AUTOMATIC VIEW CLASSIFICATION FOR CARDIAC MRI

Yan Zhou[⋆] *Zhigang Peng*[†] *Xiang Sean Zhou* [†]

[⋆] Elekta Inc., Maryland Heights, MO, USA
[†] Siemens Healthcare, Malvern, PA, USA

ABSTRACT

View classification for cardiac MR images is a new topic in medical image analysis, and can support efficient content-based filtering, browsing, and retrieval. The major difficulty lies in large variability in image appearance caused by various acquisition protocols, heart phases and disease conditions. We propose a collaborative learning approach that exploits statistical dependencies at three levels: the local image patch level, the parts-whole level, and the anatomical level. Specifically, at the local image patch level, we make redundant use of the training images in a spatial manner; At the parts-whole level, we model the relationship among the detected landmarks by a sparse configuration method to remove erroneously detected landmarks; At the anatomical level, we train a view classifier based on the DICOM header as another information source, which is then optimally incorporated in a Bayesian way as a prior. We compare the approach with various integrations of state-of-art methods. Large-scale experiments on real-world clinical datasets—over ten thousand *unseen* images, many with severe diseases—show that the approach is highly robust, outperforming other integrated approaches, and can achieve a 97.6% classification accuracy.

Index Terms— Collaborative learning, image retrieval

1. INTRODUCTION

Cardiac MR images can be taken from many different orientations. Recognition of such orientations or "cardiac views" is a prerequisite for computer-aided diagnosis and for computer-assisted workflow automation based on image contents [7][6]. The goal of this study is to automatically classify cardiac MR images with respect to what view of the heart they contain. There are three typical MRI cardiac views: vertical long axis view (VLA or two chamber view), horizontal long axis view (HLA or four chamber view), and short axis views (SA) (Figure 1). The algorithm must demonstrate high accuracy and robustness when tested on large clinical datasets (e.g., we did testing on over ten thousand MR images).

Even for the same cardiac view, images have large variability in image intensity, scale and rotation; as well as a wide variety of disease conditions such as cardiac dilation or hypertrophy, lung diseases, or excessive fat, etc. Figure 1 shows some examples from a typical clinical dataset, (a)-(b) belong to short axis view (SA) with different scales, (f)-(i) are HLA views with various rotation angles. Figure 1 (j)-(k) are normal and flipped VLA views. Some of these images are difficult to recognize even for (non-expert) human users.

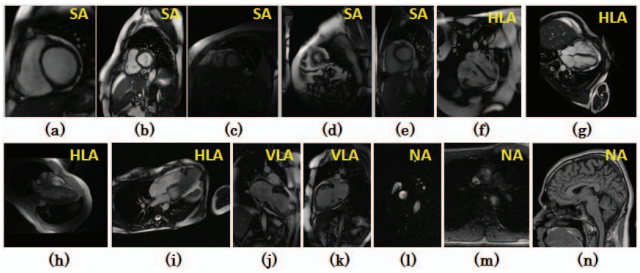


Fig. 1. Image examples from a typical clinical database. SA: short axis view; VLA: vertical long axis view; HLA: horizontal long axis view; NA: Others.

Digital Imaging and Communication in Medicine (DICOM) header includes non-image data for a standardized description of study, patient, or technical parameters. Two DICOM fields may provide cardiac view information. The first is 'SeriesDescription' field, which may contain substrings "2cv" or "2chamber" indicating a two-chamber view, "4cv" or "4chamber" indicating a four-chamber view. However, such a coding scheme, if not missing completely, may be in different languages, and are usually hospital, scanner, and/or user dependent; hence unreliable. The second DICOM field is "OrientationPatient" which encodes the orientation of the MRI slice plane in the patient coordinate. This information can serve as a good pre-filter (i.e., provide a good prior), although diseases and abnormalities abound, and some views are more easily confused with others. Furthermore, non-cardiac images may have these orientations as well. For example, Figure 1 (l-n) are not cardiac, but (m) has the same orientation as (h).

