Learning-Based Quality Control for Cardiac MR Images

Giacomo Tarroni, Ozan Oktay, Wenjia Bai, Andreas Schuh, Hideaki Suzuki, Jonathan Passerat-Palmbach, Antonio de Marvao, Declan P. O'Regan, Stuart Cook, Ben Glocker, Paul M. Matthews, and Daniel Rueckert

***Abstract*— The effectiveness of a cardiovascular magnetic resonance (CMR) scan depends on the ability of the operator to correctly tune the acquisition parameters to the subject being scanned and on the potential occurrence of imaging artifacts, such as cardiac and respiratory motion. In the clinical practice, a quality control step is performed by visual assessment of the acquired images; however, this procedure is strongly operator-dependent, cumbersome, and sometimes incompatible with the time constraints in clinical settings and large-scale studies. We propose a fast, fully automated, and learning-based quality control pipeline for CMR images, specifically for short-axis image stacks. Our pipeline performs three important quality checks: 1) heart coverage estimation; 2) inter-slice motion detection; 3) image contrast estimation in the cardiac region. The pipeline uses a hybrid decision forest method—integrating both regression and structured classification models—to extract landmarks and probabilistic segmentation maps from both long- and short-axis images as a basis to perform the quality checks. The technique was tested on up to 3000 cases from the UK Biobank and on 100 cases from the UK Digital Heart Project and validated against manual annotations and visual inspections performed by expert interpreters. The results show the capability of the proposed pipeline to correctly detect incomplete or corrupted scans (e.g., on UK Biobank, sensitivity and specificity, respectively, 88% and 99% for heart coverage estimation and 85% and 95% for motion detection), allowing their exclusion from the analyzed dataset or the triggering of a new acquisition.**

***Index Terms*— Image quality assessment, magnetic resonance imaging, motion compensation and analysis, heart.**

Manuscript received September 15, 2018; revised October 17, 2018; accepted October 17, 2018. Date of publication November 1, 2018; date of current version May 1, 2019. This work was supported in part by the EPSRC Program under Grant EP/P001009/1, in part by the British Heart Foundation under Grant NH/17/1/32725, and in part by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Imperial College Healthcare NHS Trust. This work has been conducted using the U.K. Biobank Resource [1] under Application Number 18545. The work of G. Tarroni was supported by a Marie Skodowska-Curie Fellowship. *(Corresponding author: Giacomo Tarroni.)*

G. Tarroni, O. Oktay, W. Bai, A. Schuh, J. Passerat-Palmbach, B. Glocker, and D. Rueckert are with the Department of Computing, Imperial College London, London SW7 2AZ, U.K. (e-mail: giacomo.tarroni@gmail.com).

H. Suzuki is with the Division of Brain Sciences, Faculty of Medicine, Imperial College London, London SW7 2AZ, U.K.

A. de Marvao, D. P. O'Regan, and S. Cook are with the Faculty of Medicine, MRC London Institute of Medical Sciences, Imperial College London, London W12 0NN, U.K.

P. M. Matthews is with the Division of Brain Sciences, Faculty of Medicine, Imperial College London, London SW7 2AZ, U.K., and also with the UK Dementia Research Institute, London WC1E 6BT, U.K.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2018.2878509

I. Introduction

CARDIOVASCULAR magnetic resonance (CMR) imaging presents a wide variety of different applications for the anatomical and functional assessment of the heart. The success of a CMR acquisition relies, however, on the ability of the MR operator to correctly tune the acquisition parameters to the subject being scanned [2]. Moreover, CMR can be negatively affected by a long list of imaging artefacts (caused for instance by respiratory and cardiac motion, blood flow and magnetic field inhomogeneities) [3]. Therefore, a quality control step is required to assess the usability of the acquired images. In the clinical practice this step is performed by visual inspection, usually carried out by the same operator who set up the acquisition, thus leading to highly subjective results. In the last decades, several initiatives for the acquisition of open access large-scale population studies have been launched. For example, the UK Biobank (UKBB) is a population-based prospective study, established to allow detailed investigations of the genetic and non-genetic determinants of the diseases of middle and old age. Of the 500,000 subjects enrolled in the study, CMR will be collected from 100,000 of them [1]. At the time of submission the acquisition is ongoing, with close to 20,000 subjects already scanned. Together with this trend towards the implementation of large-scale multi-centre imaging datasets, the need for fast and reliable quality control techniques for CMR images has become evident, as highlighted also by several studies aiming to define standardized criteria for this task [4]. In this scenario, quality control through visual inspection is not only subjective, but simply infeasible due to the very high throughput demanded by the acquisition pipeline. On the other hand, failure to correctly identify corrupted or unusable images could affect the results of automated analysis performed on the dataset, with undesirable effects. Consequently, the need for fully automated quality control pipelines for CMR images has arisen.

Many research efforts have been dedicated to the automated identification of quality metrics from MR images. Most of these efforts have focused on the automated estimation of noise levels [5], [6]. Still, many aspects related to the usability of

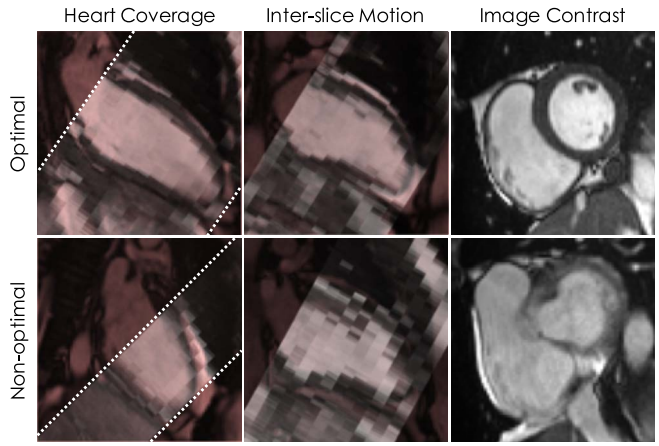

Fig. 1. Potential issues affecting CMR image acquisitions. In the first two columns, the superimposition of long-axis two-chamber views (red) and short-axis stacks (gray) is shown, while, in the last one, short-axis slices are displayed.

the acquired images are inherently modality-specific. Several automated pipelines for quality control have been proposed for brain MR imaging [7]. However, to our knowledge, no comprehensive automated quality control pipelines have been proposed so far for CMR images, in particular for the short-axis (SA) cine image stacks, which are the reference images for the structural and functional assessment of the heart. One crucial aspect of the acquisition of SA image stacks is that it requires the MR operator to identify the direction of the left ventricular (LV) long axis - the line going from the apex to the centre of the mitral valve - and to define a region of interest: the correct planning will generate a SA stack encompassing both those landmarks with slices perpendicular to the LV long axis. If this selection is incorrect, the acquired SA stack may include an insufficient number of SA slices to fully cover the LV (see first column of Fig. 1). As a consequence, any functional analysis performed on the stack (e.g. ventricular volumes estimation) may be compromised. Another important aspect involved in CMR acquisitions is that SA cine stacks are generated during multiple breath-holds (with usually 1-3 slices acquired per each breath-hold). Although the subjects are instructed to hold their breath at the same breath-holding position, in practice the heart location can vary considerably. If the differences between the breath-holding positions are too pronounced, the acquired image stack will be affected by inter-slice motion and thus will not correctly represent the cardiac shape, introducing potential errors in the following analyses and visualizations (see second column of Fig. 1). Finally, the contrast of the obtained CMR images is directly affected by the chosen acquisition parameters (as well as by potential artefacts). If the different structures of the heart are not properly contrasted, the assessment of the cardiac function can be hampered (see third column of Fig. 1).

In this paper, we present a fully-automated, learning-based quality control technique for CMR SA image stacks. Our approach uses a hybrid decision forest method to extract at once both landmark positions (LMs) and probabilistic segmentation maps (PSMs) from long-axis (LA) and SA images. LMs and PSMs are then used to perform three quality checks: 1) heart coverage estimation, 2) inter-slice motion detection, 3) image contrast estimation in the cardiac region. Our hybrid forest method is thus not intended as a novel technique for landmark detection and segmentation per se, but rather as an integral component of our pipeline. The extraction of LMs from multiple LA views and the probabilistic nature of PSMs allow the assessment of the reliability of the pipeline for each scan using dedicated sanity checks. The technique was tested on two datasets (up to 3000 cases from the UKBB study and 100 cases from the UK Digital Heart Project,[1] UKDHP) and validated against manual annotations and visual inspections.

