First published in the Proceedings of the 27th European Signal Processing Conference (EUSIPCO-2019) in 2019, published by EURASIP.

A Quality of Recognition Case Study: Texture-based Segmentation and MRI Quality Assessment

Rafael Rodrigues and Antonio M. G. Pinheiro
Instituto de Telecomunicações
Universidade da Beira Interior
Covilhã, Portugal
rafael.rodrigues@ubi.pt, pinheiro@ubi.pt

***Abstract*—Muscle texture may be used as a descriptive feature for the segmentation of skeletal muscle in Magnetic Resonance Images (MRI). However, MRI acquisition is not always ideal and the texture richness might become compromised. Moreover, the research for the development of texture quality metrics, and particularly no-reference metrics, to be applied to the specific context of MRI is still in a very early stage. In this paper, a case study is established from a texture-based segmentation approach for skeletal muscle, which was tested in a thigh Dixon MRI database. Upon the obtained performance measures, the relation between objective image quality and the texture MRI richness is explored, considering a set of state-of-the-art no-reference image quality metrics. A discussion on the effectiveness of existing quality assessment methods in measuring MRI texture quality is carried out, based on Pearson and Spearman correlation outcomes.**

***Index Terms*—Magnetic Resonance Imaging; Objective Quality Assessment; Quality of Recognition (QoR); MRI Segmentation**

I. Introduction

Magnetic Resonance Imaging (MRI) has been established as an essential tool for the diagnosis of muscle-related pathologies and studying muscle physiology and anatomy. Segmentation of anatomical structures is commonly used to obtain quantitative measurements, which play an important role in diagnosis. Interest on the development of automated or semi-automated segmentation methods has been continuously increasing, mainly because manually segmenting a large amount of data, which is often the case with 3D MRI volumes, is a very difficult and time-consuming task.

Medical image quality could be compromised by a number of factors, which may be system or context-related [1]. In the case of MRI, system-related factors include magnetic field (B_0) inhomogeneity, electrical system noise or variable coil penetration depths, whilst context factors include resonance frequency shifts between different tissues or inadequate sequence parametrization [2]. Several types of image quality impairments might be induced, including white noise artifacts [3], [4], blurring [2], [3], ghosting [4], inhomogeneities in signal intensities [2] or geometric distortions [2].

MRI quality assessment (MRI-QA), and of medical image in general, provides valuable insight into the relation between processes such as image acquisition, compression, transmission or display, and the perceived and diagnostic quality of medical visual content. The main goal of such studies is to obtain recommendations to assure image quality in clinical practice.

This research was funded by Fundação para a Ciência e Tecnologia, under the research grant SFRH/BD/130858/2017, and the Instituto de Telecomunicações - Fundação para a Ciência e Tecnologia, under the project UID/EEA/50008/2019.

II. Background & Related Work

Currently, research efforts in quality assessment of medical image are somewhat disperse, due to the extensive amount of different visual content used in medical practice and the inherent quality issues mentioned before, which hinder the definition of a reference. Moreover, subjective quality assessment should also take into account the diagnostic quality and, therefore, rely on expert evaluators [5].

Some studies on subjective MRI-QA may be found in [3], [4], [6]–[8]. As for objective quality assessment, most efforts apply common reference metrics, such as the Peak Signal-to-Noise Ratio (PSNR) [6], [9]–[12] and the Structural Similarity Index (SSIM) [6], [9], [10]. The goal of these studies has been almost invariably to evaluate the influence of artifacts and compression on the perceived image quality or to measure the performance of reconstruction or filter methods.

Image quality assessment (IQA) using full-reference or even reduced-reference metrics requires content marked as optimal, which is not always possible, especially when dealing with medical content. Although research in no-reference assessment of MRI quality is sparse, some approaches using actual clinical context MRI content may be found in [12]–[14]. Another possible approach to study the quality assessment of MRI acquisition is using phantom/test object measurements, as reported in [2].

Some other setbacks arise in objective MRI-QA using no-reference metrics, given the variety of applications in relation to which quality might be evaluated. For example, a metric strongly correlated with the perceived image quality may not effectively predict the suitability of the same image for diagnosis purposes.

