


# Liver Segmentation Quality Control in Multi-Sequence MR Studies

Yi-Qing Wang <sup>1</sup> and Giovanni Palma<sup>1</sup>

**Abstract** For an automated liver disease diagnosis system, the ability to assess the liver segmentation quality in the absence of ground truth is crucial. Because it helps detect algorithm failures at inference time so that erroneous outputs can be prevented from impacting the diagnosis accuracy. In addition, it can be used to quality check annotated data for training and testing purposes. In this paper, we introduce the concept of liver profile as the basis for an exploratory data analysis approach to identifying poorly segmented images in multi-sequence MR liver studies.

## 1 Introduction

Liver segmentation is key to automated liver disease diagnosis [7]. The ability to assess the segmentation quality in the absence of ground truth is thus of interest. It allows to detect algorithm failures at inference time, which is critically important in practice because an erroneous liver segmentation may lead to errors in downstream tasks such as lesion detection [3, 4, 12, 17, 18] and image registration [1, 9], thereby negatively impacting the overall diagnosis accuracy. Additionally, this ability can also help ensure the quality of annotated training and test datasets by identifying poor (image, segmentation mask) pairs, potentially resulting in better algorithms and more accurate performance evaluation.

The quality of liver segmentation, either by human annotators or by an algorithm, depends on the quality of the medical images under analysis. Strongly degraded images generally result in inaccurate segmentations. However, it is hard to quantify image quality in absolute terms, because images of good quality for one task can be poor for another [2]. For instance, in a medical image, it can be easy to ascertain the presence of a liver lesion. But its measurement and characterization can prove difficult if its textural details are hard to discern. For this reason, human experts tend

---

IBM Watson Health Imaging, Rue Alfred Kastler, 91400 Orsay, France  
e-mail: yi-qing.wang@ibm.com, giovanni.palma@ibm.com

to disagree with each other when it comes to rating medical images in terms of their overall quality [6, 13, 15], especially when they are trained in different medical fields [11].

How to automate segmentation quality assessment without access to ground truth has drawn some attention lately. Most existing methods are of supervised nature and need a set of well annotated samples to begin with. For example, the work [10] suggests to train a SVM-based regressor on geometric, intensity and gradient features to predict several segmentation error metrics with respect to ground truth. Similarly, a framework was presented in [16] which uses a model to evaluate an image’s segmentation quality by checking its consistency with some known annotated samples. In a few recent works [5, 8, 14], the authors proposed to use uncertainty information from model produced probabilistic segmentation maps. Although these methods can be used to detect segmentation failures at inference time, they do not lend themselves easily to training data quality control because manual annotations are almost always binary valued.

In this work, we propose an exploratory data analysis approach to assessing liver segmentation quality in MR studies of the abdomen. It is based on the following observations: 1) an MR study typically consists of multi-sequence volumes and 2) within a study, all the volumes portray the same liver. Therefore, a study’s well segmented volumes should yield consistent liver size statistics.

This paper is organized as follows. We first introduce the concept of liver profile and describe its properties. Next we propose a simple algorithm to estimate the liver profile from a multi-sequence liver study and demonstrate its effectiveness at detecting incorrectly segmented image slices. Finally we conclude and discuss future work.

## 2 Liver profile

### 2.1 Definition

Consider a human liver. We define its *profile* as a function that maps a transverse plane to its corresponding liver cross-sectional area. By definition, it thus requires an infinity of axial liver slices and cannot be computed directly. Since the liver is a smooth three dimensional object, its profile must be continuous. Therefore, it can be estimated from a medical scan produced by imaging techniques such as CT and MR, which samples liver slices on a regular interval.

