A Logarithmic Opinion Pool Based STAPLE Algorithm For The Fusion of Segmentations With Associated Reliability Weights

Alireza Akhondi-Asl*, Lennox Hoyte†, Mark E. Lockhart‡, Simon K. Warfield*, *Senior Member IEEE*

* Computational Radiology Laboratory, Department of Radiology,
Children's Hospital, 300 Longwood Avenue, Boston, MA, 02115, USA
E-mail: {Simon.Warfield,Alireza.Akhondi-Asl}@childrens.harvard.edu
† Department of Obstetrics and Gynecology,University of South Florida,
2 Tampa General Circle, 6th floor, Tampa, FL 33606, USA
E-mail: lhoyte@health.usf.edu
‡ Department of Radiology, University of Alabama at Birmingham,
1802 6th Avenue South, Birmingham, AL 35233, USA
E-mail: mlockhart@uabmc.edu

***Abstract*—Pelvic floor dysfunction is very common in women after childbirth and precise segmentation of magnetic resonance images (MRI) of the pelvic floor may facilitate diagnosis and treatment of patients. However, because of the complexity of the structures of pelvic floor, manual segmentation of the pelvic floor is challenging and suffers from high inter and intra-rater variability of expert raters. Multiple template fusion algorithms are promising techniques for segmentation of MRI in these types of applications, but these algorithms have been limited by imperfections in the alignment of each template to the target, and by template segmentation errors. In this class of segmentation techniques, a collection of templates is aligned to a target, and a new segmentation of the target is inferred. A number of algorithms sought to improve segmentation performance by combining image intensities and template labels as two independent sources of information, carrying out decision fusion through local intensity weighted voting schemes. This class of approach is a form of linear opinion pooling, and achieves unsatisfactory performance for this application. We hypothesized that better decision fusion could be achieved by assessing the contribution of each template in comparison to a reference standard segmentation of the target image and developed a novel segmentation algorithm to enable automatic segmentation of MRI of the female pelvic floor. The algorithm achieves high performance by estimating and compensating for both imperfect registration of the templates to the target image and template segmentation inaccuracies. The algorithm is a generalization of the STAPLE algorithm in which a reference segmentation is estimated and used to infer an optimal weighting for fusion of templates. A local image similarity measure is used to infer a local reliability weight, which contributes to the fusion through a novel logarithmic opinion pooling. We evaluated our new algorithm in comparison to nine state-of-the-art segmentation methods by comparison to a reference standard derived from repeated manual segmentations of each subject image and demonstrate our algorithm achieves the highest performance. This automated segmentation algorithm is expected to enable widespread evaluation of the female pelvic floor for diagnosis and prognosis in the future.**

I. Introduction

Pelvic floor dysfunction arising from the damage to pelvic muscle, nerve, and connective tissue is very common in women after childbirth. Urinary incontinence, pelvic organ prolapse, fecal incontinence, and defecatory dysfunction are some of the post-partum pelvic floor symptoms [1], [2], [3], [4], [5]. Typically, characterization of these symptoms is performed by evaluation of structural and anatomical properties of organs and tissues of the pelvis in MR images [6], [7], [8].

MR based three dimensional reconstruction has been used successfully to evaluate the female pelvic floor muscles and tissues in women with and without pelvic floor dysfunction [9], [10]. Furthermore, MR based 3D reconstructed models have been used to generate finite-element and element free computational models suitable for simulating vaginal child-birth [4], [11], offering insight into risk factors for childbirth related pelvic floor injury. Finite-Element modeling has also been used to evaluate anterior vaginal wall support, and the mechanisms underlying cystocele formation [12]. The 3D reconstructed models are generated from manually segmented label-maps, which currently require multiple hours of tedious manual segmentation to produce each reconstructed 3D model. This manual segmentation bottleneck limits the number of computational models that can be reliably produced in a timely manner, thereby limiting the number of study subjects available for the kind of statistical comparisons that can potentially lead to clinically meaningful insight. Reliable automatic segmentation has the potential to remove the manual segmentation bottleneck, which would dramatically increase the number of study

simulation models available for analysis and comparison, allowing us to draw meaningful conclusions from observations involving large groups of study subjects. As an example, automatic segmentation can permit comparison of differences in the quantity and distribution on levator ani stretch during childbirth [4].

