Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late Gadolinium enhancement MR images[☆]

Rashed Karim[a,*], Pranav Bhagirath[b], Piet Claus[c], R. James Housden[a], Zhong Chen[a], Zahra Karimaghaloo[d], Hyon-Mok Sohn[a], Laura Lara Rodríguez[e], Sergio Vera[e], Xènia Albà[f], Anja Hennemuth[g], Heinz-Otto Peitgen[g], Tal Arbel[d], Miguel A. Gonzàlez Ballester[e,i,j], Alejandro F. Frangi[h], Marco Götte[b], Reza Razavi[a], Tobias Schaeffter[a], Kawal Rhode[a]

[a] *Department of Imaging Sciences & Biomedical Engineering, King's College London, UK*
[b] *Department of Cardiology, Haga Teaching Hospital, The Netherlands*
[c] *Cardiovascular Imaging and Dynamics, Department of Cardiovascular Sciences, Universiteit Leuven, Belgium*
[d] *The Centre for Intelligence Machines, McGill University, Canada*
[e] *Alma IT Systems, Spain*
[f] *Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain*
[g] *Fraunhofer Institute for Medical Image Computing, Fraunhofer MEVIS, Germany*
[h] *Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), Department of Electronic & Electrical Engineering, University of Sheffield, Sheffield, UK*
[i] *ICREA, Spain*
[j] *SIMBIOsys Research Group, Universitat Pompeu Fabra, Spain*

ARTICLE INFO

Article history:
Received 7 May 2015
Revised 12 November 2015
Accepted 15 January 2016
Available online 28 January 2016

Keywords:
Late Gadolinium enhancement
Segmentation
Algorithm benchmarking

ABSTRACT

Studies have demonstrated the feasibility of late Gadolinium enhancement (LGE) cardiovascular magnetic resonance (CMR) imaging for guiding the management of patients with sequelae to myocardial infarction, such as ventricular tachycardia and heart failure. Clinical implementation of these developments necessitates a reproducible and reliable segmentation of the infarcted regions. It is challenging to compare new algorithms for infarct segmentation in the left ventricle (LV) with existing algorithms. Benchmarking datasets with evaluation strategies are much needed to facilitate comparison. This manuscript presents a benchmarking evaluation framework for future algorithms that segment infarct from LGE CMR of the LV. The image database consists of 30 LGE CMR images of both humans and pigs that were acquired from two separate imaging centres. A consensus ground truth was obtained for all data using maximum likelihood estimation.

Six widely-used fixed-thresholding methods and five recently developed algorithms are tested on the benchmarking framework. Results demonstrate that the algorithms have better overlap with the consensus ground truth than most of the *n*-SD fixed-thresholding methods, with the exception of the Full-Width-at-Half-Maximum (FWHM) fixed-thresholding method. Some of the pitfalls of fixed thresholding methods are demonstrated in this work. The benchmarking evaluation framework, which is a contribution of this work, can be used to test and benchmark future algorithms that detect and quantify infarct in LGE CMR images of the LV. The datasets, ground truth and evaluation code have been made publicly available through the website: https://www.cardiacatlas.org/web/guest/challenges.

1. Introduction

In recent years, the translation of image analysis tools to the clinical environment has remained limited despite their rapid development. Although algorithms are extensively validated in-house following development, it is often not clear how they compare

[☆] This paper was recommended for publication by James Duncan.

[*] Corresponding author. Tel.: +44 207 188 5437.
E-mail address: rashed.karim@kcl.ac.uk (R. Karim).

http://dx.doi.org/10.1016/j.media.2016.01.004

Table 1
Overview of previously published methods for scar quantification and segmentation.

	Reference	Model	*n*	Algorithm	Highlight
LV	Kim et al. (1999)	Canine	26	2-SD	Correlation of MRI enhancement with scar
	Amado et al. (2004)	Animal	13	1–6 SD, FWHM	FWHM correlates to histology
	Kolipaka et al. (2005)	Human	23	2,3-SD	Manual correction is necessary despite algorithm
	Positano et al. (2005)	Human	15	Clustering	Fast clustering algorithm
	Schmidt et al. (2007)	Human	47	2–6 SD	Grey-zone and core quantification
	Hennemuth et al. (2008)	Human	21	EM fitting*	Model based on scanner acquisition and reconstruction parameters
	Detsky et al. (2009)	Human	15	Clustering*	Clustering in feature space
	Tao et al. (2010)	Human	20	Otsu thresholding*	Dice overlap on chronic myocardial infarction with 2-observer manual segmentation
	Flett et al. (2011)	Human	60	2–6 SD, FWHM	Inter- and intra-observer reproducibility
	Rajchl et al. (2014)	Human	35	SD, FWHM, Max-flow	Inter- and intra-observer reproducibility on 3D CMR
	Andreu et al. (2011)	Human	12	50, 60, 70% FWHM	60% FWHM for good voltage correlation
	Lu et al. (2012)	Human	10	Graph-cuts*	Correlation with FWHM and manual segmentations on chronic myocardial infarction data
	Pop et al. (2013)	Animal	9	Mixture model	*ex-vivo* histology and high-resolution MRI
LA	Oakes et al. (2009)	Human	81	2–4 SD	LA fibrosis and correlation to recurrence
	Knowles et al. (2010)	Human	7	Maximum intensity projection	Necrosis and oedema theory for reconnection, comparison with electroanatomical data
	Ravanelli et al. (2014)	Human	10	SD, Skeletonisation*	Comparison with electroanatomical data
	Karim et al. (2014)	Human	15	Graph-cuts*	Dice with 3-observer consensus delineation
	Harrison et al. (2014)	Animal	16	2–6 SD	*ex-vivo* histology infarct volume against MR

Methods are listed in chronological order, type of data they were evaluated with and the algorithm for: left ventricle (LV) or left atrium (LA). Methods which report on a segmentation algorithm developed are marked with an asterix (*).

to other existing algorithms. Algorithm designers are faced with the challenging task of cross comparing their algorithm's performance. The absence of a common pool of data along with evaluation strategies has limited algorithm translation into the clinical workflow Moreover, as larger cohort data sets become available, the need for reducing the manual labour involved in image analysis is becoming more important

Benchmarking of algorithms on common datasets provides a fair test-bed for comparison. It is thus a very important activity as we move from bench to the bedside in the medical image processing community. In recent years, several conferences and meetings within the medical image processing community have provided a platform to benchmark algorithms from multiple research groups. These *challenges* invite participants to submit their algorithms and test them on common data. The results from the test are then evaluated and compared using common evaluation metrics. In the past, a few challenges have been organised, each with its own unique theme. There exists an index of past challenges within the medical image processing community and it can be found on the Cardiac Atlas project page in https://www.cardiacatlas.org/web/guest/challenges. In the cardiovascular imaging domain, some recent challenges include left atrial fibrosis and scar segmentation (Karim et al., 2013), left ventricle segmentation (Suinesiaputra et al., 2014), right ventricle segmentation (Petitjean et al., 2015), cardiac motion tracking (Tobon-Gomez et al., 2013) and coronary artery stenosis detection (Kirisli et al., 2013).

1.1. Motivation for left ventricle infarct segmentation

Cardiovascular magnetic resonance (CMR) imaging can be used to comprehensively assess the viability of myocardium in patients with ischaemic heart disease. Myocardial infarction can be visualised and quantified using inversion recovery imaging 10–15 min after intravenous administration of Gadolinium contrast. This imaging technique is known as late Gadolinium enhancement (LGE) imaging. Experimental models have shown excellent agreement between size and shape in LGE CMR and areas of myocardial infarction by histopathology (Kim et al., 1999; Wagner et al., 2003). Infarct size from CMR is also a primary endpoint in many clinical trials (see Desch et al., 2011 for a complete list).

Recent studies have also demonstrated how infarct size, shape and location from pre-procedural LGE can be useful in guiding ventricular tachycardias (VT) ablation (Estner et al., 2011; Andreu et al., 2011). These procedures are often time-consuming due to the preceding electrophysiological mapping study required to identify slow conduction zone involved in re-entry circuits. Post-processed LGE images provide scar maps, which can be integrated with electroanatomic mapping systems to facilitate these procedures (Andreu et al., 2011). Clinical implementation of these developments necessitates a reliable, fast, reproducible and accurate segmentation of the infarcted region. Moreover, as use of LGE-based infarct volume estimation becomes more clinically relevant, standardisation will facilitate more consistent interpretation.

1.2. State-of-the-art for cardiac infarct segmentation

A short overview of previously published infarct detection algorithms for the left ventricle (LV) is presented here. Table 1 lists the algorithms surveyed and highlights some of their important features. A common method for detecting infarct in the LV is the fixed-model approach, whereby intensities are thresholded to a fixed number of standard deviations (SD) from the mean intensity of nulled myocardium or blood pool (Flett et al., 2011). In the rest of the paper this will be known as the n-SD method, where $n = 2, 3, 4, 5$ or 6. A second common fixed-model approach is the full-width-at-half-maximum (FWHM) approach, where half of the maximum intensity within a user-selected hyper-enhanced region is selected as the fixed intensity threshold (Amado et al., 2004). Using this threshold, a region-growing process is employed from user-selected seeds. These seeds are selected to be within infarcted regions such that they can be segmented with region-growing.

