Automatic Detection and Segmentation of Focal Liver Lesions in Contrast Enhanced CT Images

Arne Militzer[1,2], Tobias Hager[1], Florian Jäger[1],
Christian Tietjen[2], Joachim Hornegger[1]
[1] *Pattern Recognition Lab, Department of Computer Science , Friedrich-Alexander-University Erlangen-Nuremberg, Martensstr. 3, 91058 Erlangen, Germany*
[2] *Siemens AG, Healthcare Sector, Siemensstr. 1, 91301 Forchheim*
arne.militzer@informatik.uni-erlangen.de

***Abstract*—In this paper a novel system for automatic detection and segmentation of focal liver lesions in CT images is presented. It utilizes a probabilistic boosting tree to classify points in the liver as either lesion or parenchyma, thus providing both detection and segmentation of the lesions at the same time and fully automatically. To make the segmentation more robust, an iterative classification scheme is integrated, that incorporates knowledge gained from earlier iterations into later decisions. Finally, a comprehensive evaluation of both the segmentation and the detection performance for the most common hypodense lesions is given. Detection rates of 77% could be achieved with a sensitivity of 0.95 and a specificity of 0.93 for lesion segmentation at the same settings.**

***Keywords*-Biomedical image processing, image segmentation, pattern classification, object detection, tumors**

I. Introduction

Liver cancer is among the most frequent types of cancerous diseases, showing responsible for the deaths of 600,000 patients worldwide in 2001 alone [1]. The incidence of liver metastases is even higher, as many common cancer types, like colorectal, lung and breast cancer, tend to metastasize into the liver. Computed tomography (CT) images, acquired after intravenous injection of a contrast agent, are widely used by clinicians for diagnosis, treatment planning and monitoring of liver tumors. These procedures, however, require information about size, shape and precise location of the tumors, for which their segmentation is a prerequisite. As the liver can stretch across well over 150 slices in a CT image and contain up to dozens of lesions, manual segmentation is tedious and prohibitively time consuming for a clinical setting. Automatic segmentation on the other hand, is a very challenging task. Despite the different contrast enhancement behavior of liver tumors and parenchyma, the image contrast between these tissues can still be low due to individual differences in perfusion and scan timing. Moreover, lesion shape, texture, and size vary considerably from patient to patient.

Proposed automatic algorithms involved, e.g., combinations of adaptive multi thresholding and morphological operators [2] or k-means clustering on mean shift filtered images [3]. However, these histogram based methods require a good contrast between lesions and parenchyma. More flexible machine learning techniques, such as AdaBoost, have been used in semi-automatic approaches to locate lesion boundaries by classifying 1-D histograms and dynamic programming [4], as well as in automatic settings to classify image textures [5]. The approach most closely related to ours was proposed by Shimizu et al. [6]. They trained two AdaBoost classifiers with a set of grey value statistical and gradient features calculated on normalized images, as well as features based on a convergence index filter, that enhances blob-like structures. One classifier was trained for segmenting large, the other for segmenting small lesions. After applying both classifiers to an image, their results were merged to a final output.

II. Methods

Object detection and segmentation are usually considered two separate tasks, where in practice often the user has to do the detection and provide some kind of initialization to a semi-automatic segmentation algorithm. On the other hand, approaches like the one by Shimizu et al. [6] or the one proposed here perform both tasks in a single step by deciding for each point in the image, whether it is part of a lesion or not. The only input that has to be provided to our system is a CT image of the liver with venous contrast enhancement. This phase shows maximum enhancement of the liver parenchyma and thus guarantees ideal conditions for lesion detection, as most liver lesions exhibit a lower uptake of contrast agent than the parenchyma.

The presented algorithm comprises four steps: First, the liver is automatically segmented. Second, the input data is normalized to compensate for variations in contrast enhancement. Next, the classification system assigns a value to each point in the liver, representing its probability of belonging to a lesion. Last, during post processing lesion candidates are generated from this probability map and returned as final detections of the system. Although we focus on mainly hypodense lesions here, the system is in principle also suitable for segmenting more complex hyperdense lesions.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.618

A. Liver segmentation

Automatic segmentation of the liver not only constrains the search space to relevant areas, saving computation time. It also reduces the complexity of the feature space making the classification task more feasible and reducing the risk of spurious detections. The method adopted here is based on [7]. Ling et al. model the liver by a hierarchical mesh-based shape representation. First, the liver is detected estimating its pose and location on the coarsest level using the marginal space learning scheme. Then, the model is refined applying a learning-based boundary localization which helps the system to become reliable to heterogeneous intensity patterns. The liver surface is decomposed into patches depending on the surrounding anatomic structures, and patch dependent classifiers are employed to cope with the different texture patterns.

