SALIENCY-BASED ROTATION INVARIANT DESCRIPTOR FOR WRIST DETECTION IN WHOLE BODY CT IMAGES

Mingchen Gao[1]*, Yiqiang Zhan*[2]*, Gerardo Hermosillo*[2]*, Yoshihisa Shinagawa*[2]*,*
Dimitris Metaxas[1]*, Xiang Sean Zhou*[2]

[1]CBIM, Rutgers University, Piscataway, NJ 08550
[2]Siemens Medical Solution USA Inc., Marlven, PA 19355

ABSTRACT

In this paper, we propose a saliency-based rotation invariant descriptor and apply it to detect wrists in CT images. The descriptor is motivated by the observation that salient landmarks around wrists usually form a characteristic spatial configuration (Fig. 1). In our framework, a set of interest points are firstly computed via scale-space analysis. For each interest point, we compute a pyramid of scale-distance 2D histograms constructed with neighboring interest points. The descriptor represents the spatial configuration among neighboring interest points in a rotation-invariant fashion. A cascade set of random forests are trained to distinguish wrist from other anatomies using this descriptor. Our algorithm shows robust and accurate performance on 41 whole body CT scans with diverse context, orientations and articulation configurations.

Index Terms— Anatomical structure detection, wrist detection, interest point descriptor, rotation invariant

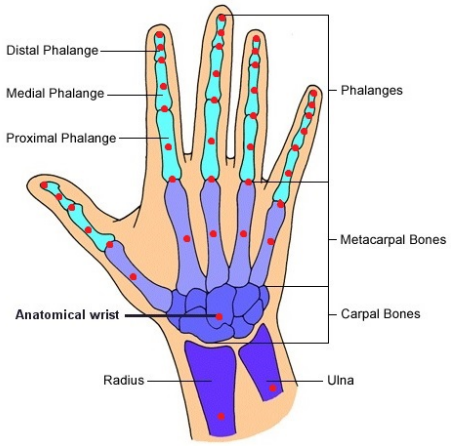

Fig. 1. Skeleton anatomy of the wrist. Red dots represent the detected interest points, which form a characteristic spatial configuration.

1. INTRODUCTION

Recent development of whole-body CT images with faster imaging speed has made it feasible in clinical workflow. It provides a comprehensive view of human anatomy. For example, in trauma cases, the radiologist can navigate through all bony structures of the body and evaluate the fracture risks. However, due to the large volume of whole-body scans, which is usually more than 500 slices and the arbitrary anatomical context of some structures, it is time consuming to manually navigate to the area wanted. Compared to other anatomies, wrist detection in whole-body CT is especially challenging, since 1) Wrists and hands can be with arbitrary positions and orientations, especially when patients are in coma or have their upper limbs injured. 2) Wrist and hands are often with highly diverse anatomical context, e.g., on the belly or beside the legs in some trauma cases; the two hands could be crossed over each other that end up with different and complex anatomical context. Fig. 2 shows some examples of the challenging cases of detecting wrist. An automatic wrist detection algorithm is thereby desired to help improving the radiological workflow. Efficiently detecting an anatomical structure (e.g., wrist, heart, liver, kidney) in medical images is an important component to build a fully automatic system. It leads to multiple applications such as semantic visual navigation, image retrieval, providing necessary initializations for the subsequent procedures, e.g., segmentation, measurements and classification. There are lots of previous work on the more general topic, object detection. It has been proved to be effective and robust in many 2D scenarios. The general approach often uses the haar feature to describe an image block and formulates it as a classification problem: whether an image block contains the target object or not [10]. However, extending this algorithm to anatomical structures detection in 3D images is not trivial due to the difficulty of effectively describing 3D features and the exponentially increased searching space with respect to the dimension of the space. Previous approaches working on 3D anatomical structures detection mainly are: classification based approaches [11, 12], regression based approaches [3, 8]. A summary survey can be found in [5].