As the aforementioned approaches require user input, making them prone to inter- and intra- observer variation, other approaches that are automatic have been developed. Hennemuth et al. (2008) modelled the intensities of homogeneous tissue in LGE CMR with a Rician distributions and an expectation-maximization (EM) algorithm was used for fitting the data. Pop et al. (2013) fitted Gaussian mixture models to myocardial tissue pixel intensities and correlated with histology. In Detsky et al. (2009), clustering in a feature space of steady-state and T_1^* intensity values provided the segmentation which was shown to provide good correlation with FWHM. Tao et al. (2010) employed automatic thresholding using the Otsu method on bi-modal intensity histograms of myocardium and blood pool. More recently, the use of the graph-cut technique in image processing has been applied to segment infarct in several methods (Lu et al., 2012; Karim et al., 2014; Karimaghaloo et al., 2012). An advantage of this technique is that constraints can be placed on the resulting segmentation,

Table 2
Image acquisition: image acquisition parameters for the challenge LGE patient and porcine datasets. Abbreviations: TI - Inversion time, TR - Repetition time, TE - Echo time, FA - Flip angle, ECG - Electrocardiogram. Imaging centres: KCL-IM - Imaging Sciences, King's College London and UL - Universiteit Leuven. Note that the patient dataset was acquired at KCL-IM and porcine dataset was acquired at UL.

	KCL-IM	**UL**
Scanner type	Philips Achieva 1.5T	Siemens Trio 3.0T
Sequence	Segmented 2D, inversion recovery gradient echo ECG triggered, breath-hold	Segmented 3D inversion recovery, gradient echo ECG triggered breath-hold
TI, TR, TE, FA	280 ms, 3.4 ms, 2.0 ms, 25°	340-370 ms, 2.19 ms, 0.78 ms, 15°
Resolution	$1.8 \times 1.8 \times 8$ mm	$1.8 \times 1.8 \times 6$ mm
Interleaving	Every R-R interval in ECG	Every other R-R interval in ECG
Subjects	Human	Porcine

allowing segmentation boundary regularization with region-based properties. It also predicts which pixels are statistically most likely to be infarct based on prior probability distribution models.

1.3. Proposed evaluation framework

In this paper we propose an evaluation framework for future algorithms that segment and quantify infarct from LGE CMR images of the LV. To demonstrate the framework, five algorithms were evaluated by comparing against a consensus segmentation of experienced observers. The algorithm and observers were both provided the myocardium segmentation. The algorithms were also provided with training data sets. Algorithms evaluated in this work were submitted as a response to the open challenge, put forth to the medical imaging community at the Medical Image Computing and Computer Assisted Intervention (MICCAI) annual meeting's workshop entitled as Delayed Enhancement MRI segmentation challenge. There were thirty LGE CMR data of the LV from both human and porcine cohorts used for the challenge. The data were divided into test ($n = 20$) and training ($n = 10$) sets. Each participant designed and implemented an algorithm which segmented the infarct in each dataset. The datasets are publicly available via the Cardiac Atlas project challenge website https://www.cardiacatlas.org/web/guest/ventricular-infarction-challenge.

2. Material and methods

2.1. Data acquisition database

LGE images were collected from two imaging centres: Imaging Sciences at King's College London (KCL-IM) and Universiteit Leuven (UL). A total of fifteen human and fifteen porcine datasets were collected, of which five in each cohort were used as a training set for the algorithms. For all datasets, a short-axis stack of DE-MRI images covering the LV were provided. The myocardial mask in each image was made available. This was delineated carefully by an expert observer using short-axis slices. A first step was to determine the basal, mid and apical slices based on the standard American Heart Association (AHA) guidelines (Cerqueira et al., 2002). The contours for epicardial and endocardial borders, excluding the papillary muscles, were carefully drawn on each slice before the enclosed region in-between them was filled to produce the mask. The images in the database were limited to the above two different types but varied in their quality. Refer to Table 2 for a summary of the two different types of data that were included in this study.

The human data ($n = 15$) were from randomly selected patients who had a known history of ischaemic cardiomyopathy and were under assessment for an implanatable cardioverter defibrillator (ICD) device for primary or secondary prevention after infarction. In addition to this, the patients chosen had a history of myocardial infarction at least three months prior to their MRI scan. There was also evidence of significant coronary artery disease on angiography and evidence of left ventricular impaired systolic function on echocardiography. The images were acquired on a clinical 1.5T MRI unit (Philips Achieva, The Netherlands). All patients gave written informed consent.

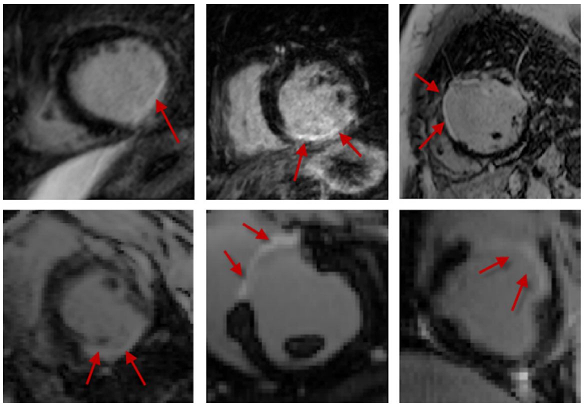

Fig. 1. Sample datasets: a sample of LGE CMR data included in the challenge. The human (top-row) and porcine (bottom-row) images are shown.

The porcine data ($n = 15$) were randomly selected from an experimental database of a pre-clinical model of chronic myocardial ischemia (Wu et al., 2011), with induced lesions obtained by occluding either the left-anterior descending or left-circumflex artery. The data were acquired six weeks after the induction of the coronary lesion on a clinical 3T MRI unit (Siemens Healthcare, Germany). Representative images are shown in Fig. 1.

Five research groups segmented the above datasets, leaving ten images aside, which were utilised for training. A brief summary of their algorithms is given in Table 3. They are described in greater detail in the sections below with a brief background on each technique implemented and details of their implementation.

2.2. Algorithm 1: Alma IT Systems - support vector machines and level sets (AIT)

2.2.1. Background:

Support vector machines (SVM) and level set methods were used to segment scar in this method. SVM is a machine learning technique which first computes the optimal hyperplane on a set of training data mapped to some feature space (Hearst et al., 1998). The hyperplane is a decision boundary which maximally separates the pre-labelled data. Once the hyperplane is obtained, the unseen data is mapped to the same feature space to see which side of the hyperplane it lies in. This labels and thus classifies the unseen data. Level-sets (Sethian, 1999) were also used in this method. In this technique a region evolves from an initial position within the region to be segmented. Level-sets have the added advantage of imposing shape constraints on the evolving region.

2.2.2. Implementation:

A number of image processing techniques were employed. In the first stage, an Otsu-based thresholding was used. Here the threshold between healthy and scar tissue was computed by

Table 3
A brief summary of algorithms that were evaluated on the proposed framework. Institution abbreviations: AIT - Alma IT Systems, and - Universitat Pompeu Fabra, MCG - McGill University, MV - Mevis Fraunhofer, KCL - King's College London.

Algorithm	Technique	Strengths and weaknesses	Key features	Interaction
AIT: Lara et al.	Otsu, support vector machines and level-sets	Post-processing improves results but increases running time	Otsu with two tissue classes. User selects seed in blood-pool	Semi-automatic
UPF: Albà et al.	Region-growing and morphology	Shapes uncharacteristic of scar are deleted but requires initialisation for every slice	Two seeds, for healthy and scar, per slice. Region-labelling step ensures smoothness, filling gaps	Semi-automatic
MCG: Karimaghaloo et al.	Conditional random fields	Hierarchical approach with two levels of processing, but uses statistics on a small neighbourhood	Posterior distribution model estimated with a direct map and not Gaussian during training	Automatic
MV: Hennemuth et al.	EM-algorithm and Watershed transformation	No fixed intensity model and the best-fit model is selected, but over-fitting can be an issue	Automated seed-selection in watershed process. Gaussian-mixture or Rician–Gaussian models for fitting intensities with EM algorithm	Semi-automatic
KCL: Karim et al.	Graph-cuts with EM-algorithm	Computes a globally optimal segmentation, but can sometimes reject good candidates	Gaussian-mixture model fits intensities with EM algorithm using three tissue classes	Semi-automatic
n-**SD**	*n* standard deviations from healthy tissue ($n = 2, 3, 4, 5, 6$)	Simple to implement, but baseline is subjective	Only involves thresholding, no region-growing as FWHM	Semi-automatic
FWHM	50% of user-selected hyper-enhanced myocardium	Validated with histology in literature but was first used to describe a phenomenon in signal analysis	Computed threshold used for region-growing from user-selected seed locations in each slice	Semi-automatic

maximising the intensity variance between the two labels in the intensity histogram (Otsu, 1975). However, as this method was subject to limitations, especially in instances where healthy and scar tissues had overlapping intensities, further steps were necessary. An ensuing connected-component analysis found groups of connected pixels. On these pixel groups, several features relevant to scar were extracted: area, bounding box, major and minor axes, eccentricity, convex-hull area and Euler number (Teague, 1980). This allowed pixel groups to be mapped to a feature space. Several classifiers were tested on the training data provided. These were namely SVM, *K*-nearest neighbours, linear Bayesian discriminant, and linear perceptron classifiers. SVM was chosen based on the best trade-off between error and sensitivity on the training data (Hearst et al., 1998).

Following classification using SVM, a further level-set-method step refined the segmentations obtained (Sethian, 1999). The contours obtained from the SVM classification step were used to initialise a level-set. The level-set was constrained by the search area obtained in the initial step of the algorithm. It evolved in a *speed image* $P(x)$ derived from the SVM classified pixels:

$$P(x) = \begin{cases} I(x) - L, & \text{if } I(x) < \frac{U+L}{2} \\ U - I(x), & \text{otherwise} \end{cases} \tag{1}$$

The values U and L were obtained from grey-level intensity $I(x)$ statistics of the SVM output, i.e. $U = \mu + 5\sigma$ and $L = \mu - 5\sigma$. These parameters are in-line with the standard deviation approach for classifying scar (Karim et al., 2013).

2.3. Algorithm 2: Universitat Pompeu Fabra - Region growing and morphology (UPF)

2.3.1. Background:

Region-growing is a well-known image processing technique which finds a group of connected pixels with intensity homogeneity. It is an iterative process which starts from a seed point, and the region increases in size by including neighbouring pixels that fit a certain pre-defined criteria. Region-growing can subsequently *leak* into neighbouring areas, which is an important limitation of the technique.