According to [15], quality of recognition (QoR) research aims at modeling quality assessment methods using recognition tasks. In the case of the referred paper, the authors study the quality of video content used for recognition tasks and task-based multimedia applications. To the best of our knowledge, there is no study correlating segmentation outcomes

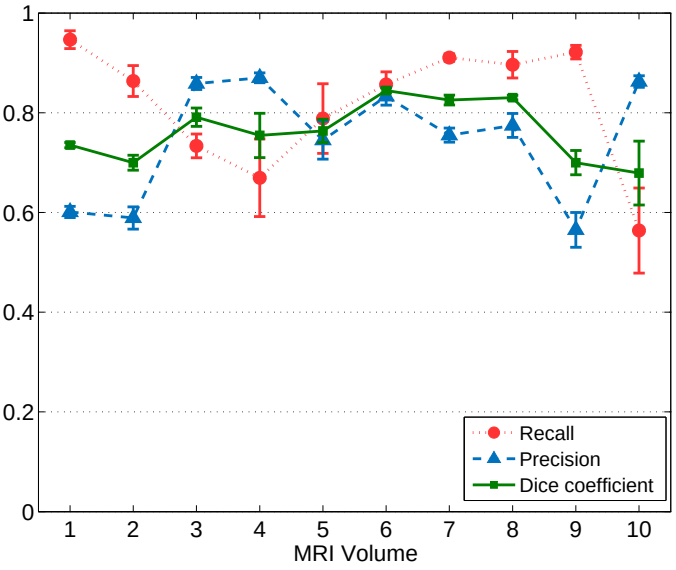


Fig. 1. Mean and standard deviation of performance measures per MRI volume. Recall, precision and the Dice overlap coefficients are plotted in red, blue and green marks, respectively.

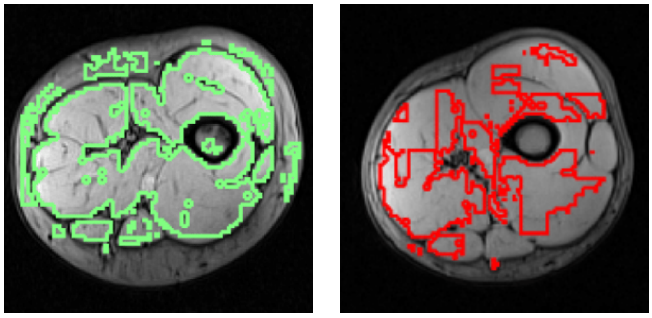
with objective MRI quality. In this paper, a QoR approach to medical IQA is proposed, using the texture-based automated segmentation framework proposed in [16] as the recognition task. The segmentation method relies on texture differences between tissues to classify a region of interest (ROI) as muscle or non-muscle tissue. A visual inspection of the available MRI data and the corresponding segmentations suggested a tendency for lower quality MRI to lead to worse results. A series of existing objective IQA metrics were correlated with the reported segmentation performance, to evaluate the adequation of texture recognition as an indirect IQA measure towards the future development of a QoR-based MRI-QA model.

III. Methods

A. MRI Dataset description

The original MRI data consisted on volumetric acquisitions of both thighs performed on a 3T scanner (Tim Trio, Siemens Healthcare, Erlangen, Germany) using a 3-point Dixon Gradient Echo sequence [17], with the following parametrization: TR = 10 ms, TE = 2.75 / 3.95 / 5.15 ms, RF flip angle = 3°, matrix: 448×224×64, field of view (FOV): 448×224×320mm^3 (voxel size: 1×1×5mm^3) [18].

For this study, the working dataset included only the Out of Phase images of the right thigh of 10 healthy subjects, with an image size of 224×224 pixels. From a total of 64 slices from each subject, a centered subset of 40 slices was considered, discarding images near the knee and the ankle. From this subset, 5 images were then randomly selected for segmentation and quality assessment. Manual segmentations of clinically relevant muscles were provided for each image and used as ground truth masks for algorithm training and performance assessment.



(a) Segmentation with higher recall rate (b) Segmentation with lower recall rate

Fig. 2. Examples of muscle tissue segmentation results.