II. Related Work

To the best of our knowledge, differently from brain MRI [7], no comprehensive quality control techniques for cardiac CMR images have been reported in the literature. One of the few studies in this direction has been recently presented by Albà *et al.* [8], who however focussed on assessing segmentation quality rather than image quality. On the other hand, automated heart coverage estimation alone has been the aim of several studies. Zhang *et al.* [9], [10] proposed to use convolutional neural networks (CNN) to perform slice classification in order to detect the presence or absence of the basal and apical slices. In their first work [9] they proposed a 2D CNN trained on UKBB data, while in their more recent one [10] they improved their previous results by using a generative adversarial network. Differently from these techniques, our approach to heart coverage estimation is based on the detection of landmarks: in our previous preliminary work [11], we proposed a decision forest method to detect the cardiac apex and the mitral valve on long-axis 2-chamber (LA 2CH) view images, and used the position of these landmarks with respect to the space encompassed by the acquired stack to estimate the coverage. The technique was applied to 3000 cases extracted from the UKBB, and was able to detect SA stacks with insufficient coverage with relatively high accuracy.

Motion detection and modeling in the thoracic area has been a highly investigated subject for more than a decade [12]. As far as inter-slice respiratory motion in CMR is concerned, most of the approaches reported in the last decade have focussed on motion correction rather than motion detection [13]–[16]. All of the cited studies focused on the compensation of inter-slice motion and in the generation of a corrected SA stack by means of rigid in-plane registration. Unfortunately, however, respiration causes a complex roto-translation of the heart in all three dimensions [17]: while most translation happens in the cranio-caudal direction (thus approximately almost perpendicularly to the long axis of the LV), big differences in subsequent breath-holding positions can cause out-of-plane motion, which would lead to an inaccurate representation of the heart in the stack. Therefore, it is important to estimate the amount of motion occurred during the acquisition of the

[1] https://digital-heart.org.

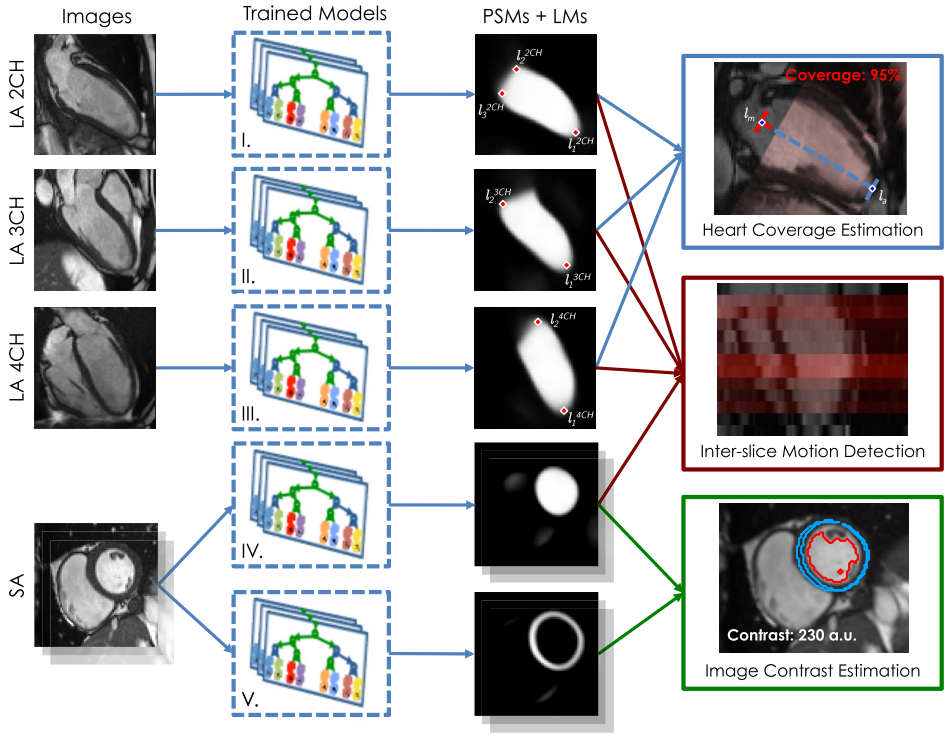

Fig. 2. Overview of the proposed pipeline. Probabilistic segmentation maps (PSMs) and landmark positions (LMs) are extracted from LA and SA images using hybrid random forests and exploited to perform three separate quality control checks.

stack to decide whether there are the grounds for the application of a motion correction technique or it is instead advisable to repeat the scan (or exclude it from subsequent analyses).

In the past, several research efforts have been made towards the correct quantification of signal-to-noise (SNR) or contrast-to-noise (CNR) ratios in MR images [5]. However, modern acquisition techniques making use of parallel imaging produce images with spatially-varying noise distributions, rendering image-based estimators unreliable [18]. To overcome this limitation, more elaborate methods have been proposed exploiting information about coil sensitivity or reconstruction coefficients [19]. Unfortunately, these data are very often not available, making the estimation of noise, and consequently of SNR and CNR, practically unfeasible in most scenarios. At the same time, image contrast between two objects - simply defined as the difference between their signal intensity - has long been used to determine their visual differentiability in the acquired MR image [20]. In CMR imaging, images with poor contrast between the LV cavity and myocardium can potentially hinder the assessment of cardiac structure and function: consequently, contrast estimation in the cardiac region can provide a useful metric for quality control purposes, either triggering the use of contrast-enhancing techniques or a new acquisition.

In this paper, we present a fully-automated, learning-based quality control pipeline for CMR SA stacks. The proposed approach builds upon our previous work [11], which used a hybrid decision forest method [21] to extract LMs from LA 2CH view images in order to perform heart coverage estimation. With respect to our previous approach as well as to state-of-the-art techniques, the main contributions of the present work can be listed as follows:

- We present the first comprehensive, fast, fully-automated quality control pipeline specifically designed for CMR SA image stacks. The checks incorporated in the pipeline are 1) heart coverage estimation, 2) inter-slice motion detection, 3) image contrast estimation in the cardiac region. To the best of our knowledge, motion detection and cardiac image contrast for the sake of quality control have not been investigated before. As for heart coverage estimation, we build on our previously published study [11] by extending LMs extraction to all long-axis views. LMs are then combined together to substantially increase the robustness and the reliability of this quality check (for details please refer to the Discussion section);
- We propose a different implementation of the previously published hybrid decision forest [21] (adopted in our previous work [11]) which allowed the joint extraction of LMs and probabilistic edge maps (PEMs). The new implementation (based on a novel mapping) allows instead the extraction of LMs and PSMs: PSMs are required to perform both inter-slice motion detection and cardiac image contrast estimation, and enable sanity checks to assess the reliability of the pipeline;
- We validate this pipeline by applying it to up to 3000 cases extracted from the UKBB study and to 100 cases from the UKDHP, showing its accuracy and robustness in real world scenarios. The pipeline could be both applied retrospectively on large-scale datasets to improve the reliability of clinical studies or deployed prospectively at acquisition sites to allow almost real-time assessment of the acquired scans.

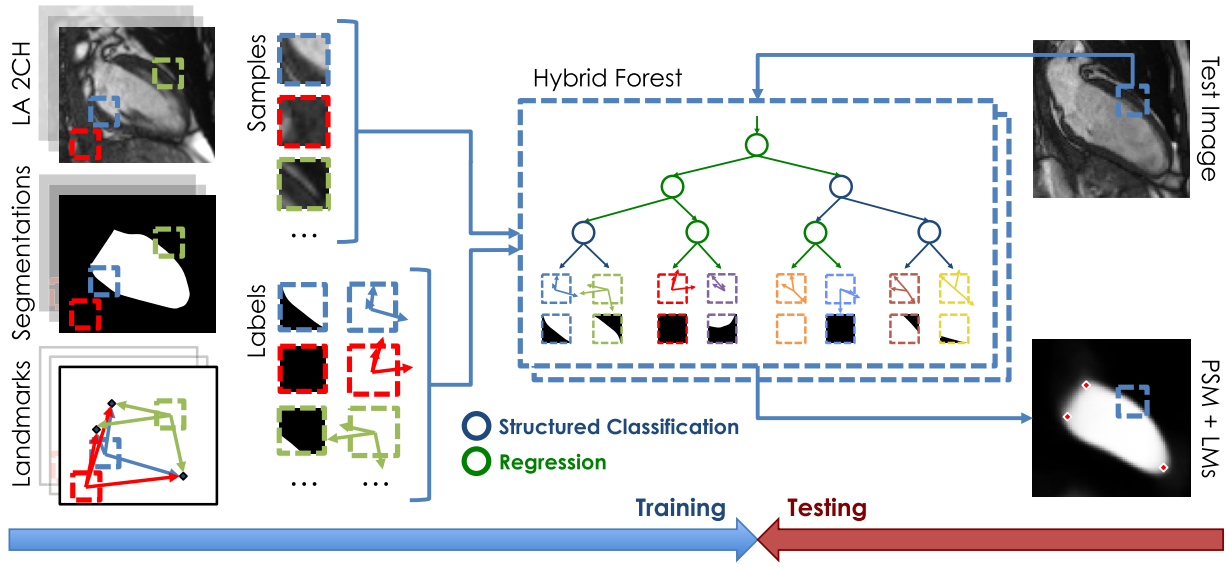

Fig. 3. Hybrid random forest. During training, randomly extracted samples with associated labels - consisting of segmentations and vector displacements - are fed to the forest, and the learnt associations are stored in the leaf nodes. During testing, each sample extracted from the test image is sent to the model, extracting both PSM and LMs at once.