2.3.2. Implementation:

Seed selection for region-growing was automatic and repeated for each slice, making it essentially a 2D technique. A minimum of two seeds were selected for each tissue class: scar and healthy. The criteria for selecting seeds for the scar tissue class was the following:

$$I > \mu_k + 2\sigma_k \tag{2}$$

where a pixel in the kth slice has intensity I and is subjected to the above test based on mean (μ) and variance (σ^2) of myocardium intensity. Individual regions satisfying the above criteria were analysed for their shape and size. Elongated and thin regions near the epicardium were deleted in an automated manner by computing the eccentricity and width (proportion to myocardial mask) of the region in question, on which a thresholding was performed based on empirical values obtained from the training set. The size of negligible regions were defined in proportion to the pixel size and size of the myocardial mask. The two largest and brightest regions were selected as the seeds. This selected seeds for the scar tissue class. For the healthy tissue class, a similar standard deviation approach was utilised (i.e. $I < \mu_k + 2\sigma_k$) and the two largest and *darkest* regions were selected as seeds. Region-growing was initiated from each seed region and these generated segmented regions for healthy or scar tissue classes. The choice of two seeds, per slice, for each tissue class is important as it generates two separate disconnected regions. However, this places a limit on the maximum number of scar or healthy regions possible (i.e. two) in each slice.

The region-growing process was followed by a region-labelling step in which pixels that were not labelled as scar or healthy tissue were analysed; if they contained any adjacent neighbour belonging to either scar and healthy classes, they were labelled as such. This was followed by a post-processing step to fill holes or small gaps in the segmentations. Also, regions that were small *islands* containing a negligible number of pixels were removed from the segmentation. Finally, dark regions that lacked contrast, but were surrounded by scar pixels were re-labelled as scar. This is characteristic of a microvascular obstruction.

2.4. Algorithm 3: McGill - conditional random fields (MCG)

2.4.1. Background:

The previous methods described are geometrical in their nature; a region's intensity and its geometrical shape are used to determine its classification. The method described in this section is different from the above approaches in that a probabilistic classifier model was used. Based on the training dataset, the classifier can infer the posterior distribution of a pixel's label to be healthy or scar given the observation. There are two sets of observations made: (1) the pixel's intensity, and (2) the pixel's neighbourhood. Since labels of neighbouring pixels are typically correlated, neighbourhood information is incorporated by building a graphical model $G(V, E)$, where voxels are represented by a set of nodes (V) and the relationships among them are represented by edges (E). In the *generative* Markov random field (MRF) (see Boykov et al., 2001), the Bayes' relationship is used to determine the posterior distribution:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \tag{3}$$

where X is the unseen image to be segmented and Y is the labelling into healthy and scar. The likelihood $p(X|Y)$ of the unseen image is estimated by assuming that the voxel intensities in X are independent given the labels. Also, a uni-modal Gaussian is often used. However, in the context of medical image segmentation, regions are not random collections of independent pixels. Instead, structures usually form coherent and continuous shapes. In this work, a conditional Markov random field (CRF) (Lafferty et al., 2001) is used which is a *discriminative* framework and the posterior $p(Y|X)$ is estimated by learning a direct map from observations to the class labels (i.e. in training images). This is how it differs from other MRF approaches used in binary classification, where the posterior is estimated using Gaussian distributions.

2.4.2. Implementation:

The CRF implemented in this work used a hierarchical approach and is described in Karimaghaloo et al. (2012). There are two levels of CRF: in the first level image intensity information was used, and in the second level, a so-called spin image feature vector derived from intensity information was used. In the first level CRF, the posterior distribution $p(Y|X)$ was estimated as in a conventional CRF (Lafferty et al., 2001):

$$p(Y|X) = \frac{1}{Z}\exp\left[\sum_{i=1}^{n}\phi(y_i|X) + \sum_{j\in N_i}\varphi(y_i, y_j|X) + \sum_{j,k\in N_i}\psi(y_i, y_j, y_k|X)\right] \tag{4}$$

where Z is a normalization term and ϕ, φ and ψ are *unary, pairwise and triplet* potentials respectively. Pairwise and triplet potentials measure the interaction between pixels that are immediate neighbours (pairwise) and neighbour's neighbours (triplet). As regions in MRI images are not random collections of independent pixels but part of coherent and continuous shapes, the pairwise and triplet potentials reinforce this notion. The unary potentials $p(y_i|\mathbf{x}_i)$ computed the inference on the healthy or scar labels (y_i) from the MRI intensity observed at pixel i. This potential was modelled from labelled training data provided within the challenge using:

$$\phi(y_i|X) = \log\ p(y_i|\mathbf{x}_i) \tag{5}$$

where y_i is the label and $\mathbf{x}_i$ is the observed intensity at voxel i. A binary classifier was employed for the purpose of distinguishing between healthy and scar. The decision boundary was learned from training data using a variant of support vector machines (SVM) known as relevance vector machines (RVM) (Tipping, 2001). The final classification of the first-level CRF was performed using a graph-cut optimization framework (Boykov et al., 2001).

In the second-level CRF, using infarction candidates from the first level, a two dimensional histogram encoding the distribution of image brightness values in the neighbourhood of a particular reference point was constructed. This is the spin image which encoded local information around infarct candidates. Besides voxel intensity, these spin image features were also used for CRF. Similar to the first-level CRF, the final inference was performed using a graph-cut optimisation framework.

2.5. Algorithm 4: Mevis Fraunhofer - EM-algorithm and watershed transformation (MV)

2.5.1. Background:

The method presented in this work assumes that the voxel intensity distribution in MR images can be modelled using statistical distribution models. Depending on acquisition parameters and the reconstruction algorithm, it can either be modelled using a Gaussian, Rayleigh or non-central χ-distribution (Dietrich et al., 2007). These distributions are also closely related to the Rician distribution, making it suitable for modelling healthy myocardium intensities. For diseased myocardium the Rician–Gaussian mixture was found to be appropriate, and for necrotic tissues, the non-central χ-distribution was shown to be suitable (Hennemuth et al., 2008).

The watershed segmentation approach was used in this method (Hennemuth et al., 2008). Watershed is a classical image segmentation technique where the gradient image is considered as a topographic surface. Structures such as scar can be assumed to have high intensity gradients at edges and low gradients in the interior. This high-low-high intensity gradient profile creates basins in the image. Once points are located inside each basin they can be segmented by following paths of decreasing altitudes on the topography of the gradient image.

2.5.2. Implementation:

In this work, three separate models were considered: Rician, Rician–Gaussian and Gaussian models. Each model was fitted to the myocardium intensity distribution in the unseen image. The model with the least mean fitting error was chosen. To achieve an optimal fit, the Expectation-Maximization (EM) algorithm was used. Two classes corresponding to healthy and scar were chosen to initialise the EM fit.

A threshold was then derived from the mixture distribution obtained from the EM-fitting process. This is the higher of the two means in the two-class mixture model. Using Euclidean distance in 3D and endocardial voxels computed from the myocardium segmenatation, voxels with intensity higher than the threshold and closer to the endocardium were chosen as seeds for the watershed process. These seeds were used to define the basins and the watershed transformation determined the extent of each basin. The basins determined each location to be labelled as scar. An ensuing connected-components analysis step removed small noisy structures.

2.6. Algorithm 5: KCL - Graph-cuts with EM-algorithm (KCL)

2.6.1. Background:

The background of the method used in this work is in some ways similar to the method proposed by MCG in Section 2.4 except that it employs a non-conditional MRF solved using graph-cuts. The image to be segmented is modelled as a graph with paths or links between neighbouring pixels. For each pixel there is also a link to two special nodes also known as source and sink nodes that correspond to scar and healthy myocardium. Each link is assigned a weight based on its intensity. The graph-cuts approach computes a partitioning to divide the graph into two sub-graphs, one containing the source node and the other the sink node. This partitioning assigns a label (source or sink) to each pixel solving the segmentation as an optimisation problem. It searches for a globally-optimal solution.

2.6.2. Implementation:

In the graph-cuts approach implemented in this work, each pixel in the myocardium was modelled as a node in the graph with links to source and sink nodes. These links were assigned weights representing the affinity to healthy (i.e. source) and scar (i.e. sink) nodes. The weights were derived from statistical distribution models developed from training images. There were separate intensity distribution models for healthy and scar tissue, both of which were derived from the training images. For scar, the ratio of delayed enhancement intensity to mean blood pool was modelled using a Gaussian distribution. For healthy tissue, a Gaussian mixture was used. The number of mixtures in the model was fixed at three. The standard EM-algorithm computed mean and variance for each mixture from the training images. In the graph-cuts framework there are also links between adjacent pixels and these were derived from a measure of intensity similarity of two pixels. Adjacent pixels with similar intensities attained a high weight. This enforced coherence in the segmentation output. The final segmentation was obtained using global optimisation over the entire image. This allowed for disjointed infarct regions to be identified in the image.

2.7. Algorithm evaluation

2.7.1. Reference standard: consensus ground truth

A reference standard for scar in each case was obtained by combining volumetric segmentations from three separate observers. All observers were cardiologists with several years' experience in CMR assessment of LV function and tissue viability. They also had several years' experience working with patients suffering from ischaemic heart diseases. For both datasets, they were blinded to the underlying clinical situation of patients and pigs. For pigs, lesions were obtained by occluding either the left-anterior descending or left-circumflex artery, and the observers were blinded to this fact. The observers were *not* instructed to look for areas of *grey zones*. For regions affected by microvascular obstructions, they were instructed to avoid these by looking for regions of significant hypo-enhancement surrounded by enhanced regions.