B. Segmentation of Skeletal Muscle

An automated method for the segmentation of skeletal muscles in MRI was proposed in [16]. To account for the rotation caused by image registration in the described algorithm, all the images in the working dataset were zero-padded across both dimensions, resulting in a final image size of 256×256. Each image was subdivided into 16x16 non-overlapping cells and a set of local features were computed within each cell. The descriptor included the Histogram of Oriented Gradients (HOG) [19] and statistical measures (mean, variance, skewness, and kurtosis) from both the grayscale image and a filtered image using a Laplacian of Gaussian (LoG) filter [20]. Moreover, the descriptor also included the Haar wavelet coefficients [21] from a 3 level decomposition, which allowed for a finer segmentation.

An AdaBoost classifier [22], [23] was trained with the proposed features, in a 10-fold cross-validation. For the classification of images from a given MRI volume, the features retrieved from images of the remaining 9 volumes were chosen for training.

The binary output was then labeled, using a probabilistic muscle atlas derived from training images in each cross-validation iteration. However, given the purpose of this study, the method performance measures take into consideration the whole muscle tissue, and not differentiated muscles. Segmentation performance was assessed by computing recall, precision, and the Dice overlap coefficient:

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Dice = \frac{2|A \cap B|}{|A| + |B|} \tag{3}$$

where TP refers to true positives, FN to false negatives and FP to false positives. In equation (3), A and B refer to muscle-labeled regions in the segmentation output and ground truth, respectively. Fig. 1 summarizes the reported cross-validation performance outcomes. In Fig. 2, two examples of muscle

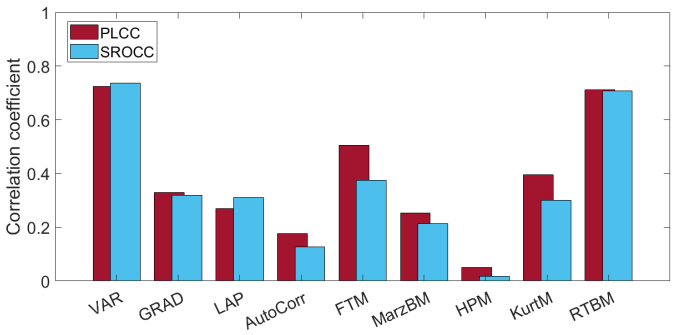
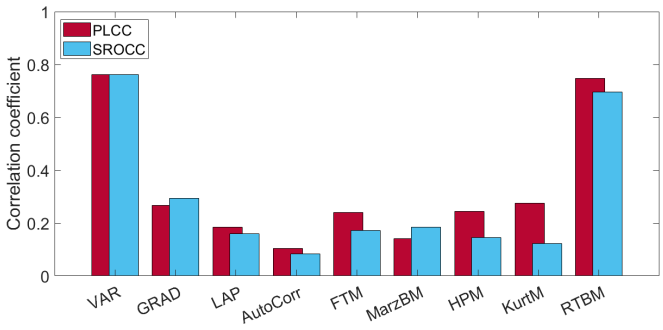
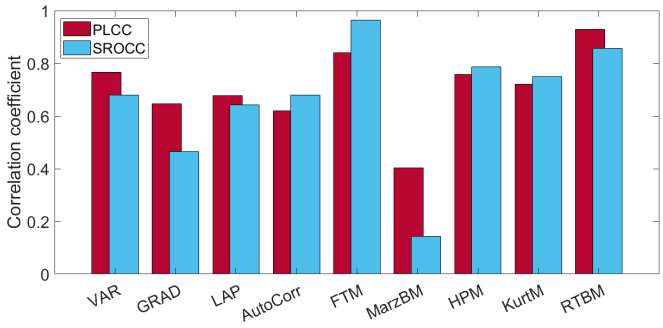


Fig. 3. Correlation coefficients for segmentation recall rates vs. IQA data (Full data).



(a) Correlation coefficients for segmentation recall rates vs. IQA data (normalized Dice Overlap Coefficients > 0.7).



(b) Correlation coefficients for segmentation recall rates vs. IQA data (normalized Dice Overlap Coefficients ≤ 0.7).