However, the structures of pelvic floor are very complex which makes manual segmentation of the structures challenging, even for the expert raters. Moreover, manual segmentation suffers from inter and intra-rater variability [6]. Hence, automatic segmentation of these structures is highly desired. Existing segmentation algorithms for pelvic floor structures have been largely semi-automatic and restricted to few anatomical structures such as rectum,vagina, and bladder, because of the complexity of structures such as obturator internus and inter-individual anatomic variability of the pelvic floor [13], [14], [15], [16].

We have developed a novel and powerful multiple-template-based segmentation algorithm to delineate pelvic floor structures with high accuracy. The fusion of multiple templates with a target is an increasingly popular family of techniques for image segmentation. In this family of techniques, a set of template images, each consisting of an image and its segmentation, is aligned to a target image and a weighted combination of the templates is formed to create a segmentation of the target image. Majority voting is the simplest fusion algorithm that can be used for this purpose [17]. However, majority voting treats the contribution of each template to the target segmentation equally, and this is only correct if each is an equally faithful reproduction of the anatomy of the target. However, for each template, both the template segmentation and the alignment of the template to the target contain errors. Therefore, the fusion algorithm should identify these errors and ameliorate their effects upon the segmentation of the target image.

Here, we briefly explain the sources of these errors. Registration can be defined as the maximization of the alignment of the corresponding features of the two images in the space of the topology-preserving maps. Diffeomorphic image registration algorithms guarantee topology-preserving maps by generating smooth and invertible transformation between the two images. The rationale is that there is one true anatomy for every tissue that is the outcome of normal development. Hence, every normal tissue can be diffeomorphically mapped to any other analogous normal tissue. However, in practice, none of the topology-preserving registration algorithms is capable of doing a perfect alignment of the anatomy of the two images and consequently, it is almost impossible to fully align a template to the target image in the registration based segmentation algorithms. There are two main reasons for this principal inadequacy of the topology-preserving registration algorithms and in particular diffeomorphic image registration methods.

First, the relation between the image intensity and the underlying anatomy is complex and imperfect. Medical imaging devices can be employed to characterize spatial distribution of some physical quantity. Nevertheless, there is an imperfect relation between the anatomy of the tissue and the measured physical quantity. This imperfectness is caused by the lack of specificity in MRI due to the similarity of MRI properties of different tissues and also, by problems such as artifacts, noise, intensity inhomogeneities, and partial voluming. Hence, while the anatomies of the corresponding two similar tissues can be mapped with some diffeomorphic maps, their corresponding intensity images may not be diffeomorphically mapped to each other. This means that the templates and the target image have uncaptureable local inter-individual differences in some regions which cannot be captured by the smooth and invertible transformations.

Second, any non-rigid registration algorithm has some parameters which need to be set. In practice, the optimal values of these parameters are unknown and differ for each pair of the images. Also, finding the global optimum of these parameters might not be practically feasible. Therefore, it is not realizable to optimize these parameters perfectly in a realistic registration problem. Moreover, even with the optimal parameters there would be intrinsic errors caused by improper model selection. This causes undesired intrinsic registration errors due to the sub-optimal assignment of the registration parameters.

Imperfect segmentation of the templates is another source of error which is independent of the registration accuracy. The segmentation of a template can be acquired manually or by an automatic segmentation algorithm. In either case, the utilized segmentation may have both systematic and random errors [18], [19], [20]. The template segmentation errors cannot be detected by measuring the local intensity similarity of the template and the target image. In other words, in the presence of a perfect registration of the templates to the target image, all of the templates will have identical intensity similarity to the target image; however, their corresponding segmentations would not be identical due to the errors in the segmentation of the templates. Therefore, a successful multiple-template-based segmentation algorithm should also take these unavoidable errors into account in the fusion process.

Warfield et al. proposed the STAPLE algorithm in which an optimal weighting of the aligned template segmentations is obtained by comparison with an estimate of the desired target segmentation. STAPLE is an expectation-maximization (EM) algorithm which iteratively estimates the target segmentation and the performance of the templates [18]. In each iteration of the EM, performance parameters of the templates are estimated based on comparison to the estimated reference standard segmentation, and these performance parameters are then used to update the reference standard segmentation. A number of generalizations of this approach have been proposed, to explore mechanisms for fusion with the goal of improving the final segmentation quality. In particular, spatially adaptive performance estimation has been found to be important [21], and effective estimation of priors is important [22], [23]. Alternative formulations of the fusion

problem have been explored [24], [25], [22], [20], [26], [21], [27], [28].