Scars in the images were segmented as follows: (1) Each slice in the LGE CMR was analysed separately in the short-axis view. The segmentation of the myocardium was loaded as an overlay. (2) The basal, mid and apical slices were identified along with the LV orientation, i.e. the posterior and anterior ends. (3) The short-axis slices were then analysed one at a time sequentially from basal to apical or apical to basal. (4) The basal slices were then examined for non-scar related enhancements (see Turkbey et al., 2012) such as the right ventricle (RV) insertion point, and partial voluming in the basal slices due to the outflow tract and appendage. The mid and apical slices were also examined for coronary arteries carrying blood that could be enhanced, and microvascular obstructions. (5) Pixels enhanced within myocardium were labelled as scar and generally noisy pixels or regions were avoided. Noise observed in the lungs was used as a reference.

Each observer was provided with the same set of guidelines as above. However, their segmentations differed in some instances. This was generally due to differences in their opinion and experience. Such inter-observer variability is now widely accepted. It was thus important to merge the segmentations and obtain a consensus ground truth. A maximum likelihood estimation of ground truth was obtained using a published algorithm known as the STAPLE (Warfield et al., 2004). For every voxel, a probabilistic estimate of the true segmentation was computed using an optimal combination of the observers' segmentations. The final consensus segmentation was then obtained by thresholding this probability above 0.7 or 70%. This is referred to in the rest of the text as the *consensus* ground truth.

2.7.2. Common algorithms: n-SD and FWHM

Quantification of scar in LGE CMR images using a fixed model is often desirable and commonly used as it includes fewer image processing steps, with some studies advocating its reproducibility (Flett et al., 2011; Amado et al., 2004). In fixed models, scar is quantified by thresholding intensities at a fixed distance from a reference intensity value. Two types of fixed models were used, namely FWHM and the *n*-SD method. FWHM is a technique where half of the maximum intensity within a user-selected hyper-enhanced region is selected as the fixed intensity threshold for an ensuing region-growing step (Amado et al., 2004). In the region-growing step, infarcted regions are segmented based on user-selected seed points. These are used to initialise the region-growing step. The *n*-SD method (where $n = 2, 3, 4, 5, 6$), uses a fixed number of standard deviations from mean signal within healthy myocardium. A manual region-of-interest (ROI) selection was required in both techniques. In FWHM, a ROI was delineated in hyper-intense myocardium. In *n*-SD, a ROI was delineated in remote myocardium. Remote myocardium was defined as a region with no enhancement and normal wall motion. Endocardial and epicardial surfaces were avoided in the delineation.

2.7.3. Evaluation metrics

Segmentations from each algorithm were compared against the reference standard for scar. As no single metric is advocated as the best metric, two different types of metric were chosen for evaluating the segmentations. These were overlap and volumetric measures, and they are briefly described below:

1. Overlap metric: The Dice similarity is a metric for segmentation overlap measuring the proportion of true positives in the segmentation:

$$s = \frac{2|X \cap Y|}{|X| + |Y|} \quad (6)$$

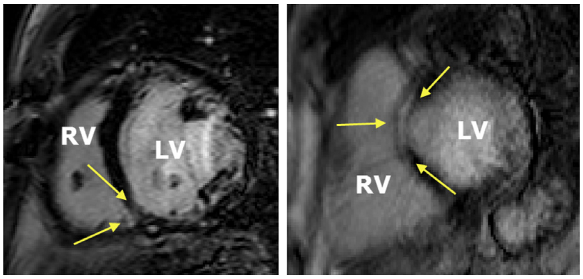

Fig. 2. Examples of pseudo infarct in the patient database. Arrows indicate enhancements due to the right ventricle insertion point (left) and outflow tract (right).

where X is the segmented region in the ground-truth and Y is the region in the challenger's algorithm.

2. Volumetric-based metric: The total volume error between the algorithm's output and reference standard was found:

$$\delta V = |V_T - V_G| \tag{7}$$

where V_T is the volume of scar in the algorithm segmentation and V_G is the volume of scar in the consensus segmentation.

2.7.4. Objective evaluation

In LGE CMR of the LV, hyper-enhanced areas not relating to scar are not uncommon (Turkbey et al., 2012). Unless the characteristic and geometry of these *pseudo infarcts* are explicitly modelled into the technique, it is challenging for an algorithm to distinguish them. Some common sources of pseudo infarcts seen in LGE CMR of the LV are: (1) the location of the RV insertion point, (2) partial voluming in basal slices due to the outflow tract and the appendage, and (3) hyper-enhanced areas due to epi- and pericardial fat. An experienced observer selected regions containing the aforementioned enhancements. These were identified using simple techniques such as checking for continuity of scar or artefact in the adjacent slices, i.e. if it continues then it is likely to be scar. Some instances of pseudo infarcts occurring in the patient dataset are shown in Fig. 2. To evaluate how the algorithms handled pseudo infarcts, each algorithm's output was evaluated separately on these regions. The percentage of voxels detected by each method in these spurious regions was determined.

A good contrast between normal myocardium, blood pool and infarct is challenging and greatly depends on achieving the optimal inversion time. Each scan in the image database was scored by five raters experienced in LGE CMR images. The rating with maximum votes determined the scan's rating. Scans in the image database were ranked into three categories: good, average and poor. The Dice metric was computed separately in each category. This indicated how robust the algorithms were against contrast enhancement quality.

3. Results and discussions

3.1. Segmentation accuracy against consensus ground truth

On the patient and porcine LGE CMR scans, segmentations from the algorithms were compared to the consensus ground truth. A consensus was available by combining segmentations from three separate observers as described in Section 2.7.1. Segmentation accuracies measured using the Dice metric are shown in Fig. 3 for the patient dataset. The Dice overlaps between algorithm and consensus were determined on an automatically-determined region-of-interest (ROI) enclosing each individual region of infarction labelled in the consensus. The medians of these individual Dice overlaps were as follows: AIT= 73, KCL= 74, MCG= 85, MV= 44, and UPF= 70. Fixed model approaches for segmenting scar (i.e. n-SD and FWHM) were also compared with the consensus ground-truth. The median Dice overlaps were: 2-SD= 47, 3-SD= 54, 4-SD= 55, 5-SD= 62, 6-SD= 64, FWHM= 78. An example of a single slice from the patient dataset is shown in Fig. 5.

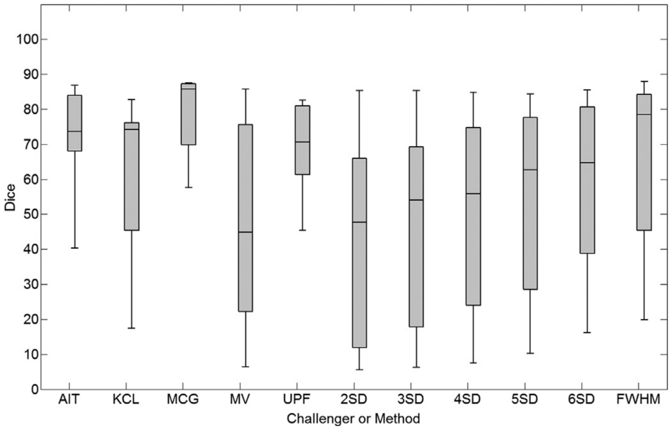

Fig. 3. Performance on patient datasets: segmentation accuracy on the patient dataset. Note the figure also displays results from 2-SD, 3-SD, 4-SD, 5-SD, 6-SD and FWHM. Dice was computed on every individual region of scar found in the consensus segmentation.

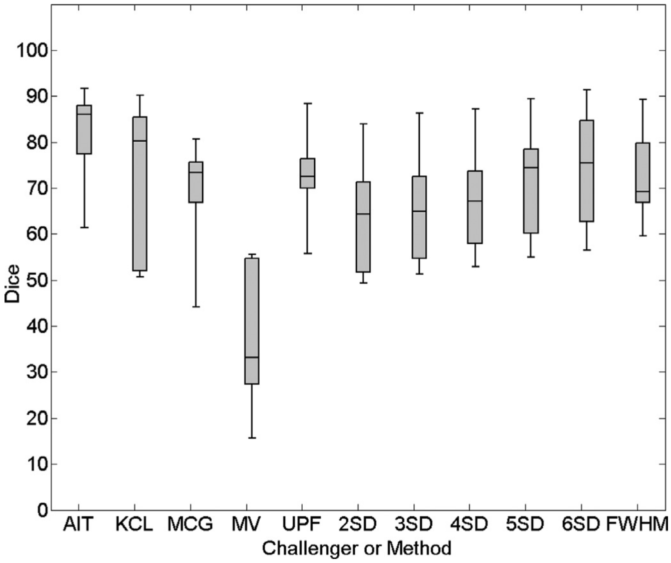

Fig. 4. Performance on porcine datasets: segmentation accuracy on the porcine dataset. Note the figure also displays results from 2-SD, 3-SD, 4-SD, 5-SD, 6-SD and FWHM. Dice was computed on every individual region of scar found in the consensus segmentation.

On the porcine LGE CMR scans segmentations from the algorithms and fixed-model approaches were compared in a similar way to the patient dataset. The Dice overlap metric is plotted in Fig. 4 for each submitted algorithm and fixed model. The Dice overlaps were determined, as above, on ROIs enclosing each region of infarction labelled in the consensus. The medians of these individual Dice overlaps were as follows: AIT= 86, KCL= 80, MCG= 73, MV= 33, and UPF= 73. Standard methods using fixed models were also compared with the consensus ground-truth and the median Dice overlaps were: 2-SD= 64, 3-SD= 65, 4-SD= 67, 5-SD= 74, 6-SD= 76, FWHM= 69. An example of a single slice from the porcine dataset is given in Fig. 6.