1.1. Previous Work

To the best of our knowledge, this is the first work addressing the cardiac view classification problem in MR. For cardiac ultrasound, one can find some related work published recently [3][11][1]. Aschkenasy et al. [1] presented a classification algorithm based on multi-scale elastic registration. Ebadollahi et al. [3] proposed a part-based approach with each view

This material is based upon work when Yan Zhou was an intern in Siemens Healthcare, Malvern, PA, USA.

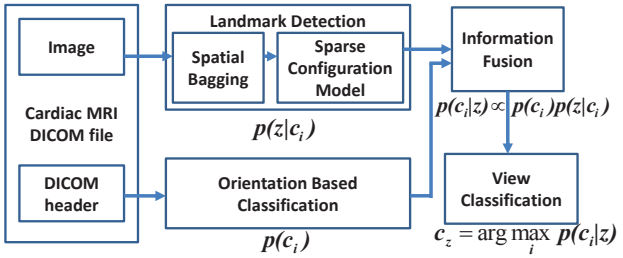


Fig. 2. Algorithm framework.

represented by a constellation of detected heart chambers, the connections of which are modeled by MRF. The accuracy is 88.35%. Zhou et al. [11] proposed a LogitBoost approach for multi-class classification based on Haar-like features. Otey et al.[8] proposed hierarchical Logistic Model Trees classifiers on bags of simple features, and achieves 92.7% accuracy. Park et al. [9] presented an extension of [11] improved the accuracy to 96%.

1.2. Proposed Approach

We propose a collaborative learning approach to solve this complex problem, in which multiple independently trained learners collaborate with one another to form a mutually enhanced and more robust consensus decision. The collaborations exist at three levels: the local image patch level, the parts-whole level, and the anatomical level. At the local image patch level, we make redundant use of the training images in a spatial manner to produce robust landmark detection results. Unlike existing bagging predictors[2] which resample with replacement, we reuse all the training images in each bag, but each time re-align to a different part of the target structure. At the parts-whole level, we model the relationship among the detected landmarks by a sparse configuration method, in which every pair or triple of landmarks form a voting group and predict among each other. This procedure is able to detect outliers and remove the falsely detected landmarks. At the anatomical level, we extract the orientation information from the DICOM header, and use it as prior knowledge. Additionally, orientation information from DICOM headers is used to identify flipped images and normalize the rotated images, which makes the learning tasks easier. The learned visual information and DICOM header information is fused by Bayesian theorem. This collaborative learning approach, as we will show later, can achieve a very high accuracy (94.6% to 99.9% for different views) in large-scale testing.

2. METHODOLOGY

In this section, we introduce our framework of collaborative learning at three levels: image patch level, parts-whole level and anatomical level. We begin with an overview of the system. Then we discuss details of the three levels.

2.1. Overview of the Algorithm

The algorithm flow is shown in Figure 2. The three cardiac views are defined as three classes $C = \{c_1, c_2, c_3\}$. Given a DICOM file, we utilize both visual information z and DICOM header information. From the image, we compute the $p(z|c_i)$ by a landmark detection algorithm. From the DICOM header, we compute the probability $p(c_i)$. It's also used to identify flipped images, and normalize rotated images. Finally the posterior probability $p(c_i|z)$ is computed, and the class label is assigned by MAP method (Equation 1).

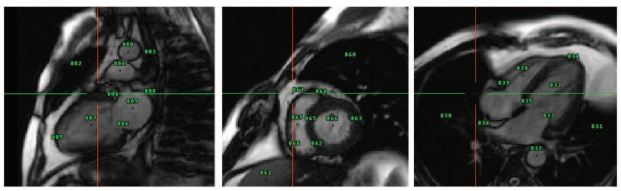


Fig. 3. Landmark training examples. For each cardiac view, 10 anatomical locations are selected and annotated by radiologists (in green).

$$c_z = \arg\max_i p(c_i|z), i = 1, 2, 3 \quad (1)$$

The landmark detection algorithm is motivated by [10], and modified to handle landmark in multi-scale context. Ten landmarks are pre-defined for each view (e.g., Figure 3) and manually annotated. Then we use a spatial bagging approach to collect the training set; this is discussed in more detail in Section 2.2. Adaboost[4] algorithm is then applied to train each landmark detector. Different detectors are trained independently across several image scales to produce a cascade of detectors.