Fig. 4. Correlation coefficients for splitted segmentation recall rates vs. IQA data, according to the corresponding Dice Overlap Coefficients.

tissue segmentation are presented (high recall in Fig. 2(a) and lower recall in Fig. 2(b)).

C. Objective Quality Assessment

A database of reference quality MRI was not available for this study. Therefore, and given the lack of MRI-specific no-reference metrics, the proposed approach for a QoR study involving MRI texture segmentation relies on a set of state-of-the-art no-reference methods which may describe image-specific properties such as sharpness, speckle noise, and intensity/contrast inhomogeneities.

The considered metrics have already been implemented in [24], namely: Variance (VAR) [25], Laplacian (LAP) [26], Gradient (GRAD) [26], Autocorrelation (AutoCorr) [26], Frequency Threshold metric (FTM) [27], Marziliano Blurring metric (MarzBM) [27], HP metric (HPM) [28], Kurtosis-based metric (KurtM) [29], and Riemannian Tensor-based metric (RTBM) [30].

As described in section III.B, the texture descriptors were based on local information, extracted from 16x16 cells, combined with finer information from wavelet decompositions. Image quality metrics were also computed within similar 16x16 non-overlapping windows, to attempt the retrieval of local variations, more closely related to the segmentation scheme. Final IQA values for each MRI scan were obtained by taking the average of cell-based metric outputs.

The chosen window size, for both texture and quality analysis, was arbitrarily defined to ensure a sufficiently large frame for meaningful information extraction.

In order to focus on muscle texture recognition, a ROI-based approach was followed. According to the ground truth images of the Dixon MRI database, quality metrics were only computed for regions marked as muscle tissue. Moreover, a correlation was only established between recall rates, i.e. sensitivity, and IQA metrics, since it represents retrieval of true positives.