The Dice scores, reported above, were evaluated within ROIs enclosing scar in the consensus segmentation. These areas can often be large sections within the image, especially if the scar is continuous and extends to several slices. This provided for a more

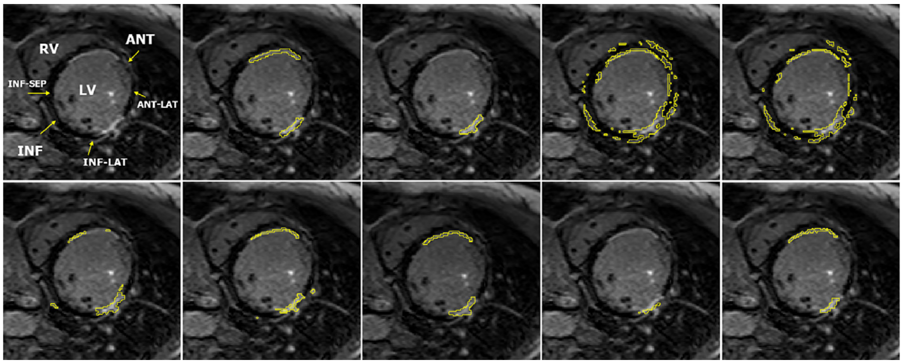

Fig. 5. Example segmentation from the patient dataset. Clockwise from top-left: original LGE CMR, consensus segmentation, FWHM, 5-SD, 6-SD, AIT, KCL, MCG, MV, UPF. Abbreviations: LV - left ventricle, RV - right ventricle, ANT - anterior, INF - inferior, INF-SEP - infero-septal, INF-LAT - infero-lateral, ANT-LAT - antero-lateral.

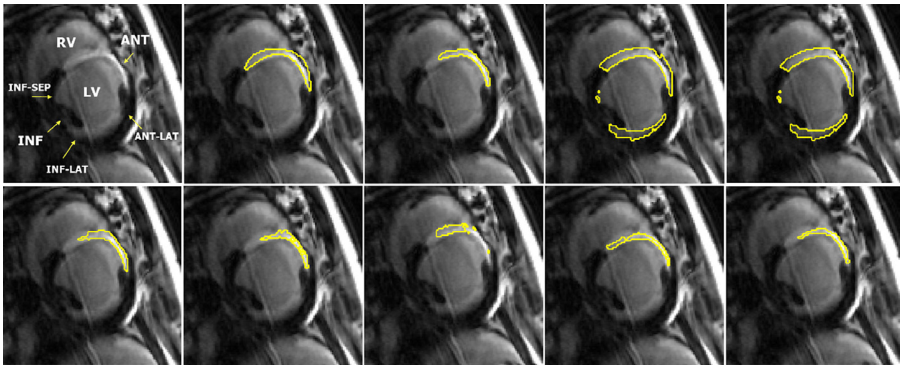

Fig. 6. Example segmentation from the porcine dataset. Clockwise from top-left: original LGE CMR, consensus segmentation, FWHM, 5-SD, 6-SD, AIT, KCL, MCG, MV, UPF. Abbreviations: LV - left ventricle, RV - right ventricle, ANT - anterior, INF - inferior, INF-SEP - infero-septal, INF-LAT - infero-lateral, ANT-LAT - antero-lateral.

Table 4
Segmentation accuracy with volume difference (δV) on patient and porcine data for submitted algorithms and fixed-models. The standard deviation of each metric is quoted in brackets.

	Patient data $\|\delta V\|$ (ml)	Porcine data $\|\delta V\|$ (ml)
AIT	0.77 (0.7)	0.84 (0.5)
KCL	1.05 (1.0)	0.73 (0.5)
MCG	1.02 (0.5)	0.54 (0.1)
MV	1.70 (2.3)	0.75 (0.3)
UPF	0.70 (0.3)	0.97 (0.7)
2-SD	8.55 (0.4)	4.00 (0.2)
3-SD	6.71 (0.3)	3.52 (0.8)
4-SD	5.20 (0.2)	2.92 (0.8)
5-SD	3.92 (0.3)	2.44 (0.1)
6-SD	2.96 (0.3)	2.08 (0.1)
FWHM	3.10 (1.0)	2.20 (0.2)

objective evaluation. The algorithm's false positives outside the ROI is not accountable. To counteract this issue, segmentations were also compared by quantifying volume differences. This was determined by measuring the difference in total volume of scar between the consensus and algorithm segmentation. An algorithm could be deemed as accurate only when it yielded a good Dice together with a small volume difference. Table 4 lists the mean volume differences and variance (as millilitres) over the entire image database for patient and porcine datasets.

To further evaluate more objectively, the Dice overlap of the algorithms' segmentations were compared to the consensus based on the slice position (basal, mid and apical. Short-axis slices were subdivided according to the standard guidelines (Cerqueira et al., 2002). The results are plotted in Fig. 7. It is not clear what should be a good Dice overlap for datasets of this type. To address this issue, the degree of agreement between observers and the computed consensus was analysed and plotted in Fig. 8. It provided for an estimation of a reasonable target (i.e. good Dice score) for the evaluated algorithms.

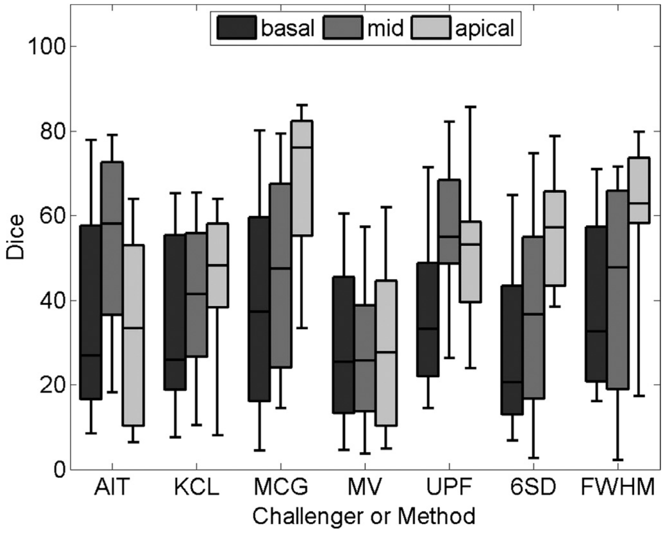

Fig. 7. Plot showing the characterization of Dice by slice location (basal, mid and apical) by combining results from the patient and porcine datasets.

3.2. Pseudo infarct regions

The algorithms were evaluated on hyper-enhanced regions which mimic scar. These pseudo infarct regions occur for several reasons mentioned in Section 2.7.4 and illustrated in Fig. 2. In each image, pseudo infarct was manually segmented by an experienced

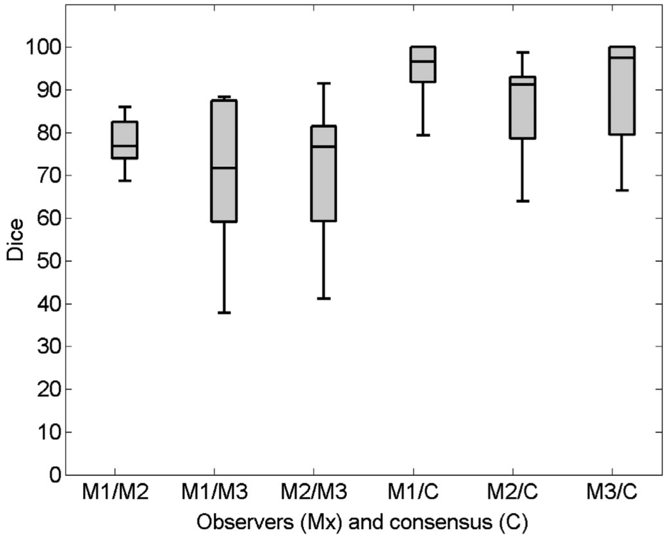

Fig. 8. Plot showing agreement between observers' segmentations (M1, M2 and M3) and consensus segmentation (C) on the combined patient and porcine datasets. For example, M1/M2 is the Dice agreement between observer's M1 and M2.

Table 5
Analysis of segmentation accuracy based on image quality (good, average and bad) on human and porcine datasets combined. The mean, standard deviation (SD) and median of the Dice for each challenger (AIT to UPF) and fixed-model method (2-SD to FWHM) is quoted.

Challengers	Poor	Average	Good
	Mean (SD), Median		
AIT	48 (19), 47	68 (23), 69	89 (9), 89
KCL	47 (22), 47	60 (23), 57	66 (20), 65
MCG	42 (25), 42	58 (18), 59	53 (24), 33
MV	41 (25), 40	32 (22), 38	38 (25), 35
UPF	46 (22), 37	52 (20), 45	44 (21), 45
2-SD	53 (22), 56	46 (22), 37	52 (20), 52
3-SD	56 (27), 61	48 (21), 39	52 (23), 54
4-SD	60 (21), 69	52 (21), 44	56 (26), 56
5-SD	66 (21), 75	55 (21), 49	59 (29), 58
6-SD	69 (21), 76	57 (19), 55	61 (32), 61
FWHM	63 (24), 64	54 (23), 51	55 (28), 54

observer. These regions were either confirmed anatomically in the case of the outflow tract or by checking adjacent slices for scar continuity in the case of partial voluming. In each image, the total volume of pseudo infarct labelled by the observer was quantified. The total volume of these spurious infarct regions present in each algorithm and fixed model segmentation was also quantified. This was possible by comparing each segmentation to the manual labellings of pseudo infarcts. Results are represented in Fig. 10. KCL and MCG had a higher proportion of manually labelled pseudo infarct regions detected on average than other methods at 21 and 23%, respectively of pseudo infarct labelled by the observers. This is in comparison to MV, AIT and UPF with only 3, 9 and 3%, respectively. Fixed models 2,3,4,5,6-SD and FWHM contained 53, 44, 36, 30, 24 and 23% respectively of manually labelled pseudo infarct volume. Pseudo infarcts were most successfully avoided in the MV and UPF algorithms and least in the 2, 3, 4 and 5-SD methods.

3.2.1. Image quality on segmentation

The LGE CMR images in the database were acquired at different imaging centres with differing protocols and scanners (see Table 2). The quality of enhancement is known to vary and it depends on a number of factors including optimal inversion times, signal-to-noise and contrast-to-noise (CNR) ratios. The images in the database were qualitatively rated by five observers experienced in LGE. Images were rated as poor, average or good depending on the overall quality of the image. The Dice overlap was measured separately in each category and these are given in Table 5. In both the good and average categories, there were 40%, 60% from the patient and porcine datasets respectively; in the poor category, there were 75%, 25% from the patient and porcine datasets, respectively. A representative set of images for each quality is shown in Fig. 9.