Specifically, consider an  $n$ -slice axial liver scan with slice spacing equal to  $s_z$  (in millimeters). Let  $S_i$  denote the area in square millimeters of the liver cross-section captured by the scan’s  $i$ -th slice. Let  $\phi$  be an interpolation through the points  $\{(i, S_i)\}_{i=0, \dots, n-1}$ . For example, the linear interpolation leads to

$$\phi(x) = (S_{\lfloor x \rfloor + 1} - S_{\lfloor x \rfloor})(x - \lfloor x \rfloor) + S_{\lfloor x \rfloor}, \quad x \in [0, n - 1) \quad (1)$$

where  $\lfloor x \rfloor$  denotes the largest integer less than or equal to  $x \in \mathbb{R}$ . The interpolation is then scaled to result in the *scan profile*  $P(t) := \phi(t/s_z)$ . Over its *support*  $\{t | P(t) > 0\}$ , the scan profile can be considered as an approximation of the liver profile. The support's length is referred to as the scan's *liver span*.

Note that in Eq.(1), it is an arbitrary choice to give the index of zero to the scan's first slice. Instead, we could set its index to another value, such as 1, and maintain the same concept of scan profile. In other words, both liver and scan profile are uniquely defined only up to a translation.

## 2.2 Properties

In the absence of major clinical events (such as a partial hepatectomy), a liver tends to have a rather static profile over a short period of time. Moreover, as long as a person's longitudinal axis points to the same direction, their liver axial cross-sectional areas are relatively insensitive to rigid body motions. As a result, a patient's various well performed scans should lead to similar-looking scan profiles, all of which resemble the same underlying liver profile.

Though the liver profiles may vary in shape from person to person (see Fig.1), they have an asymmetrical bell shape in general. It is because the liver cross-sectional area usually peaks at a transverse plane which passes through both left and right liver lobes and gradually decreases as we move the plane towards the liver's superior or inferior surfaces.

## 2.3 Estimation

Generally speaking, a scan profile is a noisy and partial estimate of its corresponding liver profile. Its approximation quality depends on the scan's slice spacing, voxel resolution and the accuracy of liver cross-sectional area measurements. A smaller slice spacing, finer voxel resolution and more accurate liver segmentation lead to a scan profile of higher approximation quality.

For patients who have undergone multiple liver scans, it is possible to obtain an even better estimation of their liver profiles than the individual scan profiles themselves. To do so, consider a patient's  $m$  scan profiles  $\{P_i\}_{i=1,\dots,m}$ . Their supports, as defined previously, are of relative value because two different scans rarely portray the same abdominal region. However, since all the scan profiles describe the same liver, we can find a common coordinate system to represent them.

Without loss of generality, let us assume that the patient's first scan profile  $P_1$  has the greatest liver span. We fix it as the reference and translate the other scan profiles to align with it individually. To this end, we use the following metric to assess the quality of alignment between two positive valued functions with finite support

$$\text{agreement}(f, g) = \frac{\int_{\{t|f(t)>0, g(t)>0\}} \min(f(t), g(t)) dt}{\int_{\{t|f(t)>0, g(t)>0\}} \max(f(t), g(t)) dt} \quad (2)$$

which is the Jaccard index of the areas underneath these two functions restricted to their common support. Aligning a scan profile thus amounts to finding its optimally translated version that has the maximum agreement with the reference.

---

**Algorithm 1: Liver Profile Estimation**


---

**Data:**  $m$  scan profiles  $\{P_i\}_{i=1, \dots, m}$   
**Result:**  $m$  aligned scan profiles and estimated liver profile  $P^*$   
 $\alpha \leftarrow 0;$   
 $i \leftarrow 0;$   
 $P_0^* \leftarrow$  the scan profile with the greatest liver span;  
**while**  $\alpha < 0.99$  **and**  $i < 5$  **do**  
    Align all the scan profiles with  $P_0^*;$   
     $P_1^* \leftarrow$  pointwise median of the aligned scan profiles i.e. Eq.(3);  
     $\alpha \leftarrow \text{agreement}(P_0^*, P_1^*);$   
     $i \leftarrow i + 1;$   
     $P_0^* \leftarrow P_1^*;$   
**end**  
 $P^* \leftarrow P_0^*;$

---

Once all the scan profiles have been aligned, we may estimate the liver profile. Specifically, the estimator's support is defined as the union of those of the aligned scan profiles and its values are set pointwise to the median of the aligned scan profiles. It leads to

$$P^*(t) = \text{median} \cup_{1 \leq i \leq m, P_i(t) > 0} \{P_i(t)\}, \quad t \in \cup_{1 \leq i \leq m} \{t | P_i(t) > 0\} \quad (3)$$

where we continue to use  $P_i$  to denote an aligned scan profile.