The direct use of intensity to weight the templates has been actively investigated by a number of authors [29], [30], [31], [32], [33], [34], [35]. These locally weighted voting methods use local intensity similarity of the templates and the target image to estimate the reliability weight of the templates. The rationale for this method is that visual inspection of the misalignment of the images suggests the presence of local intensity dissimilarities between the target image and the templates in poorly registered anatomical regions. Furthermore, the anatomical differences indicate incompatibility of the labels in the corresponding segmentation image of the aligned template and the target image. Therefore, it is possible to use the local intensity similarity of the templates and the target image to compute a voxelwise reliability weight of the templates.

While locally weighted majority voting methods use the local intensity similarities to estimate the weight of different templates to compensate for the registration errors, their performance regarding template errors is limited to the performance of the majority voting algorithm as intensity and label information are combined as two independent sources of information. In other words, even if the registration is perfect, locally weighted voting methods will be identical to majority voting. On the other hand, STAPLE based methods are able to detect and compensate for both type of the errors since the contribution of each template is assessed in comparison to a reference standard segmentation of the target image.

Here, we hypothesized that direct employment of intensity information in the STAPLE framework improves the accuracy of the estimation of the performance of templates and therefore improves the accuracy of the segmentation of the target image. Intensity information can assist the fusion algorithm for the detection and compensation of the registration errors. Therefore, we use intensity information as a new source of information in the STAPLE algorithm in order to achieve a more accurate and robust fusion algorithm.

Linear and logarithmic opinion pools are two well-known models which can be used to form weighted votes [36], [37]. All previously reported locally weighted voting approaches use the linear opinion pool to integrate reliability weights in the fusion process. In the linear opinion pool, the linear weighted average of the rater probabilities is used as the aggregation method [36]. One of the difficulties associated with the linear opinion pool is that it is typically multi-modal [36]. In addition, it can be relatively insensitive to the choice of weights and it is not externally Bayesian [36]. On the other hand, the logarithmic opinion pool is a model where the linear weighted average of the logarithm of rater probabilities is used as the aggregating method. In contrast to the linear opinion pool, the logarithmic opinion pool is typically unimodal and can be externally Bayesian [36].

Therefore, we have formulated a new decision fusion algorithm that uses a logarithmic opinion pool aggregating model for the fusion of segmentations with associated local intensity similarity based reliability weights. Our contribution is to demonstrate that a logarithmic opinion pool does allow a tractable formulation of this problem, and that it can be solved exactly with a fixed point solution, and it can be solved rapidly using a closed form approximation. Unlike previous linear opinion pooling techniques, the contribution of both intensity and vote are optimally weighted by measurement of the similarity to the evolving reference standard segmentation in our algorithm. The new algorithm that we present here compensates for both template and registration errors in the fusion procedure, which significantly improves the accuracy of the target segmentation. We applied our algorithm to segment the structures of the pelvic floor from MRI. We compare the accuracy of our method to state-of-the-art fusion methods including recently developed STAPLE based methods which use intensity information in the fusion process [38], [39]. Evaluation of the segmentation achieved indicates the superiority of the algorithm compared to the other methods.

The rest of the paper is organized as follows: Section II describes our novel fusion algorithm and the proposed method for the calculation of weights, based on the intensity information of the templates and the target image. Experimental data, procedures, and results are shown in Section III and conclusions are presented in Section IV.

II. Methods

In this section, we describe our novel fusion algorithm which can be used for the estimation of the reference standard segmentation and the segmentation generator performance parameters. In our model, we assume that the template segmentations at each voxel are associated with a reliability weight where the higher weights indicate higher credibility of the segmentations. Through Sections II-A to II-D, we describe our fusion algorithm in the general form, assuming that the reliability weights are known. Then, in Section II-E, we present our approach to calculate reliability weights based on the intensity information of the templates and the target image in the framework of multiple-template-based segmentation. Throughout the text, we use the terms rater and decision as to refer to segmentation generators and segmentations, respectively.

