. This learning based method demonstrates incremental improvement over the state of the art, rule based bone tumor segmentation method. In future work, we plan to investigate the use of more intelligent methods to segment the classifier probability output.

References

1. Mundy, G.: Metastasis to Bone: Causes, Consequences and Therapeutic Opportunities. Nat. Rev. Cancer. 2, 584–593 (2002)
2. Coleman, R.: Clinical Features of Metastatic Bone Disease and Risk of Skeletal Morbidity. Clin. Cancer Res. 12, 6243s (2006)
3. Sonpavde, G., Pond, G., Berry, W., Wit, R., Eisenberger, M., Tannock, I., Armstrong, A.: The Association Between Radiographic Response and Overall Survival in Men with Metastatic Castration-Resistant Prostate Cancer Receiving Chemotherapy. Cancer 117, 3963–3971 (2011)
4. Sadik, M., Suurkula, M., Hoglund, P., Jarund, A., Edenbrandt, L.: Quality of Planar Whole-body Bone Scan Interpretations – A Nationwide Survey. Eur. J. Nucl. Med. Mol. Im. 35(8), 1464–1472 (2008)
5. Bombardieri, E., Aktolun, C., Baum, R., Maffioli, L., Moncayo, R., Mortelmans, L., Reske, S.: Bone Scintigraphy: Procedure Guidelines for Tumour Imaging. Eur. J. Nucl. Med. Mol. Im. 30, 99–106 (2003)
6. Larson, S., Nelp, W.: The Radiocolloid Bone Marrow Scan in Malignant Disease. J. Surgical Onc. 3(6), 685–697 (1971)
7. Holder, L., Collier, D., Fogelman, I.: An Atlas of Planar and SPECT Bone Scans. CRC Press (2000)
8. Brown, M., Chu, G., Kim, H., Allen-Auerbach, M., Poon, C., Bridges, J., Vidovic, A., Ramakrishna, B., Ho, J., Morris, M., Larson, S., Scher, H., Goldin, J.: Computer-Aided Quantitative Bone Scan Assessment of Prostate Cancer Treatment Response. Nucl. Med. Commun. 33(4), 384–394 (2012)
9. Scher, H., Smith, M., Sweeney, C., Corn, P., Logothetis, C., Vogelzang, N., Smith, D., Hussain, M., George, D., Bono, J., Higano, C., Small, E., Goldin, J., Brown, M., Aftab, D., Noursalehi, M., Weitzman, A., Basch, E.: An Exploratory Analysis of Bone Scan Lesion Area, Circulating Tumor Cell change, Pain Reduction, and Overall Survival in Patients with Castration-Resistant Prostate Cancer Treated with Cabozantinib. J. Clin. Onc. 31(15), 5026 (2013)
10. Chu, G., Lo, P., Kim, H., Auerbach, M., Goldin, J., Henkel, K., Banola, A., Morris, D., Coy, H., Brown, M.: Preliminary Results of Automated Removal of Degenerative Joint Disease in Bone Scan Lesion Segmentation. In: Proc. SPIE 8670 Medical Imaging, 867007 (2013)
11. Cootes, T., Taylor, C., Cooper, D., Graham, J.: Active Shape Models - Their Training and Application. Comp. Vis. and Im. Und. 61(1), 38–59 (1995)
12. Dalal, N., Triggs, B.: Histogram of Oriented Gradients for Human Detection. In: CVPR, pp. 886–893 (2005)
13. Vedaldi, A., Fulkerson, B.: VLFeat: An Open and Portable Library of Computer Vision Algorithms, `http://www.vlfeat.org/`
14. Haralick, R., Shanmugam, K., Dinstein, I.: Textural Features for Image Classification. IEEE Trans. Sys. Man and Cyb. 6, 610–621 (1973)
15. Lindeberg, T.: Scale-Space Theory in Computer Vision. Kluwer Academic Publishers (1994)
16. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Bruncher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. J. of Mach. Learn. Res. 12, 2825–2830 (2011)
17. Breiman, L., Friedman, J., Stone, C., Olshen, R.: Classification and Regression Trees. Chapman and Hall/CRC (1984)