Next, we substitute the estimated liver profile  $P^*$  for  $P_1$  and align the entire set of scan profiles again with this new reference. These two operations are then repeated until successively obtained  $P^*$  stabilizes, which usually takes less than 5 iterations. We call the estimate  $P^*$  from the final iteration the patient's *estimated liver profile*. This procedure's pseudo code is provided in Algorithm 1.

## 3 Experiments

### 3.1 Data

Let us first describe our data. It is a private collection of 70 MR studies of adult patients from three different hospitals. The number of volumes per study varies from

4 to 11, totaling 558 volumes in all. They were acquired using various T1, T2 and diffusion weighted MR sequences with slice spacing ranging from 2mm to 11mm. A team of radiologists examined them one volume at a time and created the liver masks for the entire dataset, leading to 28458 marked slices.

Quality checking the annotated volumes one by one is tedious and cannot scale to larger datasets. We now describe how the liver profile can help us quickly identify the likely inaccurate segmentation masks. We had experimented with both linear and cubic spline interpolation schemes for constructing scan profiles. They yielded little difference. Therefore, for simplicity, we chose linear interpolation i.e. Eq.(1).

### 3.2 Exploratory analysis at the volume level

Study-wise, the scan profiles from our annotated volumes are broadly consistent. To show it, we used the agreement metric defined in Eq.(2). Specifically, for the aligned scan profiles of a study, we are interested in their individual agreement with the study's estimated liver profile. Clearly, those who agree with the estimated liver profile agree well with themselves, too. To simplify the presentation, in the following, we call an aligned scan profile's agreement with its estimated liver profile its *coherence score*, which thus also takes values in the interval  $[0, 1]$ .

A typical MR study consists of volumes with varying slice spacing and voxel resolution, leading to scan profiles of different approximation quality. As a result, their coherence scores rarely equal 1. For example, the aligned scan profiles shown in Fig.1 and Fig.4, though they agree with their respective estimated liver profile, have coherence score ranging from 0.96 to 0.99.

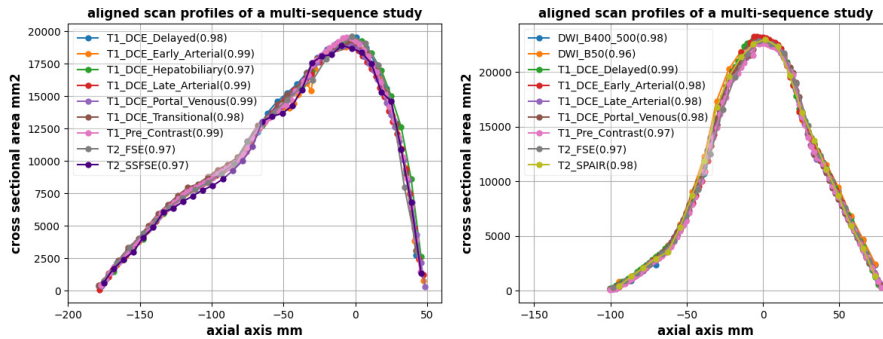


Fig. 1: aligned scan profiles of two annotated multi-sequence MR liver studies from two different patients. They are consistent within the study. In the legend, we print for each scan profile its coherence score. These two examples show that the liver profile varies in shape across people.