The existing methods are not the best to be applied to

978-1-4673-1961-4/14/$31.00 ©2014 IEEE

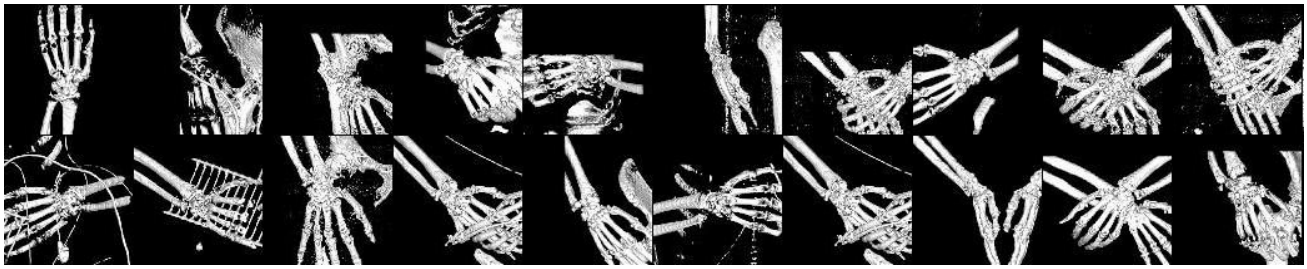

Fig. 2. Challenging cases for wrist detection in whole body CT images. Wrists are with various positions, orientations, poses and anatomical context.

this case directly. We observed that a wrist has its own characteristic, i.e., a unique spatial configuration of salient bony landmarks. A healthy wrist always consists of eight bones forming the carpal bones, with the radius and the ulna on the arm, and metacarpal bones on the hand, as illustrated in Fig. 1. Based on these observations, we propose an saliency-based rotation invariant descriptor of the wrist bones and detect the wrist based on it. Our detection algorithm starts from rotation invariant interest point detection and description. Specifically speaking, we use local extremes of difference of Gaussian (DoG) to extract interest points in images. For each interest point, a descriptor based on a pyramid of scale-distance 2D histograms is then constructed. It describes the spatial relations between neighboring interest points in a rotation-invariant fashion. It is worth noting that our descriptor is completely different from those well known rotation-invariant features, e.g., SIFT, SURF. Instead of describing local appearance features, our descriptor aims to provide a multi-scale view of spatial configurations among neighboring interest points. Both steps are guaranteed to be rotation invariant to handle the arbitrary rotated anatomical structures. This framework generates a list of "in-subject-distinctive" and "cross-subject-reproducible" interest points. The detection problem is then converted from detecting the anatomical structures in images directly to interest point classification problem, whether it is the interest corresponding to the anatomy we are looking for. Finally, a cascade of random forests [2] are learned to classify interest points to detect the point that we are interested in.

2. METHODOLOGY

Our detection system consists of three main steps: 1) Interest point detection. 2) Construction of saliency-based rotation invariant descriptors. 3) Classification on interest points. The first two steps aim to extract a set of wrist candidates. These candidates should be distinctive in an image and repeatable through different subjects as well [9], i.e., we can easily distinguish an interest point from others in one subject and the feature for the same anatomical interest point should be consistent through different subjects. In the third step, a cascade set of random forests are applied to each candidate to determine whether it is the wrist. Compared to voxel-wise classification, the classification is only applied on candidates that form a much smaller hypothesis space.

2.1. Interest Point Detection

Interest points are selected at distinctive locations in an image, such as, corners, blobs, edges. In our study, the most valuable property of an interest point detector is its repeatability. Reliability of a detector depends on its repeatability to find the same physical interest points under transformations and different viewing conditions. From the previous study on existing detectors [1, 7, 9], we choose the local extreme of DoG as the most appropriate interest point detector for our problem. It is well known that DoG is an approximation of Laplacian of Gaussian (LoG), which is the a scale invariant interest point detector. DoG is first proposed by Lindeberg [6], and has been used in lots of successful interest point detectors [7]. The scale invariant detectors and descriptions are mainly used for searching stable features across all possible scales. In our anatomical structure detection problem, scale of each interest point has anatomical meaning, which denotes the size of structure detected. We use the scale information directly as part of the description.

The process of detecting interest point is efficiently implemented as the following: Convolve the initial image with different scale of Gaussian to produce smoothed images in scale space.

$$L(x, y, z, \sigma) = G(x, y, z, \sigma) * I(x, y, z). \quad (1)$$

Adjacent image scales are subtracted to get the difference of Gaussian images.

$$D(x, y, z, \sigma) = L(x, y, z, \sigma') - L(x, y, z, \gamma). \quad (2)$$

To detect the local maxima and minima of $D(x, y, z, \sigma)$, each point is compared with its neighbors in both spatial and scale space. In our three dimensional space, it will be compared with 26 neighbors in spatial space and 27 neighbors in the scale spaces above and below, respectively. It is selected only if it is larger or smaller than all of the neighbors. The interest point detected by this process can be used as anatomically meaningful landmarks.