III. Methods

The proposed quality control pipeline is summarized in Fig. 2. All of the three quality control steps are based on the information extracted by hybrid decision forest models from the acquired images. This section of the paper starts with a brief recap on the theory behind decision forests and is followed by the description of the implementation adopted in the proposed pipeline, which allows the joint extraction of LMs and PSMs. Finally, each specific quality control step is described in detail.

A. Hybrid Decision Forests

A decision tree consists of a combination of split and leaf nodes arranged in a binary tree structure [22]. Trees route a sample $\boldsymbol{x} \in \mathcal{X}$ (in our case an image patch) by recursively branching left or right at each split node j until a leaf node k is reached, where the posterior distribution $p(y|\boldsymbol{x})$ for the output variable $y \in \mathcal{Y}$ is stored. Each split node j is associated with a binary split function $h(\boldsymbol{x}, \boldsymbol{\theta}_j) \in \{0, 1\}$ defined by the set of parameters $\boldsymbol{\theta}_j$: if $h = 0$ the node sends $\boldsymbol{x}$ to the left, otherwise to the right. Usually h is a decision stump, i.e. a single feature dimension n of $\boldsymbol{x}$ is compared with a threshold τ: $\boldsymbol{\theta} = (n, \tau)$ and $h(\boldsymbol{x}, \boldsymbol{\theta}) = [\boldsymbol{x}(n) < \tau]$. A decision forest is an ensemble of T independent decision trees: during testing, given a sample patch $\boldsymbol{x}$, the predictions of the different trees are combined into a single output by means of an ensemble model. During training, at each node the goal is to find the set of parameters $\boldsymbol{\theta}_j$ which maximizes a previously defined information gain I_j, usually defined as $I_j = H(S_j) - \sum_{i \in \{0,1\}} |S_j^i|/|S_j| \cdot H(S_j^i)$, where S_j, S_j^0 and S_j^1 are respectively the training set (comprising of samples $\boldsymbol{x}$ and associated labels y) arriving at node j, leaving the node to the left and to the right. $H(S)$ is the entropy of the training set, whose construction depends on the task at hand (e.g. classification, regression). Different types of nodes (maximizing different information gains) can be interleaved within a single tree structure (hence named "hybrid") in order to perform multiple tasks. As in previous approaches [21], [23], in the proposed technique structured classification nodes (aiming at the detection of an object close to the desired landmarks, in our case usually the LV cavity) and regression nodes (aiming at landmark localization) are combined (see Fig. 3). In particular, in the proposed framework, landmark localization is conditioned on the results of the detection of the cavity [23]. This not only leads to the extraction of two different types of information (PSMs and LMs) with only one model, but improves landmark localization by implicitly incorporating complementary information about cardiac position and shape.

1) Structured Classification and PSM Extraction: Structured classification extends the concept of classification by using structured labels for $\mathcal{Y}$ instead of integer labels. In our case, each label $\boldsymbol{y} \in \mathcal{Y}$ (associated with the image patch $\boldsymbol{x}$) consists of a segmentation of the LV cavity within $\boldsymbol{x}$. To train a structured classification node it is necessary to find a way to cluster structured labels at each split node into two subgroups depending on a similarity measure. The solution to this problem was first proposed by Dollar and Zitnick [24] and consists of two steps. First, $\mathcal{Y}$ is mapped to an intermediate space $\mathcal{Z}$ by means of the function $\Pi : \mathcal{Y} \rightarrow \mathcal{Z}$ where the distance between labels can be computed. Importantly, Π must be chosen so that similar labels $\boldsymbol{y}$ will be associated with vectors $\boldsymbol{z}$ close to each other with respect to the distance defined in $\mathcal{Z}$. Then, PCA is applied to the vectors $\boldsymbol{z}$ to map the associated labels $\boldsymbol{y}$ into a binary set of labels $c \in \mathcal{C} = \{0, 1\}$, which is achieved by applying a binary quantization to the principal component of each $\boldsymbol{z}$ vector. Finally, the Shannon entropy can be adopted [24]:

$$H_{sc}(S) = -\sum_{c \in \mathcal{C}} p(c) log\big(p(c)\big), \tag{1}$$

with $p(c)$ indicating the empirical distribution extracted from the training subset at each node. In our previous work [21], this approach has been adopted for structured labels $\mathcal{Y}$ consisting of edge maps (EMs) highlighting the contours of the

myocardium. In the case of EMs, the mapping Π can simply encode for each pair of pixels whether they belong to the same segment in the label $\boldsymbol{y}$ or not:

$$\Pi_{EM} : \boldsymbol{z} = [\boldsymbol{y}(j_1) = \boldsymbol{y}(j_2)] \quad \forall j_1 \neq j_2, \tag{2}$$

where j_1 and j_2 are indices spanning every pixel in $\boldsymbol{y}$ [24]. The resulting long binary vector $\boldsymbol{z}$ (which has a number of dimensions equal to the number of pixel pairs in $\boldsymbol{y}$) can be used to compare this particular label to the other ones by simply computing the Euclidean distance in $\mathcal{Z}$. However, the same choice for Π cannot be adopted for our task, which aims at using structured labels consisting of segmentation maps (SMs) of the LV cavity. For example, let's imagine two labels $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, the former completely outside the LV cavity and the latter completely inside: using the mapping Π_{EM}, we would obtain $\boldsymbol{z}_1 = \boldsymbol{z}_2$, which contradicts the requirement by which only similar labels will be mapped close to each other in $\mathcal{Z}$. Consequently, we implemented a different mapping:

$$\begin{aligned}\Pi_{SM} : \boldsymbol{z} = [\boldsymbol{y}(j_1) = \boldsymbol{y}(j_2) = 0] \oplus \ldots \\ \ldots [\boldsymbol{y}(j_1) = \boldsymbol{y}(j_2) = 1] \quad \forall j_1 \neq j_2,\end{aligned} \tag{3}$$

which encodes for each pair of pixels in $\boldsymbol{y}$ whether they are both equal to 0, whether they are both equal to 1 and then concatenates the two obtained binary vectors. This formulation ensures the proper computation of the distance between labels, and thus their clustering at each node based on their similarity. At the end of the training process, the label $\hat{\boldsymbol{y}}$ stored in each leaf node is the one whose $\hat{\boldsymbol{z}}$ is the medoid (i.e. that minimizes the sum of distances to all the other $\boldsymbol{z}$ at the same node). At testing time, each sample patch of the test image is sent down each tree of the forest, and the segmentation maps stored at each selected leaf node are averaged, producing a smooth segmentation map (PSM) of the LV cavity. The values in the PSM are actual probabilities (proportional to the certainty in LV cavity detection), and can be used to assess the reliability of the prediction. Of note, the introduced formulation for Π_{SM} in Eq. 3 could be easily extended to multi-label PSM generation by concatenating additional binary vectors computed for each label c_i and by performing a channel-based averaging operation at testing time.

2) Regression and Landmark Detection: To train regression nodes, it is necessary to associate with each sample patch $\boldsymbol{x}$ an additional label $\mathcal{D} = (\boldsymbol{d}^1, \boldsymbol{d}^2, \ldots, \boldsymbol{d}^L)$, where $\boldsymbol{d}^l$ represents for each of the L landmarks the N-dimensional displacement vector from the patch centre to the landmark location. Instead of the Shannon entropy defined in Eq. 1, regression nodes are trained by minimizing the determinant of the covariance matrix $|\Lambda(S)|$ defined by the landmark displacement vectors:

$$H_r(S) = \frac{1}{2} log\big((2\pi e)^d |\Lambda(S)|\big). \tag{4}$$

Landmark positions are assumed to be uncorrelated, thus only the diagonal elements of $\Lambda(S)$ are used in Eq. 4 [25]. The location predictions are stored at each leaf node k using a parametric model following a $N \cdot L$-dimensional multivariate normal distribution with $\overline{\boldsymbol{d}_k^l}$ and Σ_k^l mean and covariance matrices, respectively. At testing time, Hough vote maps are generated for each landmark by summing up the posterior distributions obtained from each tree for each patch (applying normalization factors) [23]. Assuming that pixels belonging to the LV are more informative for cardiac landmark detection than background ones, the PSM values for the LV cavity are used for each patch as weighting factor during the generation of the L Hough vote maps, effectively restricting voting rights only to pixels likely to belong to the LV cavity itself [21]. Finally, the location of a landmark is determined by identifying the pixel with the highest value on each Hough vote map.