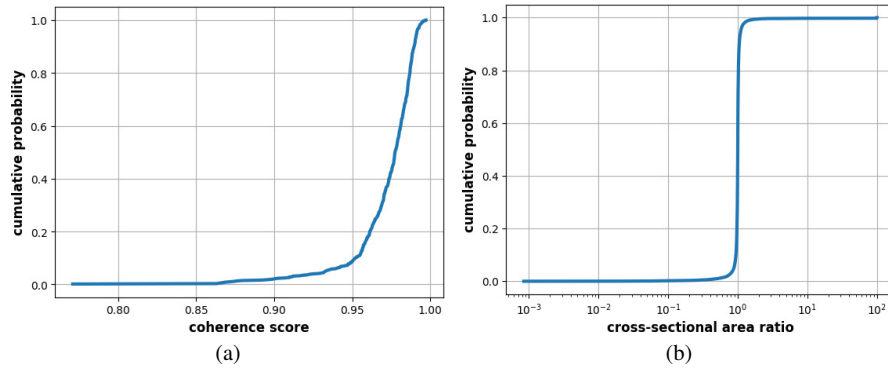


Fig. 2: (a) plots the cumulative distribution function (CDF) of the 558 coherence scores. They are mainly distributed close to 1. In fact, 90% of them exceed 0.95. (b) shows the CDF of the ratio between the segmented and expected liver cross-sectional area of the marked slices. The distribution of this slice-level statistic is concentrated around 1, indicating that the two area measures broadly agree.

We ran Algorithm 1 on our data one study at a time. It resulted in as many coherence scores as there are volumes in our data. Fig.2a plots their cumulative distribution function. Most of them indeed lied close to 1. Specifically, 90 percent of these volumes had a coherence score above 0.95. To detect inaccurate segmentations at the volume level, we thus retrieved the 55 volumes and their liver masks corresponding to the lowest 10 percent of the obtained coherence scores. They belonged to a total of 47 studies, each of which had at most 2 of these volumes. After a visual inspection, we found that among them, 43 annotated volumes with the lowest coherence scores were faulty because of either bad image quality or a visible segmentation error. See Fig.3 for an example.

### 3.3 Exploratory analysis at the slice level

For these faulty segmentations, their aligned scan profiles also help locate where the faults occur. It is because they help identify the image slices whose segmented liver cross-sectional area differs considerably from what is expected from their corresponding estimated liver profile at the same axial axis locations. In the absence of exceptional medical conditions which reduced or expanded the liver size, such a disagreement suggests a segmentation error, which may also be caused by bad image quality.

Therefore, we can use the aligned scan profiles to explore potential segmentation errors at the slice level. Specifically, for every annotated slice, we computed the ratio between its segmented liver cross-sectional area and the expected area from