3.3. Discussion

We have presented a framework which standardises evaluation of algorithms for segmenting scar in the LV. The framework was used to evaluate and compare five algorithms and six separate fixed model thresholding approaches (i.e. $n-$SD and FWHM). The algorithms were submitted as part of the STACOM challenge, a workshop organised at MICCAI in 2012. The data is publicly available via the website at:

https://www.cardiacatlas.org/web/guest/ventricular-infarction-challenge.

3.3.1. Evaluation framework

The presented evaluation framework comprises of both human and animal LV LGE CMR datasets and their respective myocardial segmentation masks. Human datasets were acquired from patients with a history of ischaemic cardiomyopathy. The animal datasets were acquired in a pig model of myocardial infarction induced by coronary stenosis. Datasets were also acquired using different

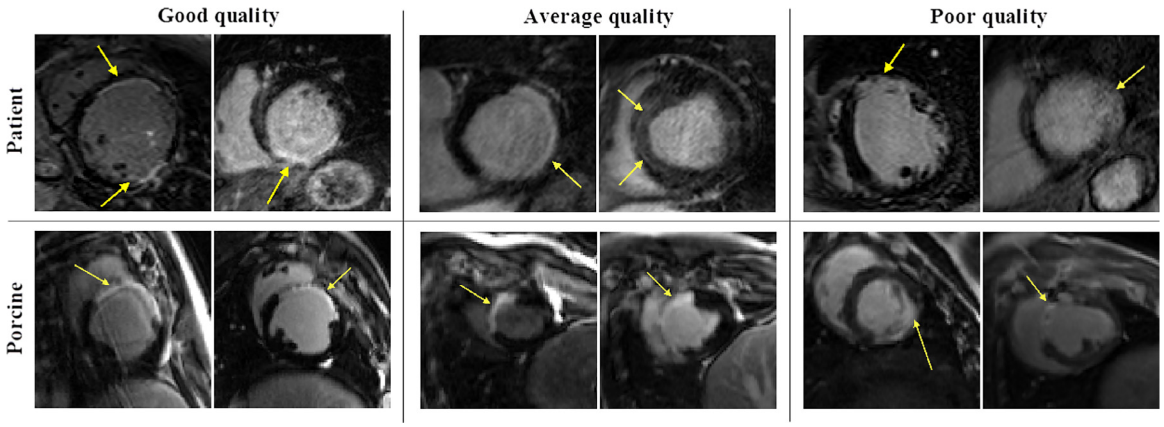

Fig. 9. Images in the patient and porcine datasets that are representative of good, average and poor quality images. The arrow labels indicate sites of possible infarction as labelled by an observer. There are two images shown for every quality.

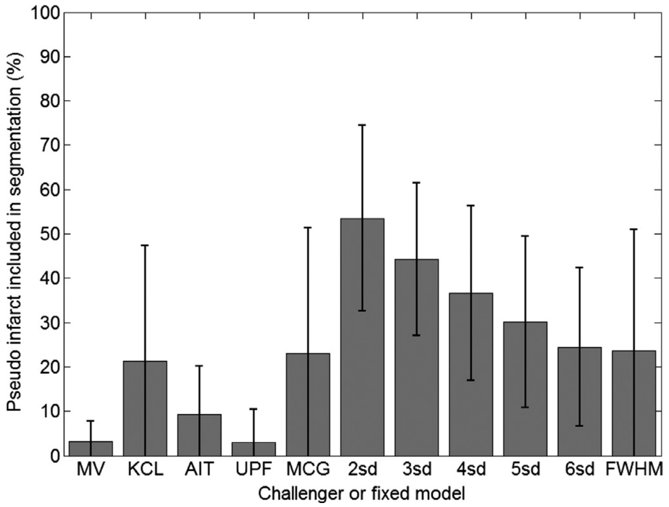

Fig. 10. The proportion of pseudo infarct manually labelled by expert observer that was detected by each method. Pseudo infarcts included hyper-enhanced regions at the right ventricle insertion points, aortic outflow tract, epi- and peri-cardial fat.

scanner vendors and resolutions. The human datasets were acquired with a 1.5T Philips scanner and the animal datasets were acquired with a 3T Siemens scanner. There were both 2D and 3D (non-isotropic) acquisitions. This ensured that algorithms evaluated on the framework were not biased to a specific acquisition protocol, scanner vendor or resolution. The proposed framework provides data acquisitions that are both commonly-used and modern, making it suitable for testing and evaluating state-of-the-art algorithms.

It is often challenging to establish ground truth on infarcted regions in LGE CMR. This makes algorithm evaluation difficult. The framework addresses this issue by proposing a reference standard against which the algorithms can be reliably evaluated. To achieve a reference standard, the human and animal datasets were manually segmented by three experienced observers provided with epi- and endo-cardial boundaries and a set of guidelines. Although, their delineations were consistent, some differences remained. The three expert delineations were combined to obtain a consensus segmentation of all three observers. The STAPLE algorithm (Warfield et al., 2004), which uses a probabilistic estimate of the true segmentation to derive the consensus, was used to obtain a consensus segmentation. The degree of agreement between observers and the computed consensus was analysed in Fig. 8 and this not only allows the assessment of agreement but also quantitatively provides for an estimation of a good Dice score in such datasets. In addition to the reference standard for scar, six commonly-used and established fixed thresholding models were used to see how they compare with the algorithms. These were namely the n-SD (where $n = 2, 3, 4, 5, 6$) and FWHM methods (Amado et al., 2004; Schmidt et al., 2007). The FWHM method is implemented as described in Amado et al. (2004), where the user clicked on hyper-enhanced regions within myocardium and an ensuing multi-pass region growing algorithm segmented infarct using the FWHM criterion.

Algorithms are often evaluated on various different metrics. This makes comparison of algorithms challenging. Most of the methods surveyed in Table 1 either use LGE volume or represent it as a percentage to evaluate detected enhancement (for example in Flett et al. (2011); Harrison et al. (2014)), or compare the amount of overlap with manual segmentation using the Dice metric (for example in Tao et al. (2010); Ravanelli et al. (2014)). The framework evaluated algorithms on both scales - volume and Dice metric. For the Dice metric, segmentations were evaluated on individual infarcted regions in the image. A Dice metric on the entire image has its pitfalls as it is difficult to ascertain within which local regions algorithms fail or succeed. This was addressed using a localised Dice evaluation strategy. Future algorithms tested on the framework will be subjected to the same metrics enabling algorithms and their segmentations to be compared in a reliable manner.

The presence of pseudo infarct, which mimics scar in LGE CMR images, poses various challenges for algorithms. Most earlier algorithms have not addressed or incorporated this into its segmentation models. The framework provided delineations of pseudo infarct regions from an experienced observer. Algorithms were assessed on the proportion of false positives due to pseudo infarct regions. This has allowed a more objective evaluation within this framework. The n-SD and FHWM fixed models segmented a large proportion of pseudo infarct labelled by the observer. The algorithms segmented significantly less pseudo infarcts than fixed models (paired t-test $p < 0.05$). Furthermore, images in the database were qualitatively rated for its quality by five different observers. Algorithms' segmentations were also evaluated separately based on the image's rating.

The proposed framework has several limitations. An important limitation is that the framework cannot be used to directly evaluate clinical utility or anatomic accuracy of the algorithms. This is since, the reference standard does not include any information about outcomes (for the patient data set) or histology (for the pig data set). Another limitation is the image database size which is 30 images, of which 20 that can be used for testing and 10 usable for training. However, within this small sample, it provides a range of datasets from different scanner vendors, scanner resolution and cohorts.

A second limitation is the dimensionality of the dataset. The human datasets are 2D acquisitions with 8 mm slice thickness. 2D images are commonly employed clinically for treatment stratification. For example based on the infarct volume and ejection fraction from 2D images, a patient could be subjected to certain therapeutic strategies, such as an implantable cardioverter defibrillator (ICD) implantation or ventricular ablation. 3D images provide more detailed quantification of infarct and only the porcine dataset within this framework are 3D non-istropic acquisitions. A third limitation is the manner in which the Dice metric is computed individually on each region of infarction labelled by the consensus. The Dice is computed only within ROIs enclosing each consensus-labelled infarct. Outside these regions, the Dice is not accountable. Thus, algorithms which over-segment can still exhibit a good Dice but poor volume error. The Dice need to be combined with the volume error to give a clearer understanding.

Intensity variation across the images due to coil shading may have an impact on segmentation, especially for methods which process absolute signal intensities. A coil sensitivity scan is a routine part of the acquisition protocol used to acquire the datasets of this study. However, no further coil sensitivity correction was carried out. This was in-line with the principal of this study to use only routine MRI scans.

A final limitation is that only one observer was employed to segment the myocardial masks. The observer was a cardiologist with several years of experience in CMR assessment of LV function assessment and ischaemic heart diseases. The issue of variability with different myocardial masks is counteracted by providing the human observers with these masks. The algorithms are also provided with the same masks. This ensures that infarct within the mask are labelled and computed. Thus, the evaluation is only carried out in the myocardial mask space.

3.3.2. Evaluated algorithms

Quantifying infarct in the LV can have important clinical implications. A 3D rendering of the LV with infarct areas can be

Table 6
The mean infarct volume (in millilitres) and average number of regions (i.e. infarct) per slice in the consensus segmentation.