3) Model Training: Each patch $\boldsymbol{x}$ is represented by several features: multi-resolution image intensity, histogram of gradients (HoG) and gradient magnitude. For a detailed description, please refer to [21]. The described hybrid random forest approach is used to build five different models (I-V) for our application (see Fig. 2): PSM estimation of LV cavity and LMs extraction for apex and mitral valve for LA images, PSM of LV cavity and LV myocardium for SA stacks. For LA 3CH and 4CH images (models II and III) only one mitral valve point is identified because in these images the LV outflow tract of the aorta can partially occlude one side of the mitral valve, making its localization inaccurate. Also, the training of the models using SA images (models IV and V) is performed by feeding the random forests with all the slices extracted from the SA image stacks: consequently, at testing time, the models are applied to each slice of the stack independently.

B. Heart Coverage Estimation

Heart coverage is estimated exploiting the landmarks identified on LA 2CH, 3CH and 4CH images using the previously trained hybrid forest models. The rationale is that a properly scanned SA stack should encompass the whole portion of space between the apex and the mitral valve. As highlighted in Fig. 2, for a specific subject we identify three landmarks for the apex (one per each LA image: $\boldsymbol{l}_1^{2CH}$, $\boldsymbol{l}_1^{3CH}$ and $\boldsymbol{l}_1^{4CH}$) and four for the mitral valve ($\boldsymbol{l}_2^{2CH}$, $\boldsymbol{l}_3^{2CH}$, $\boldsymbol{l}_2^{3CH}$, $\boldsymbol{l}_2^{4CH}$) with values in the coordinate systems of each respective LA image. Using the orientation matrix extracted from the DICOM headers of the acquired SA and LA images, it is possible to define the coordinates of these landmarks in the coordinate system of the SA stack itself. Two new "median" landmarks ($\boldsymbol{l}_a$ and $\boldsymbol{l}_m$) are then defined taking the medians of the coordinates of the landmarks for the apex and for the mitral valve, respectively, in the SA coordinate system. The extension in the z direction (i.e. along the LV long axis) of the SA stack can be easily computed from the slice thickness and slice number, which are stored in the DICOM header of the stack itself: the two extrema along this direction are defined r_a and r_m, respectively. Finally, the relative coverage can be computed by comparing the relative positions along the z direction of $\boldsymbol{l}_a$ and $\boldsymbol{l}_m$ (i.e. the space that is supposed to be covered by the SA stack) to the portion of space between r_a and r_m (i.e. the space that is actually covered). The steps for coverage estimation are listed in Algorithm 1, including the formula for the computation of the coverage (under the assumption that the apex is located at higher z compared to

Algorithm 1 Heart Coverage Estimation

Input landmarks:
Apex: $l_1^{2CH}, l_1^{3CH}, l_1^{4CH}$
Mitral Valve points: $l_2^{2CH}, l_3^{2CH}, l_2^{3CH}, l_2^{4CH}$
Change coordinate system:
Apex: $\hat{l}_1^{2CH}, \hat{l}_1^{3CH}, \hat{l}_1^{4CH}$
Mitral Valve points: $\hat{l}_2^{2CH}, \hat{l}_3^{2CH}, \hat{l}_2^{3CH}, \hat{l}_2^{4CH}$
Compute median landmarks:
$l_a = \text{median}(\hat{l}_1^{2CH}, \hat{l}_1^{3CH}, \hat{l}_1^{4CH})$
$l_m = \text{median}(\hat{l}_2^{2CH}, \hat{l}_3^{2CH}, \hat{l}_2^{3CH}, \hat{l}_2^{4CH})$
with z-components l_a and l_m, respectively
Extract SA stack extension in the z direction:
Apex: r_a
Base: r_m
Compute coverage CV:

$$CV = \begin{cases} \dfrac{max\left(0, min(r_a, l_a) - max(r_m, l_m)\right)}{l_a - l_m} & \text{if (condition)} \\ \dfrac{r_a - r_m}{l_a - l_m} & \text{otherwise} \end{cases}$$

(condition): $r_a < l_a$ or $r_m > l_m$

the mitral valve). Importantly, this technique can seamlessly be applied even if only one LA image is available. Also, while minor motion can occur between the acquisitions of LA images and of the SA stack, it is generally negligible in the z direction (the only one influencing coverage) [17] and thus registration procedures between these images were found to be unnecessary. Finally, a sanity check is performed to detect cases in which landmark detection failed: for each LA view, when either of the relative distances between the landmarks was greater or smaller than reference values by a certain threshold, the landmarks from that image were discarded, and the automated coverage estimation was performed only on the remaining landmarks (if available).

C. Inter-Slice Motion Detection

Inter-slice motion detection relies on the PSMs extracted from the acquired images. The rationale is that while LV cavity PSMs of motion-corrupted SA slices are misaligned, PSMs extracted from the LA images represent sections of the true shape of the LV cavity and can consequently be used as reference. Moreover, the amount of misalignment between the SA PSMs and LA PSMs can be used as an indicator of motion. To perform this assessment, the LA PSMs are initially rigidly registered (by 3D translation only, using sum of squared differences as dissimilarity metric) to the SA PSM stack to compensate for potential motion between different acquisitions. Then, for each slice of the SA PSM stack, the three registered LA PSMs are resampled and combined into a single image (referred to as combined LA PSM) containing the sections of the LA PSMs with respect to a specific slice (see Fig. 4). Finally, in-plane rigid registration (by translation only, using sum of squared differences as dissimilarity metric) is performed between each SA PSM slice

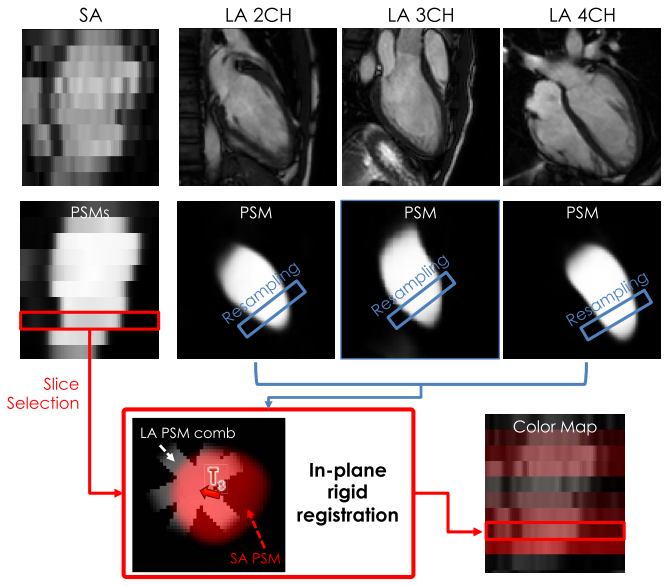

Fig. 4. Motion detection technique. For each slice of the SA stack, the corresponding portion of space in each LA PSM is resampled and combined, producing the "asterisk-shaped" LA PSM comb image. In-plane rigid registration is then performed between each SA PSM and LA PSM comb, and the translation magnitude T_s used as proxy for inter-slice motion for that slice. A color map, with the intensity of each slice proportional to the respective T_s, can be also generated.

and the associated combined LA PSM, and the magnitude of the translation T_s used as a metric for motion (i.e. differences in breath-holding positions). Of note, this step is performed only on the slices which are effectively covering the LV, condition assessed using the LA LMs as in Algorithm 1. The probabilistic nature of PSMs allows the application of a sanity check performed to detect slices with a failed PSM estimation: SA PSM slices (whose values range between 0 and 1024) with a peak probability value below a user-defined threshold are considered unreliable, and thus their T_s discarded. Also, this technique could be applied even if only two LA images were available. The steps for motion detection are listed in Algorithm 2.

Algorithm 2 Inter-Slice Motion Detection

Input PSMs:
LA images: $PSM^{2CH}, PSM^{3CH}, PSM^{4CH}$
SA slices: $PSM_s^{SA-Cav}, s = (1, \ldots, numSlices)$
Perform rigid registration of LA PSMs to SA PSM:
Output: $\overline{PSM}^{2CH}, \overline{PSM}^{3CH}, \overline{PSM}^{4CH}$
for $s = 1$ **to** $numSlices$ **do**
  Resample LA PSMs:
  Output: $\overline{PSM}_s^{2CH}, \overline{PSM}_s^{3CH}, \overline{PSM}_s^{4CH}$
  Combine resampled LA PSMs:
  Output: $\overline{PSM}_s^{LA\_comb}$
  Perform in-plane rigid registration of PSM_s^{SA-Cav} **to** $\overline{PSM}_s^{LA\_comb}$**:**
  Output: Translation magnitude T_s
end

D. Cardiac Image Contrast Estimation

Cardiac image contrast is estimated using the LV cavity and LV myocardium PSMs extracted from the SA stack. The rationale is to transform the PSMs into hard segmentations (SMs) and to use them to estimate the difference between average pixel intensity in the LV cavity and in the LV myocardium. Each cavity PSM slice is thresholded selecting the N_{cav} pixels with the highest probability values: this will maximize the probability of measuring the intensity in the actual cavity. The same happens to each myocardium PSM, thresholded selecting N_{myo} pixels. To exclude potential spurious regions from the obtained segmentation, the average centroid for the cavity segmentation is computed among the different slices, and for each slice only the connected component closest to the average centroid is kept, both for the cavity and the myocardium segmentations. Of note, this step is performed taking into account the slice-by-slice rigid transformation estimated using Algorithm 2, which amounts to performing the average centroid computation and connected components analysis on a motion-compensated stack. Finally, in order to exclude potential papillary muscles from the cavity intensity computation, a Gaussian mixture model is fitted to the distribution of intensity values inside the cavity segmentation. Since only some slices show papillary muscles, both a two-component and a one-component models are used, and only one is selected based on the Akaike information criterion [26]. If the two-component model yields the best fit, since the cavity distribution is always higher than that of papillary muscles, the mean of the component with the highest mean is used as average intensity value for the cavity. For the myocardium, the mean intensity of the pixels masked by the segmentation is computed. Cardiac image contrast is finally defined as the difference between these two values. A double sanity check is performed leveraging the probabilistic nature of PSMs: if either the peak value of either the cavity or the myocardium PSM was below a user-defined threshold or the size of either of the final hard segmentations for the cavity or the myocardium was less than a defined number of pixels, the obtained contrast was deemed unreliable. The steps for cardiac image contrast estimation are also listed in Algorithm 3.