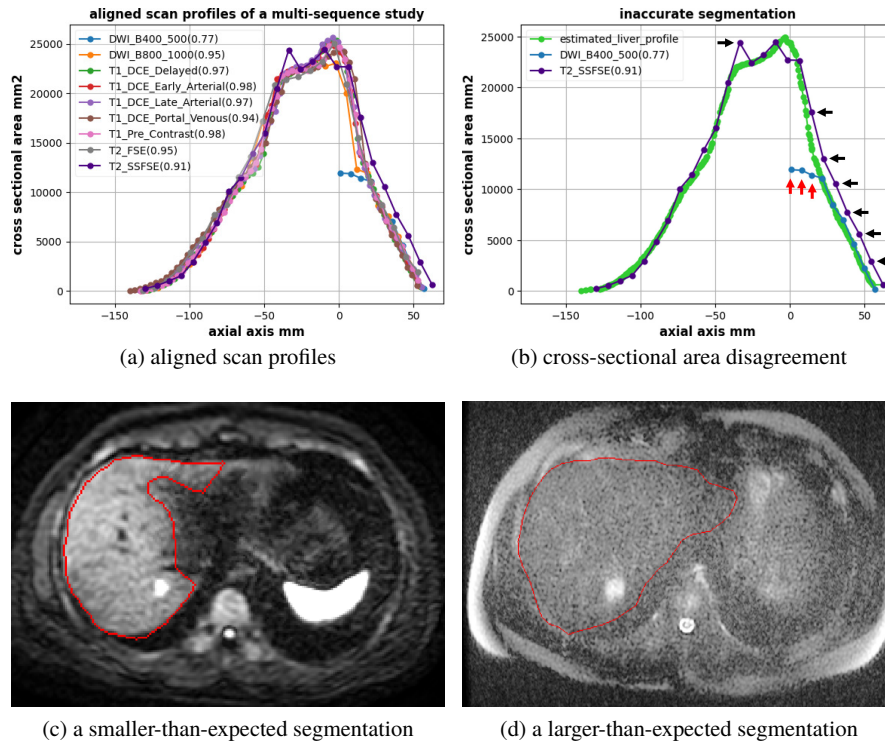


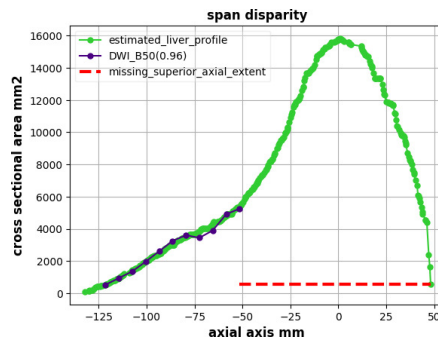
Fig. 3: (a) the aligned scan profiles of a study. A significant cross-sectional area disagreement between a scan profile and its estimated liver profile at the same axial axis location indicates a segmentation error. (b) The blue (DWI\_B400\_500) scan profile identifies a few slices with smaller than expected liver cross-sectional area (pointed by the red arrows) whereas the indigo (T2\_SSFSE) scan profile suggests over-segmentation in multiple slices (pointed by the black arrows). For example, (c) (resp. (d)) shows a detected slice with too small (resp. too large) a segmented area. Bad image quality seems to be responsible for the error in (d).

its corresponding estimated liver profile at the same axial axis location. It resulted in 28458 sample values whose cumulative distribution function is shown in Fig.2b. As expected, this statistic is concentrated around 1. Too high or too low a ratio thus indicates a segmentation error (see Fig.3).

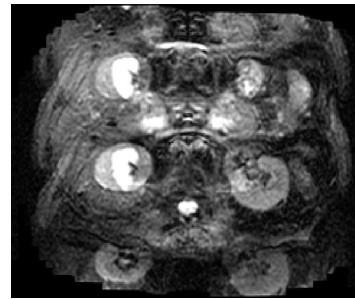
### 3.4 Span disparity

The liver profile also helps identify an additional subset of annotated volumes, which exhibit *span disparity*. It occurs when the liver is only partially observed in the image volume. This can originate from two possible causes. First, the scan's axial range is insufficient to cover the whole liver, leading to a *partial acquisition*. Second, part of the scanned liver fails to be recognized due to measurement errors or poor image quality. This latter results in a *partial segmentation*. Regardless of the cause, our approach allows to estimate such a volume's missing liver portions in a straightforward manner. See Fig.4a for an illustration.

It is also easy to detect an image volume with span disparity because the liver span of its scan profile is much shorter than that of its corresponding estimated liver profile. To differentiate between the two possible causes, the location of the discrepancy matters. If it happens at one end of the volume, with no additional image slices outside the scan profile's support, the cause can be determined to be a partial acquisition. Otherwise, it is a partial segmentation (Fig.4).



(a) span disparity



(b) cause of partial segmentation

Fig. 4: (a) shows the span disparity of a scan profile with respect to its estimated liver profile. The red line represents the axial extent of the superior liver portion missing from this volume. Additional image slices do exist to the right the scan profile' support. But they suffer from severe artefacts and were not marked by the radiologists. (b) shows one of these remaining slices.

The span disparity does not need to result in a low coherence score (Fig.4a). It is independent of cross-sectional area disagreement in that one does not necessarily entail the other. But both can happen at the same time, too.



## 4 Conclusion

In this paper, based on the concept of liver profile, we have presented an exploratory data analysis approach to liver segmentation quality control for multi-sequence MR liver studies. Our method is efficient and allows to locate inaccurately segmented image slices.

Due to its mild assumptions, this method may also carry over to the analysis of segmented liver contours arising from multi-phase CT or longitudinal studies. Furthermore, it may also be applicable to assessing the segmentation quality of other anatomies in a similar context.