B. Intensity standardization

In order to normalize the intensities and thus make the images more comparable, a non-rigid matching of the histogram of each target image to the histogram of one reference data set previously chosen from the database is conducted as proposed in [8]. Subsequently, the estimated mapping is applied to the intensities of the target image.

C. Voxel classification

In the main step of the proposed segmentation method, each voxel within the liver is classified as either lesion or parenchyma by a previously trained classifier. The highly variable appearance of parenchyma and lesions makes it difficult for a single classifier like AdaBoost or a support vector machine to globally find an appropriate model for each target class and thus an optimal decision boundary in the input space. While Shimizu et al. used two AdaBoost classifiers to be able to account for at least different lesion sizes, we chose to adopt the recently proposed probabilistic boosting tree (PBT) [9] in combination with AdaBoost and decision stumps. Due to its hierarchical nature, the PBT is able to capture the full variability within the classes. In the training stage it recursively learns a tree, where at each node a strong classifier is trained, e.g., with AdaBoost. This strong classifier is then used to split the training set into a negative and a positive subset, which form the input for training the left and right subtrees, respectively. Training samples, for which the strong classifier generates an output close to the decision boundary, however, represent hard examples and are put into both subsets. This procedure recursively subdivides the feature space, generating more detailed decision problems for nodes deeper in the tree, similar to a decision tree. Moreover, it was shown in [10] that an AdaBoost classifier H approaches logistic regression for a pattern $\mathbf{x}$ and posterior probabilities $p(y|\mathbf{x})$, $y \in \{-1, 1\}$ by

$$H(\mathbf{x}) \approx \frac{1}{2} ln \frac{p(y=+1|\mathbf{x})}{p(y=-1|\mathbf{x})}, \tag{1}$$

allowing the computation of approximate posterior probabilities for $\mathbf{x}$ as

$$q(\pm 1|\mathbf{x}) = \frac{exp(\pm 2H(\mathbf{x}))}{1 + exp(\pm 2H(\mathbf{x}))}. \tag{2}$$

To classify a new pattern with a trained PBT, the root node's strong classifier calculates its posteriors $q(\pm 1|\mathbf{x})$. Depending on the result the pattern is passed into the subtrees and the procedure is repeated recursively all the way down the tree. Each node then combines the results from its subtrees and returns them to its parent node. That way the PBT combines the classification results of its internal nodes into the overall approximate posterior distribution $\tilde{p}(y|\mathbf{x})$ at the root. This is a true probability value and can be used to trade off the tree's sensitivity vs. its specificity via a single threshold.

The features we used for classification can be grouped into three sets. The first contains grey value statistical features, like min, max, mean, and median intensities, contrast, range, variance, skewness over 3-D neighborhoods of various sizes, as well as gradients in 2-D and 3-D. The second group comprises 3-D Haar-like features in various scalings. In contrast to other approaches [4], [6], the neighborhoods are all computed not on a voxel but on a millimeter scale, making the approach robust against the use of images acquired with different CT scanners and acquisition protocols.

One drawback of the voxel classification approach is the fact, that it treats the classification of each point as an independent problem, which is obviously not true when segmenting contiguous objects. Usually, one tries to compensate for this wrong assumption by incorporating context information, e.g., by designing special features or averaging features over some neighborhood. An entirely different approach is proposed by Morra et al. in [11]. They directly use the fact that neighboring points with similar properties tend to belong to the same class. Similar to their approach, we train a cascade of PBTs, each of which receives not only the described image features as input, but also features calculated from the output probability image of the preceding classifier (Fig. 1). That way, when classifying a point, previously gained knowledge of the classes of surrounding points can function as a prior. The features we calculate from this probability image for each point are the point's own probability, the mean, median, Gaussian-weighted sums in 2-D and 3-D in some neighborhood, as well as the sum of the surrounding points without the point itself.