2.2. Saliency-based Spatial Configuration Descriptor

The second step is to generate rotation invariant description for interest points. There are a large group of descriptors proposed in the literature. SIFT provides a rotation invariant local description of the blob [7]. SURF describes a distribution of Haar-wavelet responses within the interest point neighborhood [1]. Existing features mainly are descriptions of local appearance.

In our application, we observed that, after the interest point detection, wrist, hand and arm compose a specific and consistent configuration. Motivated by this observation, a histogram based feature is proposed to describe the relationship between neighboring interest points. An interest point can be represented by a two-dimensional histogram using neighboring points around it. One dimension is the distance to neighboring interest points, and the other dimension is the scale of neighboring interest points. This is an effective and simple description of the neighboring configuration of an interest point. This description is obviously rotation invariant.

We further improve the feature description by designing a histogram pyramid for every interest point. Pyramid match kernel was proposed by Grauman. *et al*. [4]. The feature extraction function $\Phi(p)$ for an interest point p is defined as

$$\Phi(p) = [s(p), H_0(x), H_1(x), ..., H_L(x)], \qquad (3)$$

where $H_i(x)$ is a histogram vector formed over the neighboring interest points x using two-dimensional bins, whose size is 2^i, $L+1$ denotes the level of the pyramid histogram. In other words, histogram pyramid is a vector of concatenated histograms, where each subsequent component histogram has bins that double in size compared to the previous one. The scale $s(p)$ of the reference interest point is a characteristic of the size of structure detected. We put the scale directly in front of the histogram pyramid .

2.3. Classification on Interest Points

The anatomical structure detection problem is converted to a point wise classification problem. The search domain, which was the whole image, now has been significantly decreased to the interest point domain. Given the feature description and a training set of positive and negative interest points, any machine learning approach could be used to learn a classifier. In out system, we choose random forest because of its speed and resistance to overfitting.

For our application, the positive and negative samples are highly unbalanced. In general, training a classifier using highly unbalanced data yields higher false positive rate to lower error rate. Viola. *et al*. proposed a cascade strategy which will train a series of classifiers to handle unbalanced training data [10]. At each step, the threshold of each random forest classifier is adjusted so that the false negative is close to zero. It rejects many negative samples while leaving almost all positive instances. The samples pass a classifier will be used to train subsequent classifiers. During the testing stage, a sample need to pass all the classifiers to be categorized as positive. The cascade strategy increases detection accuracy while reducing computation time.

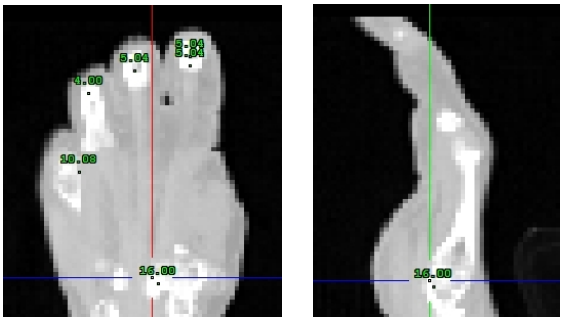

Fig. 3. Interest point detected around wrists. The scale of each interest point was marked as the number near it. The interest point on the wrist has a scale $16.0mm$.

3. EXPERIMENTS AND DISCUSSION

41 whole body CT images were collected to validate the proposed detection algorithm, each image includes one or two hands. 48 positive interest points were detected on the wrist. Among them, 38 interest points were randomly selected for training and the remaining 10 were reserved for testing. All the images were manually labeled with ground truth interest points after the interest points detection step.

Interest Point Detection: Fig. 3 shows the detected interest points. Most of the detected interest points have anatomical meanings. The set of carpal bones on the wrist were consistently detected with a relatively larger scale $12.7mm$ or $16.0mm$. The finger joints were usually detected at scale $4.0mm$ or $5.04mm$, and the middle of finger bones were detected at scale less than $4.0mm$. The wrist bones were the most consistently detected compared to other anatomically meaningful interest points. Our algorithm relies on the robustness of interest point detection. In our experiments with the whole body CT image set, we only missed wrist in one image. The interest point detection provided a stable start for further description and classification.