Algorithm 3 Cardiac Image Contrast Estimation

Input PSMs:
LV cavity: PSM_s^{SA-Cav}
LV myocardium: PSM_s^{SA-Myo}
with $s = (1, \ldots, numSlices)$
for $s = 1$ **to** *numSlices* **do**
  Threshold PSM_s^{SA-Cav} and PSM_s^{SA-Myo}**:**
    Output: SM_s^{SA-Cav} and SM_s^{SA-Myo}
  Estimate centroids $\mathbf{P}_s$ **for** SM_s^{SA-Cav}
end
Estimate mean centroid: $\mathbf{P} = \text{mean}(\mathbf{P}_s)$
for $s = 1$ **to** *numSlices* **do**
  Exclude all but one connected component per SM based on distance to P:
    Output: $\overline{SM}_s^{SA-Cav}$ and $\overline{SM}_s^{SA-Myo}$
  Fit Gaussian Mixture Model to $\overline{SM}_s^{SA-Cav}$ **to exclude papillary muscles:**
    Output: μ_s^{SA-BP}
  Compute contrast CT:
    $CT = \mu_s^{SA-BP}$ - mean$\left(SA_s\left(\overline{SM}_s^{SA-Myo}\right)\right)$
    with SA_s the s-slice of the SA stack
end

E. Performance Evaluation

1) Image Acquisition: To train and test the proposed quality control pipeline, images from two different datasets were used: the UKBB [1] and the UKDHP[1]. CMR imaging for the UKBB was performed using a 1.5T Siemens® MAGNETOM Aera system equipped with a 18 channels anterior body surface coil (45 mT/m and 200 T/m/s gradient system). 2D cine balanced steady-state free precession (b-SSFP) SA image stacks were acquired with in-plane spatial resolution 1.8×1.8 mm, slice thickness 8 mm, slice gap 2 mm, image size 198×208 and average number of slices 10. 2D cine b-SSFP LA images were acquired with in-plane spatial resolution 1.8×1.8 mm, slice thickness 8 mm and image size 162×208. Further acquisition details can be found in [1]. CMR imaging for the UKDHP was performed on healthy volunteers using a 1.5T Philips® Achieva system equipped with a 32 element cardiac phased-array coil (33 mT/m and 160 T/m/s gradient system). 2D cine balanced steady-state free precession (b-SSFP) SA image stacks were acquired with in-plane spatial resolution 1.2×1.2 mm, slice thickness 8 mm, slice gap 2 mm, image size 288×288 and average number of slices 12. 2D cine b-SSFP LA images were acquired with in-plane spatial resolution 1.5×1.5 mm, slice thickness 8 mm and image size 256×256. In both datasets, only end-diastolic frames were considered.

2) Experimental Design: A series of experiments was conducted to assess the accuracy of each portion of the pipeline. First of all, the five hybrid random forest models were trained using a randomly-generated subset of 500 cases from the UKBB. For each LA image-based model, the 500 images were used together with manually-annotated landmarks and segmentations of the LV cavity. The segmentations were obtained with a CNN-based automated tool proven to reach human-level performance [27], and then visually checked for accuracy. Each training set was quadrupled in size through data augmentation applying random rescaling (following a normal distribution with $\mu = 1, \sigma = 0.1$) and random rotation ($\mu = 0°, \sigma = 30°$). For each of the two SA stack-based models, the slices extracted from the 500 stacks were used (for a total of 5165 images) together with segmentations of the LV cavity and of the LV myocardium, respectively (obtained using the same process described for LA images). Details regarding forest training include image patch size 48×48 px for LA models and 32×32 px for SA ones, segmentation label size 16×16 px, number of samples $4{\cdot}10^6$, number of trees $T = 8$.

A first series of experiments was performed by evaluating the trained pipeline on a separate testing set consisting of 3000 cases randomly extracted from the UKBB. To evaluate

the accuracy of the proposed heart coverage estimation technique, two experiments were conducted. First, for each of the three LA views, the positions of the landmarks were manually annotated on 100 randomly selected cases. The automatically detected LMs were compared to the manually identified ones by measuring the Euclidean distance between the two sets of points. Then, the 3000 SA stacks were visually inspected (sometimes using LA images as reference) to identify cases with insufficient coverage, defined as such when at least one full slice was missing. Automated heart coverage estimation was then performed on the same dataset. To instruct the previously described sanity check, the mean and standard deviation of the relative distances between manually annotated landmarks were computed on the 100 images ($\overline{l_2 - l_1} = 89 \pm 12$ mm, $\overline{l_3 - l_2} = 32 \pm 5$ mm, $\overline{l_3 - l_1} = 87 \pm 12$ mm): then, for each LA view, when either of the relative distances between the automatically detected LMs was over 2 standard deviations greater or smaller than the respective mean distance value (thus covering roughly 95% of the measured variability), the LMs from that image were discarded and the automated coverage estimation was performed only on the remaining ones (if available). Finally, the accuracy of the technique was assessed against the performed visual inspection performing a standard binary classification test using a threshold for insufficient coverage optimized automatically with an ROC analysis. To evaluate the accuracy of the motion detection technique, two experiments were conducted. First, for each of the three LA views as well as for the SA stacks, the automatically extracted PSMs were compared to hard segmentations obtained using the previously-described CNN-based automated tool [27] on 1000 randomly selected cases. While this experiment was aimed at assessing the accuracy of the PSMs, it is worth noting that the PSMs are never directly thresholded for segmentation purposes in the pipeline, which on the contrary exploits their probabilistic nature. For the sake of this comparison, the PSMs were turned into hard segmentation by applying a global threshold and compared to the reference ones by computing the Dice coefficient (DSC). The global threshold was optimized automatically using an ROC analysis. Then, 1500 SA stacks were visually inspected (sometimes using LA images as reference) to identify cases with noticeable motion corruption. Automated motion detection was then performed on the same dataset. To implement the previously described sanity check, PSM slices with peak probability values below 600 were considered not reliable for motion detection, and thus their T_s (i.e. the estimated translation magnitude) discarded; if less than 2 T_s values were left, the motion detection analysis was not performed on the specific stack. Accuracy of the automated technique was assessed against visual inspection with a standard binary classification test using the following criterion: a stack was deemed motion-corrupted if either the average T_s was above a first threshold T_A or at least two T_s were above a second threshold T_B. This double criterion aimed at the detection of both stacks with a few, clearly misaligned slices as well as stacks with poor general alignment. Both T_A and T_B were optimized automatically using an ROC-like approach. To evaluate the accuracy of the cardiac image contrast estimation technique, 100 random slices from as many random SA stacks were manually annotated selecting regions of interests (ROIs) within the LV cavity and the LV myocardium. Cardiac image contrast was estimated both from the original images and from the images after contrast normalization using a randomly selected reference image stack. Automated contrast estimation was then performed on the same dataset, both before and after normalization, using $N_{cav} = 450$ px and $N_{myo} = 200$ px. To implement the previously described sanity check, contrast extracted from slices with PSMs (either for the cavity or for the myocardium) with peak values below 150 or with respective hard segmentations with a size of less than 32 mm^2 (i.e. 10 pixels) was deemed unreliable and excluded from the analysis. Automatically estimated and manually computed contrast values were compared using Pearson's correlation coefficient, linear regression and Bland-Altman analyses.