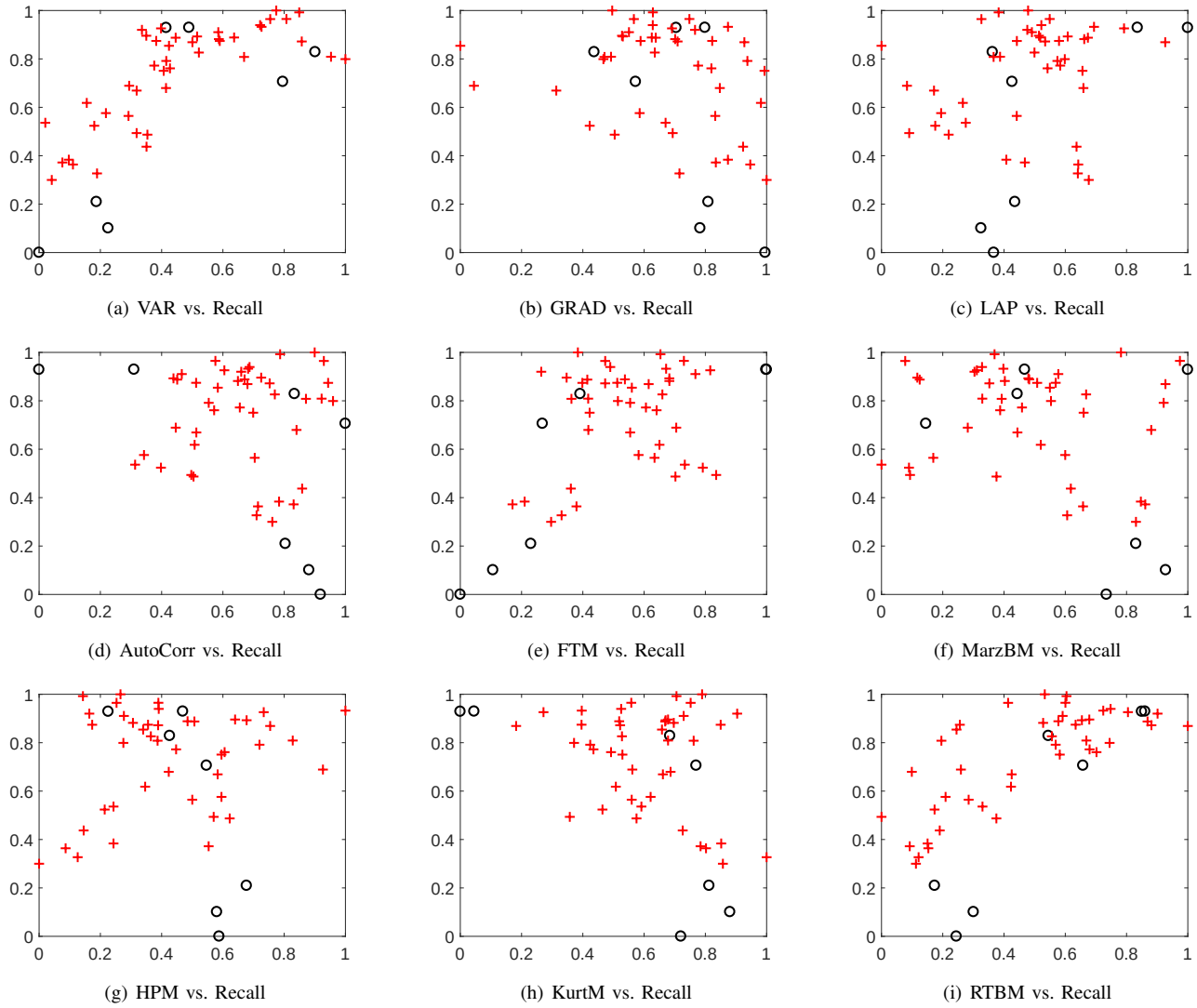
IV. Results & Discussion

In this study, IQA outcomes were compared to texture segmentation performance measures, considering each individual image. The cross-validation scheme in [16] leads to a total of 50 data pairs for each IQA metric vs. performance case (10 volumes*5 images).

In Fig. 3 the absolute values of the Pearson Linear Correlation (PLCC) and the Spearman Rank Order Correlation (SROCC) coefficients of recall rates vs. IQA metrics are shown. These results consider the full set of segmentation results. The top-performing metrics are the Variance metric (PLCC = 0.724, SROCC = 0.737) and RTBM (PLCC = 0.712, SROCC = 0.707). Taking into account the remaining metrics, the results drop significantly.

Analyzing recall vs. IQA scatter plots, a highly non-linear region may be observed. As stated in section III.C, only recall rates are correlated with IQA in this study, given the proposed task for the QoR approach, i.e. texture recognition. Metric computation also relied on averaging over a local-approach, which may also play a role in the observed non-linearity. However, it does not seem probable that taking a single quality value over the entire image could produce a more strongly correlated output. A major drawback would be the presence of different tissues in each MRI slice, which translates into texture diversity. The texture-based segmentation outcomes are coherent with this assumption.

Typically, there is a trade-off between recall and precision, observable in classification and segmentation tasks, which may be confirmed in Fig. 1. An increase in the true positive rate may also increase the number of false positives, which affects the precision rate. In terms of segmentation evaluation, the



(a) VAR vs. Recall (b) GRAD vs. Recall (c) LAP vs. Recall

(d) AutoCorr vs. Recall (e) FTM vs. Recall (f) MarzBM vs. Recall

(g) HPM vs. Recall (h) KurtM vs. Recall (i) RTBM vs. Recall

Fig. 5. Scatter plots of normalized recall rates (x-axis) vs. normalized IQA Metrics (y-axis). Circle points represent cases with Dice ≤ 0.7 and cross points refer to Dice > 0.7.

Dice overlap coefficient provides a more specific measure of segmentation accuracy. To try to reduce the non-linearity in the reported correlations, the Dice coefficient was used to divide recall data considering a threshold of 0.7.

The results were then analyzed considering the resulting subsets. Fig. 4(a) shows correlation coefficients for segmentation outputs with Dice > 0.7. On the other hand, Fig. 4(b) only takes into account segmentation outputs with Dice ≤ 0.7. In Fig. 5, recall vs. IQA scatter plots are shown, also considering the referred subsets.