A. Reliability Weight Based Fusion Framework

We assume that there are J raters and their corresponding decision matrix $\mathbf{D} = \{\mathbf{D}_0, \ldots, \mathbf{D}_i, \ldots, \mathbf{D}_N\}$ where $\mathbf{D}_i = \{D_{i1}, \ldots, D_{ij}, \ldots, D_{iJ}\}$ and D_{ij} is the decision by template j at voxel i. The decision at each voxel i can be any of S labels in $\{1, \ldots, s, \ldots, S\}$. The goal is to estimate the hidden ground truth $\mathbf{T} = \{T_1, \ldots, T_i, \ldots, T_N\}$ and the performance parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_j, \ldots, \boldsymbol{\theta}_J\}$ where $\boldsymbol{\theta}_j$ is a matrix of size $S \times S$ and $\theta_{js's} = f(D_{ij} = s'|T_i = s)$. In addition, we define weight matrix $\boldsymbol{\omega} = \{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_i, \ldots, \boldsymbol{\omega}_N\}$ where $\boldsymbol{\omega}_i = \{\omega_{i1}, \ldots, \omega_{ij}, \ldots, \omega_{iJ}\}$ and ω_{ij} is the reliability weight of rater j at voxel i. It

is assumed that $0 \leq \omega_{ij} \leq 1$ where $\omega_{ij} = 1$ indicates the highest credibility. The weights can be normalized such that $\sum_j \omega_{ij} = 1$; however, this is not a requirement.

We are interested in estimation of the hidden reference standard segmentation by maximization of $f(\mathbf{D}, \mathbf{T}|\boldsymbol{\theta}, \boldsymbol{\omega})$, the probability density function (pdf) of the complete data. Since the performance parameters are unknown, we use an EM algorithm to solve our optimization problem. Therefore, given the estimation of the performance parameters at step t, we maximize the following function to estimate the parameters at step $t+1$:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) &= E\left[\log f(\mathbf{D}, \mathbf{T}|\boldsymbol{\theta}, \boldsymbol{\omega})|\mathbf{D}, \boldsymbol{\theta}^t, \boldsymbol{\omega}\right] \\ &= \sum_i \sum_{T_i} \log f(\mathbf{D}_i, T_i|\boldsymbol{\theta}, \boldsymbol{\omega}_i) f(T_i|\mathbf{D}_i, \boldsymbol{\theta}^t, \boldsymbol{\omega}_i) \end{aligned} \tag{1}$$

where we assumed spatial independence of the reference segmentation. In this equation, $W_{si}^t = f(T_i = s|\mathbf{D}_i, \boldsymbol{\theta}^t, \boldsymbol{\omega}_i)$ can be calculated using the Bayes rule:

$$\begin{aligned} W_{si}^t &= f(T_i = s|\mathbf{D}_i, \boldsymbol{\theta}^t, \boldsymbol{\omega}_i) \\ &= \frac{f(\mathbf{D}_i|T_i = s, \boldsymbol{\theta}^t, \boldsymbol{\omega}_i) f(T_i = s)}{\sum_{s'} f(\mathbf{D}_i|T_i = s', \boldsymbol{\theta}^t, \boldsymbol{\omega}_i) f(T_i = s')} \end{aligned} \tag{2}$$

In the next step, we need to define an appropriate pooling model for $f(\mathbf{D}_i|T_i = s, \boldsymbol{\theta}, \boldsymbol{\omega}_i)$ to combine the decision probabilities with the associated reliability weights. We propose to model the differing performance of each template. In addition, we are interested to have less sensitivity to template decisions at voxels with the smaller reliability weight, as rater decisions at those voxels are less credible. We also know that templates are aligned and rate independently of the other templates. To achieve these goals, we use a logarithmic opinion pool approach to model $f(\mathbf{D}_i|T_i = s, \boldsymbol{\theta}, \boldsymbol{\omega}_i)$:

$$\begin{aligned} & f(\mathbf{D}_i|T_i = s, \boldsymbol{\theta}, \boldsymbol{\omega_i}) \\ &= \frac{\prod_j f(D_{ij}|T_i = s, \boldsymbol{\theta}_j)^{\omega_{ij}}}{\sum_{s_1} \cdots \sum_{s_J} \prod_j f(D_{ij} = s_j|T_i = s, \boldsymbol{\theta}_j)^{\omega_{ij}}} \\ &= \frac{\prod_j f(D_{ij}|T_i = s, \boldsymbol{\theta}_j)^{\omega_{ij}}}{\prod_j \sum_{s''} f(D_{ij} = s''|T_i = s, \boldsymbol{\theta}_j)^{\omega_{ij}}} \\ &= \frac{\prod_j \left(\theta_{jD_{ij}s}\right)^{\omega_{ij}}}{\prod_j [\sum_{s''} \left(\theta_{js''s}\right)^{\omega_{ij}}]} \end{aligned} \tag{3}$$

Using this model it is easy to see that:

$$W_{si}^t = \frac{\prod_j \left[\left(\theta_{jD_{ij}s}^t\right)^{\omega_{ij}} \zeta_{jsi}^t\right] \rho_{si}}{\sum_{s'} \prod_j \left[\left(\theta_{jD_{ij}s'}^t\right)^{\omega_{ij}} \zeta_{js'i}^t\right] \rho_{s'i}} \tag{4}$$

where $\rho_{si} = f(T_i = s)$ and $\zeta_{jsi} = \frac{1}{\sum_{n'} \left(\theta_{jn's}\right)^{\omega_{ij}}}$.

In the case of $\omega_{ij} = 0$, template j is not reliable in the decision process at voxel i. Using our model, it can be seen that in this case, $\forall s, \left(\theta_{jD_{ij}s}\right)^{\omega_{ij}} \zeta_{jsi} = \frac{1}{S}$ which means that the template j does not provide any contribution in the estimation of the W_{si}^t.

In the next step, we maximize the function Q with respect to the performance parameters. To this end, using $W_{si}^t = f(T_i = s|\mathbf{D}_i, \boldsymbol{\theta}^t, \boldsymbol{\omega}_i)$ and the logarithmic opinion pool model in Eq. 3, Eq. 1 can be reformulated as:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) &= \sum_{s,i} W_{si}^t \log f(\mathbf{D}_i, T_i = s|\boldsymbol{\theta}, \boldsymbol{\omega}_j) \\ &= \sum_{s,i} W_{si}^t \log \Big(f(\mathbf{D}_i|T_i = s, \boldsymbol{\theta}, \boldsymbol{\omega}_i) f(T_i = s) \Big) \\ &= \sum_{s,i} \left[W_{si}^t \log \left(\prod_j \frac{\left(\theta_{jD_{ij}s}\right)^{\omega_{ij}}}{\sum_{s'} \left(\theta_{js's}\right)^{\omega_{ij}}} \right) + W_{si}^t \log(\rho_{si}) \right] \\ &= \sum_{s,i,j} W_{si}^t \omega_{ij} \log(\theta_{jD_{ij}s}) + \sum_{i,j,s} W_{si}^t \log(\zeta_{jsi}) + K \end{aligned} \tag{5}$$

where $K = \sum_{s,i} W_{si}^t \log(\rho_{si})$ is a constant which does not depend on the optimization parameters. Hence, only $Q'(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \sum_{s,i,j} \left[W_{si}^t \omega_{ij} \log\left(\theta_{jD_{ij}s}\right) + W_{si}^t \log(\zeta_{jsi})\right]$ is related to the performance parameters and can be optimized using the constraint $\sum_{s'} \theta_{js's} = 1$. Using the Lagrange multipliers, it is possible to maximize this expression by formulating the constrained optimization problem:

$$\begin{aligned} 0 = \frac{\partial}{\partial \theta_{js's}} \Big[& \sum_{s,i,j} W_{si}^t \omega_{ij} \log(\theta_{jD_{ij}s}) \\ & + \sum_{i,j,s} W_{si}^t \log(\zeta_{jsi}) + \lambda(\sum_{n'} \theta_{jn's} - 1) \Big] \\ = & \frac{\sum_{i:D_{ij}=s'} W_{si}^t \omega_{ij}}{\theta_{js's}} - \sum_i \frac{W_{si}^t \omega_{ij} \theta_{js's}^{(\omega_{ij}-1)}}{\sum_{n'} \theta_{jn's}^{\omega_{ij}}} + \lambda \\ = & \sum_{i:D_{ij}=s'} W_{si}^t \omega_{ij} - \sum_i \frac{W_{si}^t \omega_{ij} \theta_{js's}^{\omega_{ij}}}{\sum_{n'} \theta_{jn's}^{\omega_{ij}}} + \lambda \theta_{js's} \end{aligned} \tag{6}$$