A second series of experiments was then performed by evaluating the pipeline trained on UKBB on a separate testing set consisting of 100 cases randomly extracted from the UKDHP. Since the scans in UKDHP were acquired with a different scanner and with different parameters from those used for UKBB, these experiments were aimed at assessing the generalization properties of the proposed pipeline. To harmonize the differences between training and testing datasets, the images in UKDHP were pre-processed through intensity normalization [28], spatial resampling and image reorientation. The 100 cases were then visually inspected and manually annotated following the same criteria described for the previous experiments to provide the ground truth for estimation coverage, motion detection and contrast estimation. Since the visual assessment for heart coverage estimation returned no sub-optimal cases, a procedure was implemented to simulate coverage issues and allow a more meaningful evaluation of the pipeline. Stacks were randomly picked following a uniform distribution (10% chances of being picked), and a number of slices were deleted (either from the top or the bottom of the stack with equal probability), with this number randomly selected from a normal distribution ($\mu = 1$, $\sigma = 2$). Coverage was then visually re-assessed on the whole dataset. It is important to note that while this corruption procedure altered the properties of the dataset with respect to coverage, it did not affect the images on which the learning-based portion of the pipeline is applied (i.e. the LA images) but only the SA stacks, which influence the coverage estimation by means of their size and spatial orientation. The pipeline was applied with the same settings used for the previous dataset except for the threshold for the sanity check for contrast estimation relative to the peak PSM value, which was moved from 150 to 100 to account for the slightly lower overall response in the PSMs. The evaluation strategy for the three checks was the same as for the previous set of experiments.

For all the experiments, manual annotations and visual inspections used as ground truth were performed internally by G. T. (medical imaging researcher with 10 years of experience in cardiac imaging) and H. S. (experienced cardiologist), both blinded to the results of the automated analyses: more specifically, H. S. visually inspected the 3000 SA stacks from

the UKBB dataset to identify cases with insufficient coverage, and G. T. performed all of the remaining assessments.

IV. Results

The experiments were initially run on a single core of an Intel® Xeon CPU E5-1650 v3 @ 3.50GHz with 64 GB of memory to assess the speed of the current pipeline. Average time required to extract PSMs and LMs (when included in the model) was 1.3s per SA stack (of roughly 10 slices) and 0.85s per LA image. Average times required to perform the quality control checks were 0.26s per SA stack for coverage estimation, 9s per SA stack for motion detection (in this case using parallelization on 6 cores to evaluate multiple slices from one stack at once) and 0.6s per slice for contrast estimation.

In Table I are reported the localization errors for landmark detection on UKBB for the three LA views.[2] Of note, the landmarks extracted from one image per LA view were identified as outliers and thus excluded from the reported results. Mean DSC values between thresholded PSMs (using a threshold of 450) and reference segmentations were respectively 0.90 ± 0.07 for the SA stacks, 0.94 ± 0.08 for LA 2CH, 0.94 ± 0.08 for LA 3CH, and 0.94 ± 0.07 for LA 4CH.

First are reported the results for quality control on UKBB. For accuracy assessment of heart coverage estimation, 3 of the 3000 cases were excluded from the analysis: one due to the lack of LA images, and two for failing the sanity check on all the LA images. The ROC analysis performed on the remaining 2997 images returned an optimal threshold of 90%. The results of the binary classification test are reported in Table II. For accuracy assessment of motion detection, 3 of the 1500 cases were excluded from the analysis: one due to the lack of the SA stack and two for failing the sanity check. An ROC-like analysis was performed on the remaining 1497 images to select the thresholds T_A and T_B. The results of the binary classification test, obtained for $T_A = 3.4$ mm and $T_B = 6$ mm, are reported in Table III. For accuracy assessment of contrast estimation, 3 of the 100 images were excluded from the analysis for failing the sanity check. In Table IV are reported Pearson's correlation coefficients as well as the results of linear regression and Bland-Altman analyses between automatically and manually estimated contrast values.[3] In Figs. 5, 6, 7 are shown examples of the results obtained for the three checks.[4]

[2]A box plot is also presented in Fig. 8 (Media/Supplementary Material).

[3]Plots are also presented in Fig. 10 (Media/Supplementary Material).

[4]Further examples are also presented in Fig. 9 (Media/Supplementary Material).

TABLE I
LANDMARK LOCALIZATION ERRORS IN mm (MEAN ± STD)

Landmark Detection			
Localization Error	LA 2CH	LA 3CH	LA 4CH
Apex	4.2 ± 2.5	4.5 ± 3.0	4.6 ± 3.1
Mitral Valve (Side I)	3.6 ± 2.6	3.5 ± 2.6	3.2 ± 2.3
Mitral Valve (Side II)	3.9 ± 2.7		

TABLE II
CLASSIFICATION RESULTS FOR HEART COVERAGE ESTIMATION ON UKBB USING A 90% COVERAGE THRESHOLD. POSITIVE CASES CORRESPOND TO CASES WITH INSUFFICIENT COVERAGE

Heart Coverage Estimation (UKBB)				
Sensitivity	Specificity		Visual Assessment	
88%	99%	Proposed Technique	49 (TP) 7 (FN)	15 (FP) 2926 (TN)

TABLE III
CLASSIFICATION RESULTS FOR MOTION DETECTION ON UKBB USING $T_A = 3.4$ mm AND $T_B = 6$ mm. POSITIVE CASES CORRESPOND TO MOTION-CORRUPTED CASES

Inter-Slice Motion Detection (UKBB)				
Sensitivity	Specificity		Visual Assessment	
85%	95%	Proposed Technique	213 (TP) 39 (FN)	58 (FP) 1187 (TN)

TABLE IV
CORRELATION COEFFICIENT (R), BIAS AND STD FOR BLAND-ALTMAN ANALYSIS, LINEAR REGRESSION COEFFICIENTS (A AND B) AND MEAN MEASURED VALUE BETWEEN AUTOMATICALLY AND MANUALLY ESTIMATED CARDIAC IMAGE CONTRAST ON UKBB, BOTH ON ORIGINAL IMAGES AND AFTER HISTOGRAM NORMALIZATION, IN A.U.

Cardiac Image Contrast Estimation (UKBB)						
	R	Bias	Std	A	B	Mean
Original Images	0.95	-0.6	12.1	0.96	7.8	190
Normalized Images	0.94	-0.7	12.4	0.97	5.3	169

Then are reported the results on UKDHP. For accuracy assessment of heart coverage estimation, all cases passed the sanity check. The ROC analysis returned an optimal threshold of 92% coverage, and the results of the subsequent binary classification test are reported in Table V. For accuracy assessment of motion detection, 1 of the 100 cases was excluded from the analysis for failing the sanity check. The ROC-like analysis was performed on the remaining 99 images to select the thresholds T_A and T_B. The results of the binary classification test, obtained for $T_A = 3$ mm and $T_B = 6$ mm, are reported in Table VI. For accuracy assessment of contrast estimation, 9 of the 100 images were excluded from the analysis for failing the sanity check, and in Table VII are reported the results obtained on the remaining ones.[3]

V. Discussion

The results obtained for the landmark localization experiment show that the average localization error is around 3.9 mm (roughly two pixels) and is thus small compared to the reconstructed slice thickness in both datasets (10 mm), suggesting the reliability of the landmark detection technique for the sake of heart coverage estimation. The proposed hybrid decision forest method is based upon a previous implementation [21] which consisted of a multi-stage approach devised to increase the robustness to large variations in distances and orientation of the landmarks. It is worth mentioning that initial experiments performed using this approach showed no measurable

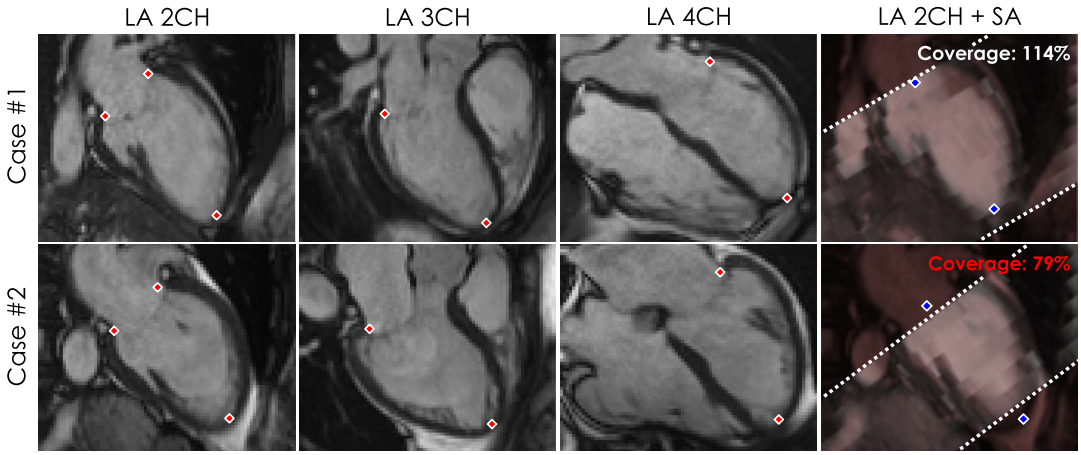

Fig. 5. Results for heart coverage estimation on UKBB in two cases, one with sufficient (case #1) and one with insufficient coverage (case #2). In the first three columns, the results for landmark detection in the three LA views. In the last column, a mix view with the LA two-chamber view and the SA stack together with the median landmarks for the apex and the mitral valve.