	Patient data	Porcine data
Mean infarct volume (ml)	5.38 (6.73)	13.81 (8.70)
Average regions per slice	1.2 (0.5)	1.0 (0.1)

integrated into electroanatomical systems for facilitating catheter ablation. As the resolution and SNR of LGE CMR continues to improve, detailed quantification of infarct is becoming possible. The pitfalls of fixed thresholding models advocated in past literature (Amado et al., 2004; Kim et al., 1999) have been highlighted in recent studies (Harrison et al., 2014). Fixed threshold model makes crude assumptions about the contrast levels between nulled blood pool and infarct, deeming a fixed cut-off threshold. However, as these contrast levels are directly dependent on the inversion time selected in LGE CMR, the preset threshold often requires user readjustments.

The algorithms were evaluated based on the slice position (basal, mid and apical) (see Fig. 7). In the analysis, there was no significant difference between the basal and mid slices. The apical slices showed better overlap for some algorithms. However, apical slices enclose a smaller myocardial area and thus the overlap assessments in these regions can be biased. However, it is important to note that the Dice overlap used here was slice-based and not region-based as the other results in this work. In general, with Dice scores, it is difficult to ascertain what is a good Dice for datasets of the nature included in this study. The analysis of agreement between the observers' segmentations (see Fig. 8) provide for a reasonable estimation and target for the algorithms.

The algortihms' comparison to common algorithms is important. The difference with FWHM remains small except for MCG, which was able to provide high accuracy in the patient dataset, and AIT providing the same in the porcine set. Both methods have considerable strengths, with the former using a state-of-the-art probabilistic technique for image segmentation, and the latter benefitting from post-processing steps which rectify the segmentation. The Dice results reflect the strengths of these methods. On the patient datasets, algorithms AIT, MCG and UPF performed similarly while KCL and MV also performed similarly but with a lower average Dice. This was due to greater variability in Dice for KCL and MV. However, AIT and UPF are both capable of rectifying errors in its segmentation with post-processing steps. AIT employs level-sets following SVM classification and UPF employs shape discriminants. Both KCL and MV rely heavily on its core segmentation process, with no post-processing. As a result, spurious regions are included. Models that are sub-optimal were able to benefit from post-processing.

The algorithms were also evaluated on the total infarct volume it segmented (see Table 4) and these volumes were compared to the consensus volumes. This is important as Dice computed in this work has the aforementioned limitations. Also when evaluating the myocardium, quantification of infarct volume is an important step. The average volume error in challenger's algorithms were 1.04 ml and 0.76 ml for patient and porcine datasets respectively (from Table 4). This was low compared to the overall average infarct volume in the datasets (see Table 6).

The algorithms evaluated on the framework have common traits – most employ region-based image processing techniques, for example level-set (AIT), region-growing (UPF and FWHM) and watershed (MV). This is justifiable as the algorithms are meant to segment infarct that are contiguous regions. However, key considerations such as the shape of candidate regions, are not always taken into account. UPF searches for regions that are elongated, as this is a strong characteristic of LV infarcts. A second important consideration is the seed selection step. If only a single seed is allowed per slice for capturing the infarct (for example UPF, see Table 3), other infarct areas on the same slice cannot be included. The average number of infarct regions per slice was computed for both patient and porcine datasets in Table 6. With the average number of regions found to be 1.2 in the patient dataset, more than a single seed may be necessary.

A second consideration is the spatial positioning of the scar candidate in relation to the image slices or 17-segment model of the AHA (Cerqueira et al., 2002). Enhancement in the basal slices due to the outflow tract or RV insertion point should be discriminated as a pseudo infarct. None of the algorithms or fixed models, have classified enhancement based on its location. Thus, pseudo infarcts have not been addressed in the evaluated methods.

A third consideration is the extent of scarring. Sub-classification of infarct as sub-endocardial, mid-wall and epicardial helps stratify treatment. But first and foremost, these formations are indicative of scar, one which the algorithms should be able to distinguish based on Euclidean distances measured on the myocardium segmentation. Equipped with this information, algorithms should be able to better distinguish scar, especially when enhancements arise due to partial voluming or a fat-related cause.

LGE CMR for the LV can be acquired either in 2D or 3D, with the former being more common as they can be obtained relatively quickly. However, 3D acquisitions are preferred over 2D when post-processing involves detailed quantification. As scanner engineering and technology continue to improve, 3D acquisitions will become more common. All algorithms, except UPF, evaluated within this framework and those surveyed in Table 1 uses 3D techniques that also work on 2D datasets. The UPF technique performs region-growing with seed selection on a slice-by-slice basis. For the porcine 3D datasets, it chooses a particular slice orientation (*x*, *y* or *z*) to work on; and an increasing load on the operator for seed-selection in each 3D slice. The framework supplies with both types of acquisitions to enable future algorithms to be evaluated separately.

3.3.3. Future algorithms

Infarct quantification in the LV is an important assessment criteria for many cardiac therapies. Furthermore, heterogeneity within infarct, especially in the peri-infarct regions, was shown to be a predictor of tachycardia and sudden cardiac death (Schmidt et al., 2007). This work proposes an evaluation framework for future algorithms which segment and quantify LV infarct. To demonstrate its usability, five different algorithms were evaluated on the framework. Three of which have been published (Hennemuth et al., 2008; Karimaghaloo et al., 2012; Karim et al., 2014). Six different fixed-model approaches were also evaluated. The framework provides thirty datasets, of which ten are for algorithm training and the rest for testing. Although they represent a specific pulse sequence, some algorithms evaluated here could be re-trained on new sequences. The consensus ground truths are derived from manual segmentations of three separate observers. Future algorithms can be evaluated both objectively with overlap metrics or less objectively and conventionally with pixel volumes. Most importantly, they can be compared and benchmarked against existing algorithms. To our knowledge, this is the first proposed framework for evaluating LV infarct segmentation and quantification algorithms from LGE CMR images. For the left atrium, a benchmarking evaluation framework already exists (Karim et al., 2013).

4. Conclusions

CMR continues to play an important role in imaging and quantifying infarct in the LV. Several algorithms have been proposed for its quantification but it is not clear how they compare or perform relative to one another. Furthermore, algorithms have only been tested on centre- and vendor-specific images. The translation of such algorithms into the clinical environment thus remains challenging. Benchmarking frameworks, providing a common dataset and evaluation strategies, is important for clinical translation of these algorithms. The proposed benchmarking framework provides thirty datasets, with fifteen datasets in each cohort: patient and porcine. Datasets in the two separate cohorts were acquired using different scanner vendors and field strength (1.5T and 3T), resolutions and acquisition protocols (2D and 3D). The ground truth is often absent in such datasets, and to this end, the framework provides with a powerful expert observers' consensus ground truth. The proposed framework remains publicly available for accessing the image database, uploading segmentations for evaluation and contributing manual segmentations for improving the consensus ground truth on the datasets.

Acknowledgements

The author acknowledges support from the King's College London Centre of Excellence in Medical Engineering funded by the Wellcome Trust and EPSRC (WT 088641/Z/09/Z). This research was also supported by the National Institute for Health Research (NIHR) Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. Miguel A. Gonzalez Ballester was partially funded by project TIN2013-47913-C3-1-R, from the Spanish Ministry of Economy and Competitiveness.

References

Amado, L., Gerber, B., Gupta, S., Rettmann, D., Szarf, G., Schock, R., Nasir, K., Kraitchman, D., Lima, J., 2004. Accurate and objective infarct sizing by contrast-enhanced magnetic resonance imaging in a canine myocardial infarction model. J. Am. Coll. Cardiol. 44 (12), 2383–2389.

Andreu, D., Berruezo, A., Ortiz-Péréz, J.T., Silva, E., Mont, L., Borràs, R., de Caralt, T.M., Perea, R.J., Fernández-Armenta, J., Zeljko, H., et al., 2011. Integration of 3d electroanatomic maps and magnetic resonance scar characterization into the navigation system to guide ventricular tachycardia ablation. Circ.: Arrhythm. Electrophysiol. 4 (5), 674–683.

Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 1222–1239.

Cerqueira, M.D., Weissman, N.J., Dilsizian, V., Jacobs, A.K., Kaul, S., Laskey, W.K., Pennell, D.J., Rumberger, J.A., Ryan, T., Verani, M.S., et al., 2002. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart a statement for healthcare professionals from the cardiac imaging committee of the council on clinical cardiology of the american heart association. Circulation 105 (4), 539–542.

Desch, S., Eitel, I., de Waha, S., Fuernau, G., Lurz, P., Gutberlet, M., Schuler, G., Thiele, H., 2011. Cardiac magnetic resonance imaging parameters as surrogate endpoints in clinical trials of acute myocardial infarction. Trials 12 (1), 204.

Detsky, J., Paul, G., Dick, A., Wright, G., 2009. Reproducible classification of infarct heterogeneity using fuzzy clustering on multicontrast delayed enhancement magnetic resonance images. IEEE Trans. Med. Imaging 28 (10), 1606–1614.

Dietrich, O., Raya, J.G., Reeder, S.B., Reiser, M.F., Schoenberg, S.O., 2007. Measurement of signal-to-noise ratios in mr images: influence of multichannel coils, parallel imaging, and reconstruction filters. J. Magn. Reson. Imaging 26 (2), 375–385.

Estner, H.L., Zviman, M.M., Herzka, D., Miller, F., Castro, V., Nazarian, S., Ashikaga, H., Dori, Y., Berger, R.D., Calkins, H., et al., 2011. The critical isthmus sites of ischemic ventricular tachycardia are in zones of tissue heterogeneity, visualized by magnetic resonance imaging. Heart Rhythm 8 (12), 1942–1949.

Flett, A.S., Hasleton, J., Cook, C., Hausenloy, D., Quarta, G., Ariti, C., Muthurangu, V., Moon, J.C., 2011. Evaluation of techniques for the quantification of myocardial scar of differing etiology using cardiac magnetic resonance. JACC: Cardiovasc. Imaging 4 (2), 150–156.