In the testing phase, a cascade of detectors are used in a coarse-to-fine manner. The coarse scale landmark detector is first employed to slide over the image, generating detection scores at each location. Then candidate positions with responses higher than a threshold λ are obtained. The threshold λ is automatically computed in the training phase. Subsequently, the finer level detectors are employed at previously obtained candidate positions. When the process stops, a single winner with a sufficiently high response is retained. However, at this stage, there are some unavoidable erroneous detections. We use a sparse configuration model to remove erroneous detection, this is discussed in more detail in Section 2.3. Finally, a detection score is computed for each view as $p(z|c_i) = \frac{n_i}{N_i}$, where n_i is the number of eligible landmarks for the ith view, and N_i is the total number of defined landmarks. This score is used by Section 2.4.

2.2. Image Patch Level: Spatial Bagging

Existing bagging method makes redundant use of the training set by re-sampling with replacement. Given a training set $T = \{(y_n, x_n), n = 1, ...N\}$ where y_n is the class label and x_n is the input, and a learning algorithm that uses this set to form a classifier $f(x, T)$, a traditional bagging predictor [2] can be formulated in Equation 2, where T^B's are the bootstrap samples or bags, and avg_B denotes averaging or voting among the classifiers.

$$f_B(x) = avg_B f(x, T^B) \quad (2)$$

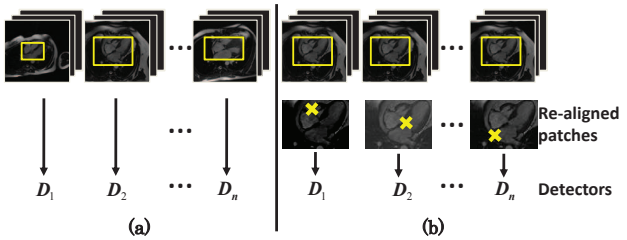


Fig. 4. (a) Existed bagging method; (b) Spatial bagging.

We also make redundant use of the training images, but in a spatial manner. To learn visual patterns, where x is an input image, the boostraping can be done *spatially*, with each bag using all the images in the training set, but re-aligned by different parts of the target pattern. Denoting the re-alignment process as $R(T)$, the formulation becomes Equation 3. Figure 4 compares the existing bagging method with the spatial bagging method. The spatial bagging method makes sense because it resembles the human's foveal vision system [5], which makes repeated examinations of the same scene with different focus points. In our application, the focus points are the annotated landmark points.

$$f_R(x) = avgf(x, R(T)) \tag{3}$$

2.3. Parts-whole Level: Sparse Configuration Model

This module is designed to remove false positive landmarks based on configurations. Denote the conditional probability of one landmark location l_i given other landmarks as $P(l_i|L_{\bar{i}})$. If all the elements in the set $L_{\bar{i}}$ are correct, this probability can be used to measure the quality of l_i. However, we don't know if any member in $L_{\bar{i}}$ is erroneous, so we adopted a voting strategy. The vote received by landmark point l_i is denoted by $v(l_i|L_v)$, where L_v is a voting group. The vote is defined as l_i's likelihood of being accepted by L_v based on the conditional distribution estimation of the annotated training set. Assuming the training set follows a Gaussian distribution with mean μ and covariance matrix $\sum$, the vote is computed in Equation 4, where $D \in \{2, 3\}$ is the dimensionality of the image.

$$v(l_i|L_v) = \frac{1}{(2\pi)^{D/2}|\sum|^{1/2}} e^{-(l_i-\mu_i)^T \sum^{-1}(l_i-\mu_i)} \tag{4}$$