TABLE V
CLASSIFICATION RESULTS FOR HEART COVERAGE ESTIMATION ON UKDHP USING A 92% COVERAGE THRESHOLD. POSITIVE CASES CORRESPOND TO CASES WITH INSUFFICIENT COVERAGE

Heart Coverage Estimation (UKDHP)				
Sensitivity	Specificity		Visual Assessment	
100%	100%	Proposed Technique	5 (TP) 0 (FN)	0 (FP) 95 (TN)

TABLE VI
CLASSIFICATION RESULTS FOR MOTION DETECTION ON UKDHP USING $T_A = 3$ mm AND $T_B = 6$ mm. POSITIVE CASES CORRESPOND TO MOTION-CORRUPTED CASES

Inter-Slice Motion Detection (UKDHP)				
Sensitivity	Specificity		Visual Assessment	
78%	90%	Proposed Technique	14 (TP) 4 (FN)	8 (FP) 73 (TN)

TABLE VII
CORRELATION COEFFICIENT (R), BIAS AND STD FOR BLAND-ALTMAN ANALYSIS, LINEAR REGRESSION COEFFICIENTS (A AND B) AND MEAN MEASURED VALUE BETWEEN AUTOMATICALLY AND MANUALLY ESTIMATED CARDIAC IMAGE CONTRAST ON UKDHP, BOTH ON ORIGINAL IMAGES AND AFTER HISTOGRAM NORMALIZATION, IN A.U.

Cardiac Image Contrast Estimation (UKDHP)						
	R	Bias	Std	A	B	Mean
Original Images	0.94	-12.3	27.3	0.98	-4.9	335
Normalized Images	0.94	-15.5	38.8	0.96	5.4	498

improvement with respect to the single-stage one (perhaps due to the size of the training set and to the consistency of the orientation of the images), which was thus preferred.[5] The high DSC values obtained for the PSMs suggest their reliability for both motion detection and contrast estimation. The fact that the PSMs of the SA stacks are slightly worse than those of the LA images (0.90 vs 0.94) is mainly due to a lower response of the model in the apical slices, where a different thresholding value would have been beneficial. However, this does not cause a direct problem on the proposed pipeline, which never thresholds PSMs for segmentation purposes and instead exploits their probabilistic nature.

[5] A more in-depth comparison between the two implementations is presented in Fig. 11 (Media/Supplementary Material).

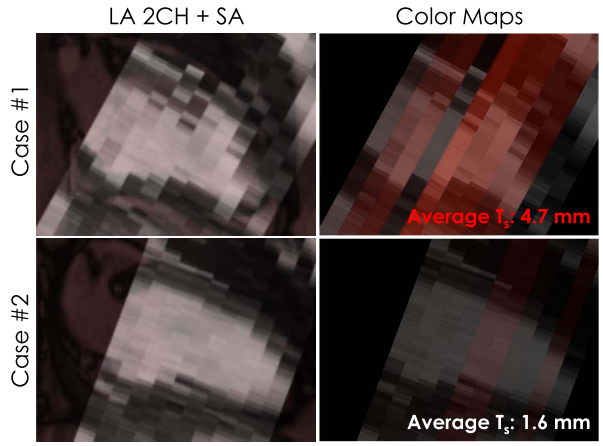

Fig. 6. Results for motion detection on UKBB in two cases, one with (case #1) and one without motion corruption (case #2). In the second column, the color maps of the translation magnitude for each slice are overlaid on top of the SA stacks.

The first set of experiments involving the whole pipeline was aimed at assessing its accuracy on UKBB. The binary classification test on coverage estimation performed on 2997 cases from UKBB indicates the high accuracy of the proposed technique, with sensitivity = 88% and specificity = 99%. The interpretation of these results is hindered by the strong class imbalance between cases with sufficient and insufficient coverage, and thus a more detailed analysis of the reported confusion matrix is required. By applying the proposed automated technique, it is possible to correctly detect 88% of the cases with insufficient coverage, and thus to lower the percentage of undetected wrongly imaged cases from 1.9% to 0.2%. This comes at the price of having to visually check an additional 0.5% of cases that actually featured a sufficient coverage. Notably, several of the 15 FP cases actually had a sub-optimal coverage, but not of the amount required to be considered as wrongly imaged following the criterion adopted during visual inspection. Compared to our

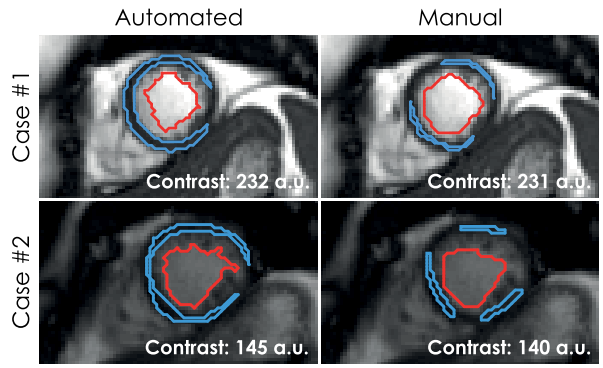

Fig. 7. Results for contrast estimation on UKBB in two cases, one with high (case #1) and one with low contrast (case #2). The ROIs from which the mean intensities are estimated are shown in red and cyan.

previous work [11], the present approach makes use of three LA images instead of just one. The redundancy offered by exploiting all the available LA views allows a more robust and reliable estimation: this is suggested by the higher sensitivity and specificity achieved (88% vs 73% and 99% vs 98%, respectively, although a direct comparison is not completely fair since the UKBB subset used in [11] was different from the present one) and by the lower number of cases excluded due to failing the sanity check (down from 89 to 3). Of note, this check is able to indirectly detect and exclude LA images with high noise levels, wrong acquisition planning or wrong file naming that make the landmark localization unreliable, and in the present implementation only cases in which all the three LA views yielded bad landmark detection had to be excluded from the coverage assessment. Zhang *et al.* [9] addressed coverage estimation by performing fully-supervised CNN-based slice classification to detect stacks with missing basal (MBS) or apical slices (MAS). In their later work [10], they acknowledged the need for a large amount of labelled data during training to achieve good generalization: to mitigate this issue, the authors have increased the size of the training set using generative networks (reaching average accuracies of 93% for MAS and 89% for MBS on a dataset of 3400 cases from UKBB). The use of different subsets of data from the UKBB and the different validation strategies (detection of missing slices separately in the apical and in the basal region vs detection of overall non-optimal cases) make the comparison between the two approaches not straightforward. The main advantage of the approach of Zhang *et al.* is that it can detect problematic scans using only the SA stack, while our pipeline relies on the presence of at least one of the LA views (which are, however, routinely acquired in most CMR protocols). On the other hand, we believe there is a clinical and practical advantage in measuring the relative coverage instead of performing binary classification: cases with only slightly sub-optimal coverage could still be included in the following analyses, especially when the lack of coverage is in the apical area. Moreover, while their approach completely relies on feature extraction from single slices and thus small image perturbations can potentially lead to misclassification, our approach is designed to exploit the redundancy offered by the multiple LA views for greater robustness.

The reported results on UKBB for motion detection indicate that the proposed approach achieves sensitivity = 85% and specificity = 95% over 1497 cases. By applying the proposed automated technique, it is possible to lower the percentage of undetected motion-corrupted cases from 16.8% to 2.6%. This comes at the price of having to visually check 3.9% cases that were visually deemed motion-free. It is worth to note that the binary classification of stacks based on the visual assessment of motion is a difficult task in itself, limiting the measurable accuracy of any technique.

The accuracy of the contrast estimation technique on UKBB is indicated by very high correlation coefficients and regression lines near unity both for images before and after contrast normalization. Bland-Altman analyses show negligible biases and narrow limits of agreement with respect to the mean measured values, suggesting the high accuracy of the technique.

The second set of experiments was aimed at assessing the accuracy of the pipeline (trained on UKBB) on the UKDHP dataset, thus testing its generalization properties, and yielded encouraging results. Regarding heart coverage estimation, our technique was able to correctly identify all sub-optimal cases. Regarding motion detection, it returned slightly lower values for sensitivity and specificity than those obtained on UKBB: while this might be due to a lower accuracy of the extracted PSMs, we noted that motion in the UKDHP dataset is considerably less pronounced than on UKBB, so it is easier to misclassify borderline cases. Regarding contrast estimation, the technique showed again very high correlation coefficients and regression lines near unity. The increased difficulty in dealing with a testing dataset different from the training one can be seen in the slightly higher number of cases failing the sanity check (up from 3 to 9) and in bigger biases, still however negligible when compared to the mean measured values. In general, the small size of the UKDHP dataset should be taken in consideration when evaluating these results, especially for binary classification tests where the misclassification of a single case can have a very large influence on the accuracy figures. However, we believe the reported results show that the proposed approach generalizes well to previously unseen datasets, coping with differences in the acquisition protocols.