In the case of Dice ≤ 0.7, it should be noted that only 7 points were retrieved, from a total of 50 recall-IQA pairs. Stronger correlations were obtained with all metrics, with the best performances being obtained with RTBM (PLCC = 0.928, SROCC = 0.857) and FTM (PLCC = 0.840, SROCC = 0.964).

When the best segmentation results are taken into account, considering the Dice overlap coefficient, correlation values are similar to those obtained with the full dataset (Fig 4(a)). The Variance metric, with PLCC = 0.761 and SROCC = 0.762 (Fig. 5(a)), and RTBM, with PLCC = 0.748 and SROCC = 0.6956 (Fig. 5(i)) remain the top-performing IQA metrics.

The RTBM and VAR metrics showed an acceptable correlation with the recall of the segmentations in all tested scenarios (full data and Dice-based data selection), while the remaining IQA metrics did not perform well. In the scatter plots of RTBM and VAR (Figs. 5(a) and 5(i), respectively), there is a tendency for points with higher recall to appear in a higher quality region. Similarly, points with lower recall tend to appear in a lower quality region. For these extreme cases, the texture recognition task was able to predict the IQA outcome, while the correlation between mid-range recall/IQA is slightly less significant.

V. Conclusions

A QoR approach to the image quality assessment of MRI was presented in this paper, following the results of a local-

based texture segmentation method. This study was deployed as an initial effort towards the development of no-reference task-based quality assessment models for MRI.

Task-based approaches may present a solution for the problem of objective MRI quality assessment, as well as medical imaging in general. In this study, texture recognition showed a reasonable potential to model the outcome of existing no-reference IQA metrics.

The tested metrics are designed for general purpose IQA and typically do not yield a very high performance. Therefore, future research should consider metrics specifically designed for MRI or, at least, medical image applications. Also, a more comprehensive MRI dataset should be tested to further validate these preliminary findings.

ACKNOWLEDGEMENT

The authors would like to thank Dr. Pierre Carlier and his research team at the NMR Laboratory of the Institut de Myologie, Paris, for providing the MRI database for this study.

REFERENCES

[1] U. Reiter, K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank, "Factors influencing quality of experience," in *Quality of experience*. Springer, 2014, pp. 55–72.

[2] M. Davids, F. G. Zöllner, M. Ruttorf, F. Nees, H. Flor, G. Schumann, and L. R. Schad, "Fully-automated quality assurance in multi-center studies using MRI phantom measurements," *Magnetic resonance imaging*, vol. 32, no. 6, pp. 771–780, 2014.

[3] L. S. Chow, H. Rajagopal, R. Paramesran, and Alzheimer's Disease Neuroimaging Initiative, "Correlation between subjective and objective assessment of magnetic resonance (MR) images," *Magnetic resonance imaging*, vol. 34, no. 6, pp. 820–831, 2016.

[4] H. Liu, J. Koonen, M. Fuderer, and I. Heynderickx, "The relative impact of ghosting and noise on the perceived quality of MR images," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3087–3098, 2016.

[5] L. Lévêque, H. Liu, S. Baraković, J. B. Husić, M. Martini, M. Outtas, L. Zhang, A. Kumcu, L. Platisa, R. Rodrigues *et al.*, "On the subjective assessment of the perceived quality of medical images and videos," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018, pp. 1–6.

[6] B. Kumar, S. P. Singh, A. Mohan, and H. V. Singh, "MOS prediction of SPIHT medical images using objective quality parameters," in *2009 International Conference on Signal Processing Systems*. IEEE, 2009, pp. 219–223.

[7] Y. Gaudeau, J. Lambert, N. Labonne, and J.-M. Moureaux, "Compressed image quality assessment: Application to an interactive upper limb radiology atlas," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 501–505.

[8] J. Miao, F. Huang, S. Narayan, and D. L. Wilson, "A new perceptual difference model for diagnostically relevant quantitative image quality evaluation: a preliminary study," *Magnetic resonance imaging*, vol. 31, no. 4, pp. 596–603, 2013.

[9] J. A. Dowling, B. M. Planitz, A. J. Maeder, J. Du, B. Pham, C. Boyd, S. Chen, A. P. Bradley, and S. Crozier, "Visual quality assessment of watermarked medical images," in *Medical Imaging 2007: Image Perception, Observer Performance, and Technology Assessment*, vol. 6515. International Society for Optics and Photonics, 2007, p. 65151L.