2.4. Anatomical Level: Bayesian Fusion

According to the DICOM standard, the "Image Orientation" (0020,0037) field specifies the direction cosines of the first row and the first column with respect to the patient. Figure 5 (a) shows the orientation distribution of 180 DICOM files selected from different cardiac views, some of them are flipped cases (e.g.,Figure 1 (j) and (k)). Each point on the graph represents the dot product of the row value and column value. The clusters are well separated, thus provide valuable evidence for automatic flip correction and orientation normalization. We train the orientation for all the five clusters as five classes. For each testing case, a class label is assigned by k-nearest-neighbor or k-NN($k = 13$) method. The identified

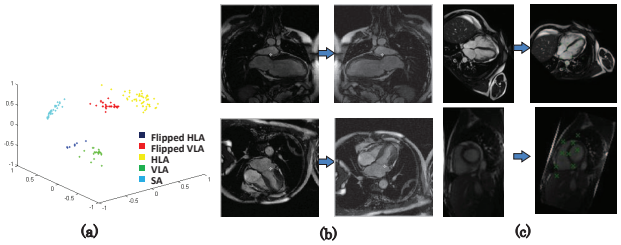


Fig. 5. (a) Orientation distribution of 180 DICOM files; (b) Corrected flipped cases; (c) Normalized rotated cases.

flipped images will be corrected before the landmark detection step (shown in Figure 5 (b)). Similarly, in-plane rotations are normalized as shown in Figure 5 (c).

The k-NN step described in the previous section produces a DICOM header based classification score computed as $p(c_i) = \frac{s_i}{k}$, s_i is the number of cases in class i among the nearest $k = 13$ neighbors. The posterior probability $p(c_i|z)$ is then obtained by $p(c_i|z) \propto p(c_i)p(z|c_i)$.

3. EXPERIMENTS

3.1. Experimental Settings

The DICOM files for the experiments are acquired by Siemens MAGNETOM scanners including Sonata 1.5T, Symphony 1.5T and Avanto 1.5T, which mainly contain CINE sequences for the diagnosis of Ischemic heart disease, Arrhythmogenic Right Ventricular Dysplasia, Thalassemia etc. We conduct extensive experiments to evaluate the performance of the approach from the following perspectives:

- *Evaluation of the landmark detection method*: The purpose of the evaluation is to test the approach with only visual information, when no DICOM header information is involved. The testing is performed on 'DB1' database consisting of 6258 standard cardiac view MR images from 52 patients. 'DB1' contains 1456 VLA, 2034 HLA and 2795 SA images, from which 144 VLA, 197 HLA and 273 SA images are randomly selected and annotated for training.
- *Comparisons of various integrations of methods*: The purpose of the evaluation is to compare various integrations of methods. We use 'DB2' database for the testing. It consists of 1571 unseen images from 57 patients including standard view, non-standard view, and non-cardiac MR images. No information is used in training.
- *Practical usability on a large scale unseen database*: 'DB3' consists of 10741 unseen images from 23 patients including standard view, non-standard view, non-cardiac MRI, and non-MR images (Figure 6). No information is involved in the training.

3.2. Experimental Results

Table 1 (a) shows the confusion matrix of 'DB1'. The landmark detection algorithm achieves 99.8% accuracy. This shows that when all the images from the dataset belong to

Table 1. (a) Confusion matrix for 'DB1'; (b) Classification accuracy for 'DB2' with various integrations of methods (see text); (c) Confusion matrix for 'DB3'.

(a)

	SA	VLA	HLA	Others
SA	2739	2	3	8
VLA	0	1443	0	0
HLA	0	0	2030	0

(b)

	O	L	LF	OL	OLR
SA	79.7%	85.1%	99.1%	100.0%	98.6%
VLA	97.8%	88.5%	90.6%	98.4%	99.2%
HLA	61.9%	68.9%	72.9%	92.3%	97.9%
Others		68.4%	80.8%	95.7%	96.9%

(c)