Our approach to quality control does not attempt to directly classify sub-optimal cases for three reasons. First, this allows the complete circumvention of any potential class-imbalance issues, since the only learning-based portions of our pipeline aim at the identification of structures that are present in every image. Second, landmark extraction and probabilistic segmentation allow the assessment of the reliability of the pipeline by means of simple sanity checks, less trivial to implement in classification approaches like [9]. Third, our pipeline does not work as a "black-box": each quality check produces quantitative metrics with a clear meaning, which can be of great value in informing the MR operators on the type and the entity of the identified issues. Importantly, the proposed pipeline could be adopted also using different techniques for landmark detection and probabilistic segmentation. One major requirement for these alternative methods would be the generation of fuzzy segmentations maps providing a probabilistic representation of the target structures: this allows the assessment of their

reliability for both motion detection and contrast estimation, otherwise unfeasible with standard, hard segmentations.

The main limitation affecting our approach is that no quality check is performed on the manual selection of the imaging planes for LA and SA images, which can be subject to error. However, countermeasures have been implemented to deal with this issue. Regarding coverage estimation, the redundancy offered by exploiting all the three LA views and the adoption of a sanity check helps to minimize the issue. Regarding motion detection, a slightly off-axis LA image still correctly represents the cardiac anatomy, and the initial 3D registration step will position it correctly with respect to the SA stack.

VI. Conclusion

In this paper, a fully-automated, learning-based pipeline for quality control of CMR images has been presented. The implemented quality checks are heart coverage estimation, inter-slice motion detection and cardiac image contrast estimation for short-axis image stacks. The pipeline uses hybrid random forests to extract probabilistic segmentation maps and identify landmarks on long- and short-axis images, and then leverages these information to perform the quality checks. It was tested on up to 3000 cases from the UKBB as well as on 100 cases from the UKDHP, and compared to the results of visual or manual analyses to evaluate its accuracy. The results suggest that the proposed approach is able to perform the quality checks with a high accuracy across different datasets. With the recent launch of several initiatives for the acquisition of large-scale CMR datasets, there is a strong need for robust quality control tools in order to facilitate and ensure the reliability of the analyses performed as part of clinical studies. In addition, the low computational time required by the proposed pipeline makes it potentially deployable at the acquisition site, allowing the almost real-time assessment of the scan and the potential triggering of a new acquisition.

References

[1] S. E. Petersen *et al.*, "UK Biobank's cardiovascular magnetic resonance protocol," *J. Cardiovascular Magn. Reson.*, vol. 18, no. 1, p. 8, Jan. 2016.

[2] J. Zhuo and R. P. Gullapalli, "MR artifacts, safety, and quality control," *RadioGraphics*, vol. 26, no. 1, pp. 275–297, Jan. 2006.

[3] P. F. Ferreira, P. D. Gatehouse, R. H. Mohiaddin, and D. N. Firmin, "Cardiovascular magnetic resonance artefacts," *J. Cardiovascular Magn. Reson.*, vol. 15, no. 1, p. 41, May 2013.

[4] V. Klinke *et al.*, "Quality assessment of cardiovascular magnetic resonance in the setting of the European CMR registry: Description and validation of standardized criteria," *J. Cardiovascular Magn. Reson.*, vol. 15, no. 1, p. 55, Jun. 2013.

[5] P. Coupé, J. V. Manjeón, E. Gedamu, D. Arnold, M. Robles, and D. L. Collins, "Robust Rician noise estimation for MR images," *Med. Image Anal.*, vol. 14, no. 4, pp. 483–493, Aug. 2010.

[6] I. I. Maximov, E. Farrher, F. Grinberg, and N. J. Shah, "Spatially variable Rician noise in magnetic resonance imaging," *Med. Image Anal.*, vol. 16, no. 2, pp. 536–548, Feb. 2012.

[7] E. L. Gedamu, D. L. Collins, and D. L. Arnold, "Automated quality control of brain MR images," *J. Magn. Reson. Imag.*, vol. 28, no. 2, pp. 308–319, Aug. 2008.

[8] X. Albà, K. Lekadir, M. Pereañez, P. Medrano-Gracia, A. A. Young, and A. F. Frangi, "Automatic initialization and quality control of large-scale cardiac MRI segmentations," *Med. Image Anal.*, vol. 43, pp. 129–141, Jan. 2018.

[9] L. Zhang *et al.*, "Automated quality assessment of cardiac MR images using convolutional neural networks," in *Simulation Synthesis Medical Imaging—SASHIMI* (Lecture Notes in Computer Science), vol. 9968. Cham, Switzerland: Springer, Oct. 2016, pp. 138–145.

[10] Le Zhang, A. Gooya, and A. F. Frangi, "Semi-supervised assessment of incomplete LV coverage in cardiac MRI using generative adversarial nets," in *Simulation Synthesis Medical Imaging—SASHIMI* (Lecture Notes in Computer Science), vol. 10557. Cham, Switzerland: Springer, Sep. 2017, pp. 138–145.

[11] G. Tarroni *et al.*, "Learning-based heart coverage estimation for short-axis cine cardiac MR images," in *Functional Imaging Modelling Heart—FIMH* (Lecture Notes in Computer Science), vol. 10263. Cham, Switzerland: Springer, Jun. 2017, pp. 73–82.

[12] J. R. McClelland, D. J. Hawkes, T. Schaeffter, and A. P. King, "Respiratory motion models: A review," *Med. Image Anal.*, vol. 17, no. 1, pp. 19–42, Jan. 2013.

[13] J. Lötjönen, M. Pollari, S. Kivistö, and K. Lauerma, "Correction of movement artifacts from 4-D cardiac short- and long-axis MR data," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 3217. Cham, Switzerland: Springer, Sep. 2004, pp. 405–412.

[14] O. Oktay *et al.*, "Respiratory motion correction for 2D cine cardiac MR images using probabilistic edge maps," in *Proc. Comput. Cardiol. Conf. (CinC)*, Mar. 2017, pp. 129–132.

[15] M. Sinclair, W. Bai, E. Puyol-Antón, O. Oktay, D. Rueckert, and A. P. King, "Fully automated segmentation-based respiratory motion correction of multiplanar cardiac magnetic resonance images for large-scale datasets," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 10434. Cham, Switzerland: Springer, Sep. 2017, pp. 332–340.

[16] D. Yang, P. Wu, C. Tan, K. M. Pohl, L. Axel, and D. Metaxas, "3D motion modeling and reconstruction of left ventricle wall in cardiac MRI," in *Functional Imaging and Modelling of the Heart—FIMH* (Lecture Notes in Computer Science), vol. 10263. Cham, Switzerland: Springer, Jun. 2017, pp. 481–492.

[17] K. McLeish, D. L. G. Hill, D. Atkinson, J. M. Blackall, and R. Razavi, "A study of the motion and deformation of the heart due to respiration," *IEEE Trans. Med. Imag.*, vol. 21, no. 9, pp. 1142–1150, Sep. 2002.

[18] O. Dietrich, J. G. Raya, S. B. Reeder, M. F. Reiser, and S. O. Schoenberg, "Measurement of signal-to-noise ratios in MR images: Influence of multichannel coils, parallel imaging, and reconstruction filters," *J. Magn. Reson. Imag.*, vol. 26, no. 2, pp. 375–385, 2007.

[19] S. Aja-Fernández, G. Vegas-Sánchez-Ferrero, and A. Tristán-Vega, "Noise estimation in parallel MRI: GRAPPA and SENSE," *Magn. Reson. Imag.*, vol. 32, no. 3, pp. 281–290, Apr. 2014.

[20] S. D. Wolff and R. S. Balaban, "Assessing contrast on MR images," *Radiology*, vol. 202, no. 1, pp. 25–29, Jan. 1997.

[21] O. Oktay *et al.*, "Stratified decision forests for accurate anatomical landmark localization in cardiac images," *IEEE Trans. Med. Imag.*, vol. 36, no. 1, pp. 332–342, Jan. 2017.

[22] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Found. Trends Comput. Graph. Vis.*, vol. 7, nos. 2–3, pp. 81–227, Mar. 2011.

[23] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2188–2202, Nov. 2011.

[24] P. Dollár and C. L. Zitnick, "Fast edge detection using structured forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1558–1570, Aug. 2015.

[25] A. Criminisi *et al.*, "Regression forests for efficient anatomy detection and localization in computed tomography scans," *Med. Image Anal.*, vol. 17, no. 8, pp. 1293–1303, 2013.

[26] G. Celeux and G. Soromenho, "An entropy criterion for assessing the number of clusters in a mixture model," *J. Classif.*, vol. 13, no. 2, pp. 195–212, Sep. 1996.

[27] W. Bai *et al.*, "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks," *J. Cardiovascular Magn. Reson.*, vol. 20, no. 65, pp. 1–12, Sep. 2018.

[28] L. G. Nyul, J. K. Udupa, and X. Zhang, "New variants of a method of MRI scale standardization," *IEEE Trans. Med. Imag.*, vol. 19, no. 2, pp. 143–150, Feb. 2000.