[10] J. Mohan, Y. Guo, V. Krishnaveni, and K. Jeganathan, "MRI denoising based on neutrosophic wiener filtering," in *Imaging Systems and Techniques (IST), 2012 IEEE international conference on*. IEEE, 2012, pp. 327–331.

[11] S. Ravishankar and Y. Bresler, "Sparsifying transform learning for compressed sensing MRI," in *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*. IEEE, 2013, pp. 17–20.

[12] H. Luong, B. Goossens, J. Aelterman, L. Platiša, and W. Philips, "Optimizing image quality in MRI: on the evaluation of k-space trajectories for under-sampled MR acquisition," in *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*. IEEE, 2012, pp. 25–26.

[13] M. B. A. Haghighat, A. Aghagolzadeh, and H. Seyedarabi, "A non-reference image fusion metric based on mutual information of image features," *Computers & Electrical Engineering*, vol. 37, no. 5, pp. 744–756, 2011.

[14] A. Krishn, V. Bhateja, and A. Sahu, "Medical image fusion using combination of PCA and wavelet analysis," in *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*. IEEE, 2014, pp. 986–991.

[15] W. Heng, T. Jiang, and W. Gao, "How to assess the quality of compressed surveillance videos using face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[16] R. Rodrigues and A. M. G. Pinheiro, "Segmentation of skeletal muscle in thigh Dixon MRI based on texture analysis," *arXiv preprint arXiv:1904.04747*, 2019.

[17] G. Glover and E. Schneider, "Three-point dixon technique for true water/fat decomposition with b0 inhomogeneity correction," *Magnetic resonance in medicine*, vol. 18, no. 2, pp. 371–383, 1991.

[18] Y. Barnouin, G. Butler-Browne, T. Voit, D. Reversat, N. Azzabou, G. Leroux, A. Behin, J. S. McPhee, P. G. Carlier, and J.-Y. Hogrel, "Manual segmentation of individual muscles of the quadriceps femoris using mri: a reappraisal," *Journal of Magnetic Resonance Imaging*, vol. 40, no. 1, pp. 239–247, 2014.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[20] S. Agaian and A. Almuntashri, "Noise-resilient edge detection algorithm for brain MRI images," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE, 2009, pp. 3689–3692.

[21] Y. Zhang, Z. Dong, L. Wu, and S. Wang, "A hybrid method for MRI brain image classification," *Expert Systems with Applications*, vol. 38, no. 8, pp. 10 049–10 053, 2011.

[22] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *International Conference on Machine Learning*, vol. 96, 1996, pp. 148–156.

[23] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *International Journal of Computer Vision*, vol. 63, no. 2, pp. 153–161, 2005.

[24] P. Hanhart, M. V. Bernardo, M. Pereira, A. M. Pinheiro, and T. Ebrahimi, "Benchmarking of objective quality metrics for HDR image quality assessment," *EURASIP Journal on Image and Video Processing*, vol. 2015, no. 1, p. 39, 2015.

[25] S. Erasmus and K. Smith, "An automatic focusing and astigmatism correction system for the SEM and CTEM," *Journal of Microscopy*, vol. 127, no. 2, pp. 185–199, 1982.

[26] C. F. Batten, "Autofocusing and astigmatism correction in the scanning electron microscope," *Mphill thesis, University of Cambridge*, 2000.

[27] A. V. Murthy and L. J. Karam, "A MATLAB-based framework for image and video quality evaluation," in *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*. IEEE, 2010, pp. 242–247.

[28] D. Shaked and I. Tastl, "Sharpness measure: Towards automatic image enhancement," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 1. IEEE, 2005, pp. I–937.

[29] N. F. Zhang, A. Vladar, M. T. Postek, and R. D. Larrabee, "A kurtosis-based statistical measure for two-dimensional processes and its applications to image sharpness," Tech. Rep., 2003.

[30] R. Ferzli and L. J. Karam, "A no reference objective shapness metric using Riemannian tensor," *Simulation*, vol. 1, p. 1, 2007.