Saliency-based Spatial Configuration Descriptor: The interest point description was built using both one layer of the scale-distance histogram and and histogram pyramid. We compared their performances in Fig. 5. 12 bins of scale and 12 bins of distance were used to build the lowest layer of the scale-distance histogram. For the histogram pyramid, each higher level doubled the size of bins. We used three layers of histogram to build the histogram pyramid. We compared three settings of the proposed method in Fig. 5. In the non-cascaded experiment settings, we randomly selected a smaller

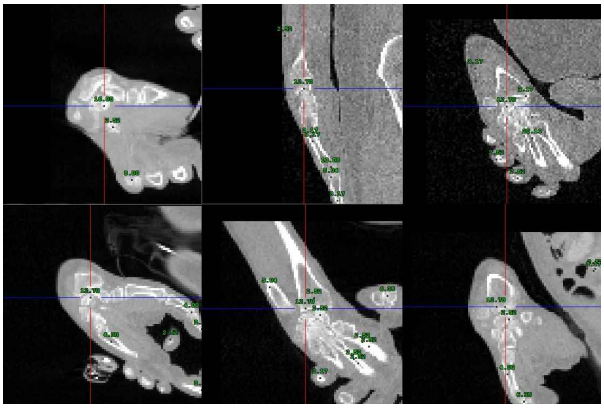

Fig. 4. Interest point detected and classified as positive.

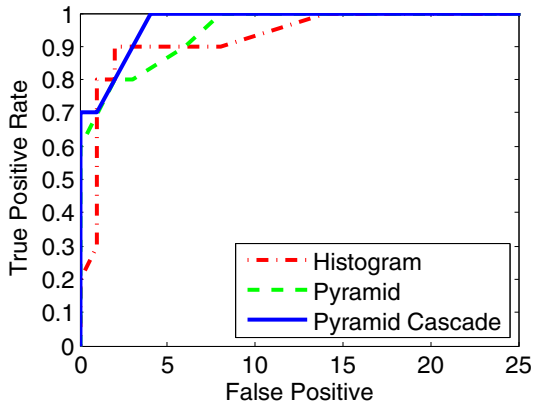

Fig. 5. ROC curve for our wrist detector. The red, green and blue curves represent the performance using one scale-distance histogram without cascade mode, histogram pyramid feature without cascade mode and the histogram pyramid feature with cascade mode, respectively.

set of negative interest points for training. Pyramid histogram with cascaded mode performed the best. The experiment was performed on a 2.40GHz Intel Core2 Quad computer with 8G RAM. The most time consuming part was the interest point detection. Typically it took about 10 seconds to detection all the interest points in a whole-body CT image. After that, the positive interest point can be classified in real time.

4. CONCLUSION

The wrist detection problem in whole-body CT images has been cast here as anatomically meaningful landmark detection, descriptor and classification problems. The core idea is to design a saliency based rotation invariant descriptor to distinguish anatomical landmarks. Our algorithm has two major advantages. First, both the interest point detection and descriptor are "in-subject-distinctive" and "cross-subject-reproducible". Therefore, the wrist detection system is able to achieve accurate detection results. Second, compared to voxel-wise classification, our classification is only applied on interest points that form a much smaller searching space, such to facilitate the detection speed. The framework can be extended in two directions, detection of other anatomical structures, and more feature-based applications, e.g., align arbitrarily rotated anatomies in medical images, content based medical image retrieval, etc.

References

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer vision and image understanding*, 110(3):346–359, 2008.

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

[3] A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu. Regression forests for efficient anatomy detection and localization in CT studies. In *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, pages 106–117. Springer, 2011.

[4] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV 2005.*, volume 2, pages 1458–1465. IEEE, 2005.

[5] S. Kevin Zhou. Discriminative anatomy detection: Classification vs regression. *Pattern Recognition Letters*, 2013.

[6] T. Lindeberg. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics*, 21(1-2):225–270, 1994.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[8] O. Pauly, B. Glocker, A. Criminisi, D. Mateus, A. M. Möller, S. Nekolla, and N. Navab. Fast multiple organ detection and localization in whole-body MR Dixon sequences. In *MICCAI 2011*, pages 239–247. Springer, 2011.

[9] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(3):177–280, 2008.

[10] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR 2001.*, volume 1, pages I–511. IEEE, 2001.

[11] Y. Zhan, X. S. Zhou, Z. Peng, and A. Krishnan. Active scheduling of organ detection and segmentation in whole-body medical images. In *MICCAI 2008*, pages 313–321. Springer, 2008.

[12] Y. Zheng, B. Georgescu, and D. Comaniciu. Marginal space learning for efficient detection of 2D/3D anatomical structures in medical images. In *IPMI*, pages 411–422. Springer, 2009.