## References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging* **38**(8), 1788–1800 (2019)
2. Barrett, H.H., Myers, K.J., Hoeschen, C., Kupinski, M.A., Little, M.P.: Task-based measures of image quality and their relation to radiation dose and patient risk. *Physics in Medicine & Biology* **60**(2), R1 (2015)
3. Bilic, P., Christ, P.F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.W., Han, X., Heng, P.A., Hesser, J., et al.: The liver tumor segmentation benchmark (LiTS). *arXiv preprint arXiv:1901.04056* (2019)
4. Chlebus, G., Schenk, A., Moltz, J.H., van Ginneken, B., Hahn, H.K., Meine, H.: Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing. *Scientific reports* **8**(1), 1–7 (2018)
5. DeVries, T., Taylor, G.W.: Leveraging uncertainty estimates for predicting segmentation quality. *arXiv preprint arXiv:1807.00502* (2018)
6. Esses, S.J., Lu, X., Zhao, T., Shanbhogue, K., Dane, B., Bruno, M., Chandarana, H.: Automated image quality evaluation of T2-weighted liver MRI utilizing deep learning architecture. *Journal of Magnetic Resonance Imaging* **47**(3), 723–728 (2018)
7. Gotra, A., Sivakumaran, L., Chartrand, G., Vu, K.N., Vandenbroucke-Menu, F., Kauffmann, C., Kadoury, S., Gallix, B., de Guise, J.A., Tang, A.: Liver segmentation: indications, techniques and future directions. *Insights into imaging* **8**(4), 377–392 (2017)
8. Hoebel, K., Andrearczyk, V., Beers, A., Patel, J., Chang, K., Depeursinge, A., Müller, H., Kalpathy-Cramer, J.: An exploration of uncertainty information for segmentation quality assessment. In: *Medical Imaging 2020: Image Processing*, vol. 11313, p. 113131K. International Society for Optics and Photonics (2020)
9. Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C.M., Emberton, M., et al.: Weakly-supervised convolutional neural networks for multimodal image registration. *Medical image analysis* **49**, 1–13 (2018)
10. Kohlberger, T., Singh, V., Alvino, C., Bahlmann, C., Grady, L.: Evaluating segmentation error without ground truth. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 528–536. Springer (2012)
11. Ledenius, K., Svensson, E., Stålhammar, F., Wiklund, L.M., Thilander-Klang, A.: A method to analyse observer disagreement in visual grading studies: example of assessed image quality in paediatric cerebral multidetector CT images. *The British journal of radiology* **83**(991), 604–611 (2010)
12. Li, Z., Zhang, S., Zhang, J., Huang, K., Wang, Y., Yu, Y.: MVP-Net: multi-view FPN with position-aware attention for deep universal lesion detection. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 13–21. Springer (2019)

13. Ma, J.J., Nakarmi, U., Kin, C.Y.S., Sandino, C.M., Cheng, J.Y., Syed, A.B., Wei, P., Pauly, J.M., Vasanaawala, S.S.: Diagnostic image quality assessment and classification in medical imaging: Opportunities and challenges. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 337–340. IEEE (2020)
14. Roy, A.G., Conjeti, S., Navab, N., Wachinger, C.: Inherent brain segmentation quality control from fully convnet Monte Carlo sampling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 664–672. Springer (2018)
15. Sujit, S.J., Coronado, I., Kamali, A., Narayana, P.A., Gabr, R.E.: Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks. *Journal of Magnetic Resonance Imaging* **50**(4), 1260–1267 (2019)
16. Valindria, V.V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B.: Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. *IEEE transactions on medical imaging* **36**(8), 1597–1606 (2017)
17. Vorontsov, E., Tang, A., Pal, C., Kadoury, S.: Liver lesion segmentation informed by joint liver segmentation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1332–1335. IEEE (2018)
18. Yan, K., Wang, X., Lu, L., Summers, R.M.: DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging* **5**(3), 036501 (2018)