Harrison, J.L., Jensen, H.K., Peel, S.A., Chiribiri, A., Grøndal, A.K., Bloch, L.Ø., Pedersen, S.F., Bentzon, J.F., Kolbitsch, C., Karim, R., et al., 2014. Cardiac magnetic resonance and electroanatomical mapping of acute and chronic atrial ablation injury: a histological validation study. Eur. Heart J. 35 (22), 1486–1495.

Hearst, M.A., Dumais, S., Osman, E., Platt, J., Scholkopf, B., 1998. Support vector machines. Intell. Syst. Appl. IEEE 13 (4), 18–28.

Hennemuth, A., Seeger, A., Friman, O., Miller, S., Klumpp, B., Oeltze, S., Peitgen, H.-O., 2008. A comprehensive approach to the analysis of contrast enhanced cardiac mr images. IEEE Trans. Med. Imaging 27 (11), 1592–1610.

Karim, R., Arujuna, A., Housden, R.J., Gill, J., Cliffe, H., Matharu, K., Rinaldi, C.A., O'Neill, M., Rueckert, D., Razavi, R., Schaeffter, T., Rhode, K., 2014. A method to standardize quantification of left atrial scar from delayed-enhancement mr images. J. Transl. Eng. Health Med. 2.

Karim, R., Housden, R., Balasubramaniam, M., Chen, Z., Perry, D., Uddin, A., Al-Beyatti, Y., Palkhi, E., Acheampong, P., Obom, S., Hennemuth, A., Lu, Y., Bai, W., Shi, W., Gao, Y., Peitgen, H.-O., Radau, P., Razavi, R., Tannenbaum, A., Rueckert, D., Cates, J., Schaeffter, T., Peters, D., MacLeod, R., Rhode, K., 2013. Evaluation of current algorithms for segmentation of scar tissue from late gadolinium enhancement cardiovascular magnetic resonance of the left atrium: an open-access grand challenge. J. Cardiovasc. Magn. Reson. 15 (105).

Karimaghaloo, Z., Shah, M., Francis, S.J., Arnold, D.L., Collins, D.L., Arbel, T., 2012. Automatic detection of gadolinium-enhancing multiple sclerosis lesions in brain mri using conditional random fields. IEEE Trans. Med. Imaging 31 (6), 1181–1194.

Kim, R.J., Fieno, D.S., Parrish, T.B., Harris, K., Chen, E.-L., Simonetti, O., Bundy, J., Finn, J.P., Klocke, F.J., Judd, R.M., 1999. Relationship of mri delayed contrast enhancement to irreversible injury, infarct age, and contractile function. Circulation 100 (19), 1992–2002.

Kirisli, H., Schaap, M., Metz, C., Dharampall, A., Meijboom, W., Papadopoulou, S., Dedic, A., Nieman, K., de Graaf, M., Meijs, M., Cramer, M., Broersen, A., Cetin, S., Eslami, A., Flórez-Valencia, L., Lor, K., Matuszewski, B., Melki, I., Mohr, B., Öksüz, I., Shahzad, R., Wang, C., Kitslaar, P., Unal, G., Katouzian, A., Orkisz, M., Chen, C., Precioso, F., Najman, L., Masood, S., Ünay, D., van, L.V., Moreno, R., Goldenberg, R., Vuçini, E., Krestin, G., Niessen, W., van, T.W., 2013. Standardized evaluation framework for evaluating coronary artery stenosis detection, stenosis quantification and lumen segmentation algorithms in computed tomography angiography. Med. Image Anal. 8 (17), 856–876.

Knowles, B., Caulfield, D., Cooklin, M., Rinaldi, C., Gill, J., Bostock, J., Razavi, R., Schaeffter, T., Rhode, K., 2010. 3-D visualization of acute RF ablation lesions using MRI for the simultaneous determination of the patterns of necrosis and edema. IEEE Trans. Biomed. Eng. 57 (6), 1467–1475.

Kolipaka, A., Chatzimavroudis, G.P., White, R.D., O'Donnell, T.P., Setser, R.M., 2005. Segmentation of non-viable myocardium in delayed enhancement magnetic resonance images. Int. J. Cardiovasc. Imaging 21 (2-3), 303–311.

Lafferty, J., McCallum, A., Pereira, F.C., 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, pp. 282–289.

Lu, Y., Yang, Y., Connelly, K.A., Wright, G.A., Radau, P.E., 2012. Automated quantification of myocardial infarction using graph cuts on contrast delayed enhanced magnetic resonance images. Quant. Imaging Med. Surg. 2 (2), 81.

Oakes, R., Badger, T., Kholmovski, E., Akoum, N., Burgon, N., Fish, E., Blauer, J., Rao, S., DiBella, E., Segerson, N., et al., 2009. Detection and quantification of left atrial structural remodeling with delayed-enhancement magnetic resonance imaging in patients with atrial fibrillation. Circulation 119 (13), 1758–1767.

Otsu, N., 1975. A threshold selection method from gray-level histograms. Automatica 11 (285-296), 23–27.

Petitjean, C., Zuluaga, M.A., Bai, W., Dacher, J.-N., Grosgeorge, D., Caudron, J., Ruan, S., Ayed, I.B., Cardoso, M.J., Chen, H.-C., Jimenez-Carretero, D., Ledesma-Carbayo, M.J., Davatzikos, C., Doshi, J., Erus, G., Maier, O.M., Nambakhsh, C.M., Ou, Y., Ourselin, S., Peng, C.-W., Peters, N.S., Peters, T.M., Rajchl, M., Rueckert, D., Santos, A., Shi, W., Wang, C.-W., Wang, H., Yuan, J., 2015. Right ventricle segmentation from cardiac mri: a collation study. Med. Image Anal. 19 (1), 187–202. http://dx.doi.org/10.1016/j.media.2014.10.004.

Pop, M., Ghugre, N.R., Ramanan, V., Morikawa, L., Stanisz, G., Dick, A.J., Wright, G.A., 2013. Quantification of fibrosis in infarcted swine hearts by ex vivo late gadolinium-enhancement and diffusion-weighted mri methods. Phys. Med. Biol. 58 (15), 5009.

Positano, V., Pingitore, A., Giorgetti, A., Favilli, B., Santarelli, M.F., Landini, L., Marzullo, P., Lombardi, M., 2005. A fast and effective method to assess myocardial necrosis by means of contrast magnetic resonance imaging. J. Cardiovasc. Magn. Reson. 7 (2), 487–494.

Rajchl, M., Stirrat, J., Goubran, M., Yu, J., Scholl, D., Peters, T.M., White, J.A., 2014. Comparison of semi-automated scar quantification techniques using high-resolution, 3-dimensional late-gadolinium-enhancement magnetic resonance imaging. Int. J. Cardiovasc. Imaging 1–9.

Ravanelli, D., dal Piaz, E., Centonze, M., Casagranda, G., Marini, M., Del Greco, M., Karim, R., Rhode, K., Valentini, A., 2014. A novel skeleton based quantification and 3d volumetric visualization of left atrium fibrosis using late gadolinium enhancement magnetic resonance imaging. IEEE Trans. Med. Imaging 33 (2).

Schmidt, A., Azevedo, C., Cheng, A., Gupta, S., Bluemke, D., Foo, T., Gerstenblith, G., Weiss, R., Marban, E., Tomaselli, G., et al., 2007. Infarct tissue heterogeneity by magnetic resonance imaging identifies enhanced cardiac arrhythmia susceptibility in patients with left ventricular dysfunction. Circulation 115 (15), 2006–2014.

Sethian, J.A., 1999, 3. Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science. Cambridge university press.

Suinesiaputra, A., Cowan, B.R., Al-Agamy, A.O., Elattar, M.A., Ayache, N., Fahmy, A.S., Khalifa, A.M., Medrano-Gracia, P., Jolly, M.-P., Kadish, A.H., et al., 2014. A collaborative resource to build consensus for automated left ventricular segmentation of cardiac mr images. Med. Image Anal. 18 (1), 50–62.

Tao, Q., Milles, J., Zeppenfeld, K., Lamb, H.J., Bax, J.J., Reiber, J.H., van der Geest, R.J., 2010. Automated segmentation of myocardial scar in late enhancement mri using combined intensity and spatial information. Magn. Reson. Med. 64 (2), 586–594.

Teague, M.R., 1980. Image analysis via the general theory of moments*. JOSA 70 (8), 920–930.

Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1, 211–244.

Tobon-Gomez, C., Craene, M.D., McLeod, K., Tautz, L., Shi, W., Hennemuth, A., Prakosa, A., Wang, H., Carr-White, G., Kapetanakis, S., Lutz, A., Rasche, V., Schaeffter, T., Butakoff, C., Friman, O., Mansi, T., Sermesant, M., Zhuang, X., Ourselin, S., Peitgen, H.-O., Pennec, X., Razavi, R., Rueckert, D., Frangi, A., Rhode, K., 2013. Benchmarking framework for myocardial tracking and deformation algorithms: an open access database. Med. Image Anal. 17 (6), 632–648.

Turkbey, E., Nacif, M., Noureldin, R., Sibley, C., Liu, S., Lima, J., Bluemke, D., 2012. Differentiation of myocardial scar from potential pitfalls and artefacts in delayed enhancement mri. Br. J. Radiol. 85 (1019).

Wagner, A., Mahrholdt, H., Holly, T.A., Elliott, M.D., Regenfus, M., Parker, M., Klocke, F.J., Bonow, R.O., Kim, R.J., Judd, R.M., 2003. Contrast-enhanced mri and routine single photon emission computed tomography (spect) perfusion imaging for detection of subendocardial myocardial infarcts: an imaging study. The Lancet 361 (9355), 374–379.

Warfield, S., Zou, K., Wells, W., 2004. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23 (7), 903–921.

Wu, M., D'hooge, J., Ganame, J., Ferferieva, V., Sipido, K.R., Maes, F., Dymarkowski, S., Bogaert, J., Rademakers, F.E., Claus, P., 2011. Non-invasive characterization of the area-at-risk using magnetic resonance imaging in chronic ischaemia. Cardiovasc. Res. 89 (1), 166–174.