D. Post Processing

The final probability map generated by the classification step is smoothed by means of a median filter. Then, a morphological opening operation with a kernel size of

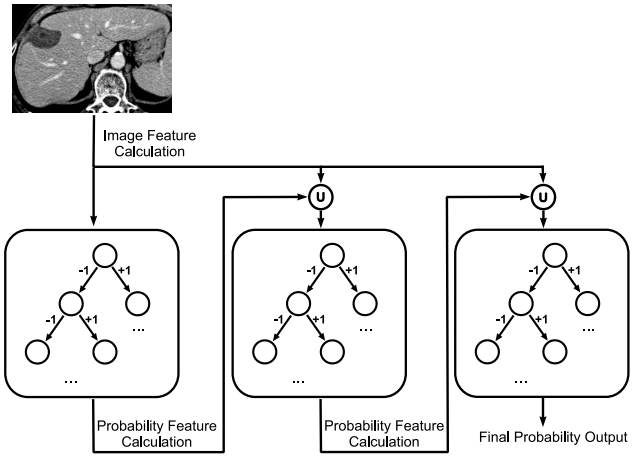


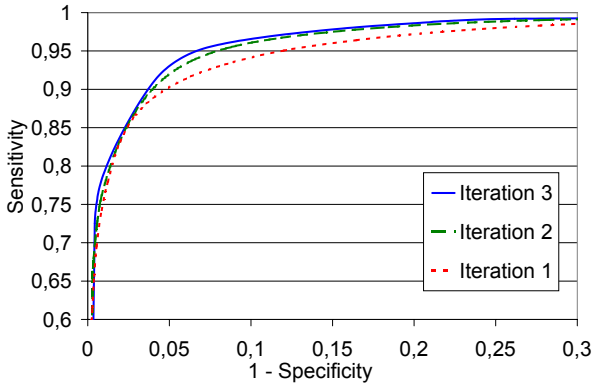
Figure 1. In the iterative classification, output probabilities of each step function as additional features for the next one.

$3 \times 3 \times 3$ is performed to eliminate small and isolated false positive detections. Finally, the image is converted into a lesion candidate mask by thresholding the probability values.

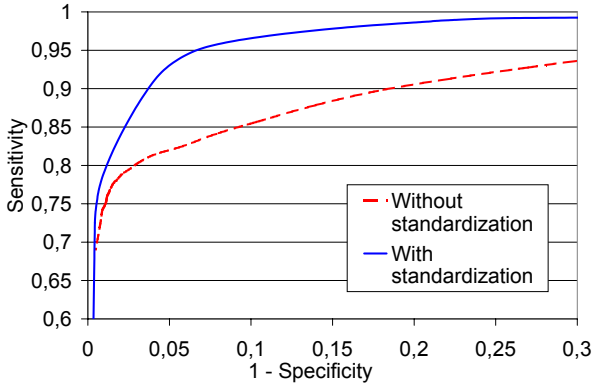
III. Results and Discussion

The system was evaluated using a total of 15 CT liver datasets with venous phase contrast enhancement from three different clinical sites. The voxel resolution varied from 0.547 to 0.832 mm in x and y directions, and 1 to 3 mm in z direction. For training and testing, datasets were sub sampled to a slice thickness of 3 mm where applicable. The contained lesions were all hypodense, i.e. they appeared darker than the parenchyma in the images. Apart from some (benign) cysts, most of the lesions were malignant ones such as hepatocellular carcinoma or metastases of various sizes, which form the most common malignant focal liver lesions in the clinic.

For each reported experiment a 5-fold cross-validation was performed, using 12 datasets for training and the remaining three for testing. For segmentation quality assessment, receiver operator characteristics (ROC) curves were calculated by varying a threshold on the probability map generated by the classifier. Thus, the curves represent the classification performance, with no post processing. Fig. 2(a) shows the effect of the iteration scheme. The calculated probability features were extensively used by the classifiers, such that the second and third classification stage both outperformed the first one, while the performance gain in the first step was bigger than in the second. The improvement achieved by training more than three iterations turned out to be marginal. From Fig. 2(b) it is clearly visible how crucial the proposed intensity standardization is. For these curves, the outputs of two classifiers at the last stage of the iterative scheme were compared, one of them trained and applied with standardization, the other without. The standardization



(a)



(b)

Figure 2. Effect of the iterative scheme (a) and the intensity standardization (b) on the classifier's performance.

increased the system's robustness and thus improved the classification. At a false positive rate (1-specificity) of 0.05, e.g., the sensitivity improved from 0.82 to 0.92.