	SA	VLA	HLA	Others
SA	3444	0	0	2
VLA	0	735	0	19
HLA	0	0	4138	0
Others	1	0	237	2165

standard views, the landmark detection algorithm is able to achieve high accuracy based on only visual information. However, when the dataset includes ambiguous samples that do not belong to any of the standard views, visual information is not sufficient. 'DB2' is such kind of a dataset. Table 1 (b) shows the classification accuracy of various integrations of methods. We start the testing by using only landmark detection method (L) or using only orientation information from DICOM header (O). Both give limited performance. Then we incrementally integrate flip identification (LF) to correct flipped cases, the accuracy increases accordingly. By integrating Bayesian fusion (OL) and rotation identification (OLR) methods, the accuracy increases to over 96.9% for all the classes. Finally we test the approach with all the components involved in a large clinical dataset 'DB3' to validate real-world usability. This is the most challenging dataset in our testing because it contains all types of 'awkward' images including flipped/rotated images, ambiguous positioning / cut-planes, and various non-cardiac images. Some examples are shown in Figure 6. Table 1 (c) shows the confusion matrix. Our approach achieves 94.6% to 99.9% accuracy for different views, and an overall accuracy of 97.6%.

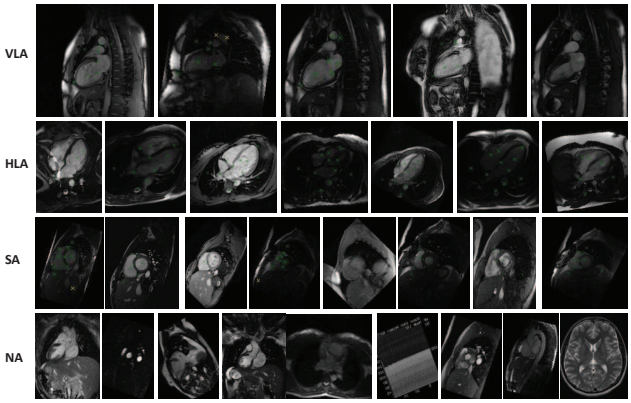


Fig. 6. Results for 'DB3'. Yellow marks are detected outliers.

4. REFERENCES

[1] S. V. Aschkenasy, C. Jansen, R. Osterwalder, A. Linka, M. Unser, S. Marsch, and P. Hunziker. Unsupervised image classification of medical ultrasound data by multiresolution elastic registration. *Ultrasound in Medicine and Biology*, 32(7):1047 – 1054, 2006.

[2] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[3] S. Ebadollahi, S.-F. Chang, and H. Wu. Automatic view recognition in echocardiogram videos using parts-based representation. *CVPR*, 2:2–9, 2004.

[4] Y. Freund and R. E. Schapire. A short introduction to boosting. *Japonese Society for Artificial Intelligence*, 14(5):771–780, 1999.

[5] H.Hunziker. The eye of the reader: Foveal and peripherial perception from letter recognition to the joy of reading. *Transmedia Zurich*, 2006.

[6] T. Lehmann, M. Guild, C. Thies, B. Fischer, K. Spitzer, D. Keysers, H. Ney, M. Kohnen, H. Schubert, and B. Wein. Content-based image retrieval in medical applications. *Methods of Information in Medicine*, pages 354–361, 2004.

[7] H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler. A review of content-based image retrieval systems in medical applications – clinical benefits and future directions. *Int'l Journal of Medical Infomatics*, 73(1):1–23, 2004.

[8] M. Otey, J. Bi, S. Krishna, B. Rao, J. S. A. S. Katz, J. Han, and S. Parthasarathy. Automatic view recognition for cardiac ultrasound images. *Int'l Workshop on Computer Vision for Intravascular and Intracardiac Imaging*, pages 187–194, 2006.

[9] J. H. Park, S. K. Zhou, C. Simopoulos, J. Otsuki, and D. Comaniciu. Automatic cardiac view classification of echocardiogram. *ICCV*, pages 1550–5499, 2007.

[10] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *CVPR*, 1:511 – 518, 2001.

[11] S. Zhou, J. Park, B. Georgescu, D. Comaniciu, C. Simopoulos, and J. Otsuki. Image-based multiclass boosting and echocardiographic view classification. *CVPR*, pages 1559 – 1565, 2006.