To allow a fair judgement of the system, lesion detection has to be assessed separately. Unfortunately, this is rarely done in literature. Massoptier and Casciaro [3] do report figures on lesion detection performance, however, they do not clearly state their criteria for considering a lesion correctly detected or not, which makes a comparison difficult. Here, we consider a lesion candidate a false positive if it is covered by the reference lesion mask by less than 50%. A reference lesion r is considered detected (true positive) if (1) its centre of gravity lies inside a candidate and that candidate's centre of gravity lies inside r, or if (2) r is covered by lesion candidates by more than 50% (taking into account only candidates that were not marked false positive). The entire

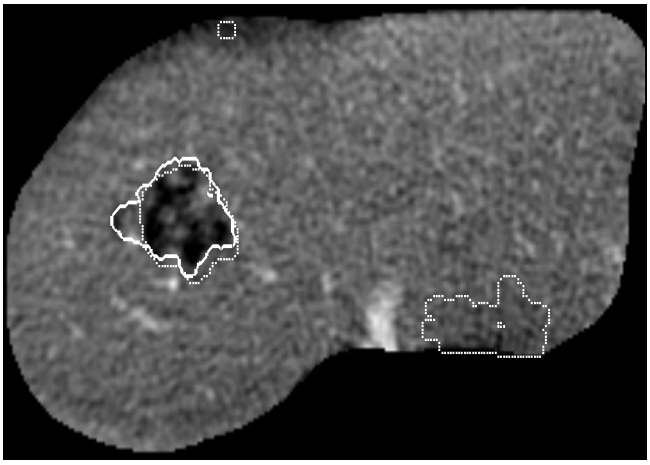


Figure 3. Segmentation result (dotted contours) with ground truth (solid).

analysis is restricted to objects with a volume larger than 0.125 ml. Applying these criteria to the final output of the system at threshold 0.44, we achieved a detection rate of 71% at a precision of 0.17, meaning that there were 4.8 false positive detections for each true positive or 14.4 false positives per patient. While this false positive rate appears very high, closer examination shows, that most of these false positives were located at the liver boundary or in fissures and exhibit characteristic shapes, and can thus be filtered using another classifier. The segmentation result at this threshold on the other hand was 0.95 sensitivity and 0.93 specificity. Figure 3 shows a typical detection result of the system.

IV. Conclusion and Outlook

The presented system for detection and segmentation of focal liver lesions is able to reliably segment the lesions in the used patient database. Successive training of several classifiers using additional probability features proved useful, as did the proposed standardization method. For satisfying lesion detection, however, false positive rates have to be further reduced. An integration of multiple contrast phases into the classification process might also be helpful when adapting the system for the segmentation of hyperdense lesions.

References

[1] "The World Health Report," *World Health Organization*, p. 188, 2002.

[2] M. Bilello, S. B. Göktürk, T. Desser, S. Napel, R. B. Jeffrey, and C. F. Beaulieu, "Automatic detection and classification of hypodense hepatic lesions on contrast–enhanced venous–phase CT," *Medical Physics*, vol. 31, no. 9, pp. 2584–2593, 2004.

[3] L. Massoptier and S. Casciaro, "A new fully automatic and robust algorithm for fast segmentation of liver tissue and tumors from CT scans," *European Radiology*, vol. 18, no. 8, pp. 1658–1665, 2008.

[4] Y. Li, S. Hara, and K. Shimura, "A machine learning approach for locating boundaries of liver tumors in CT images," in *Proc. ICPR*, vol. 1, 2006, pp. 400–403.

[5] D. Pescia, N. Paragios, and S. Chemouny, "Automatic detection of liver tumors," in *Proc. ISBI*, 2008, pp. 672–675.

[6] A. Shimizu, T. Narihira, D. Furukawa, H. Kobatake, S. Nawano, and K. Shinozaki, "Ensemble segmentation using AdaBoost with application to liver lesion extraction from a CT volume," *MIDAS Journal: Grand Challenge Liver Tumor Segmentation (MICCAI Workshop)*, 2008.

[7] H. Ling, S. Zhou, Y. Zheng, B. Georgescu, M. Sühling, and D. Comaniciu, "Hierarchical, learning-based automatic liver segmentation," in *Proc. CVPR*, 2008, pp. 1–8.

[8] F. Jäger and J. Hornegger, "Nonrigid registration of joint histograms for intensity standardization in magnetic resonance imaging," *IEEE Trans. Medical Imaging*, vol. 28, no. 1, pp. 137–150, Jan. 2009.

[9] Z. Tu, "Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering," in *Proc. ICCV*, vol. 2, 2005, pp. 1589–1596.

[10] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *Annals of Statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[11] J. H. Morra, Z. Tu, L. G. Apostolova, A. E. Green, A. W. Toga, and P. M. Thompson, "Automatic subcortical segmentation using a contextual model," in *Proc. MICCAI*. Springer-Verlag, 2008, pp. 194–201.