We can see that:

$$\begin{aligned} 0 &= \sum_{s'} \Big[\sum_{i:D_{ij}=s'} W_{si}^t \omega_{ij} - \sum_i \frac{W_{si}^t \omega_{ij} \theta_{js's}^{\omega_{ij}}}{\sum_{n'} \theta_{jn's}^{\omega_{ij}}} + \lambda \theta_{js's} \Big] \\ &= \sum_i W_{si}^t \omega_{ij} - \sum_i \frac{W_{si}^t \omega_{ij} \sum_{s'} \theta_{js's}^{\omega_{ij}}}{\sum_{n'} \theta_{jn's}^{\omega_{ij}}} + \lambda \sum_{s'} \theta_{js's} \\ &= \sum_i W_{si}^t \omega_{ij} - \sum_i W_{si}^t \omega_{ij} + \lambda \\ &= \lambda \end{aligned} \tag{7}$$

Hence:

$$\frac{\sum_{i:D_{ij}=s'} W_{si}^t \omega_{ij}}{\theta_{js's}} - \sum_i \frac{W_{si}^t \omega_{ij} \theta_{js's}^{(\omega_{ij}-1)}}{\sum_{n'} \theta_{jn's}^{\omega_{ij}}} = 0 \tag{8}$$

Therefore, maximization of Eq. 5 with respect to the performance parameters leads to:

$$\theta_{js's}^{t+1} = \frac{\sum_{i:D_{ij}=s'} \omega_{ij} W_{si}^t}{\sum_i \zeta_{jsi}^{t+1} \omega_{ij} \left(\theta_{js's}^{t+1}\right)^{(\omega_{ij}-1)} W_{si}^t} \tag{9}$$

The fixed point solution of this equation can be achieved through iterative application of the equation and normalization of $\theta_{js's}^{t+1}$ using the constraint $\sum_{s'} \theta_{js's}^{t+1} = 1$. Note

that it is not feasible to use linear opinion pooling in this framework as the maximization step will be intractable. For example, a possible linear opinion pool model is:

$$f(\mathbf{D}_i|T_i=s,\boldsymbol{\theta},\boldsymbol{\omega_i}) = \frac{\sum_j \omega_{ij} f(D_{ij}|T_i=s,\boldsymbol{\theta}_j)}{\sum_j \omega_{ij}} \tag{10}$$

from which we derive:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \sum_{s,i}\left[W_{si}^t \log\left(\frac{\sum_j \omega_{ij} f(D_{ij}|T_i=s,\boldsymbol{\theta}_j)}{\sum_j \omega_{ij}}\right) + W_{si}^t \log(\rho_{si})\right] \tag{11}$$

Optimization of this function using the Lagrange multipliers and the constraint on the performance parameters leads to the following set of equations for j,s, and s':

$$\sum_{i:D_{ij}=s'} \frac{W_{si}^t \omega_{ij}}{\sum_{j'} \omega_{ij'} \theta_{j's's}} = \sum_n \sum_i \frac{W_{si}^t \omega_{ij} \theta_{jns}}{\sum_{j'} \omega_{ij'} \theta_{j'ns}} \tag{12}$$

Unfortunately, Eq. 12 permits no simple solution for the performance parameters of each rater as it does not have a closed form or a fixed point solution. In this section, we introduced our fusion algorithm logarithmic opinion pool based STAPLE (LOP-STAPLE) which generalizes the STAPLE algorithm for the cases where a reliability weight is associated with the decisions. It is clear that if $\forall i,j$ we set $\omega_{ij}=1$, this formulation will be identical to the classic STAPLE algorithm where there is no reliability information. It is also possible to find an approximate closed form solution for the problem which significantly decreases the computational complexity of the problem. Assuming that ω_{ij} is close to 0 or 1, $\zeta_{jsi}^{t+1}\omega_{ij}(\theta_{js's}^{t+1})^{(\omega_{ij}-1)} \approx \omega_{ij}$ which leads to a closed form approximate solution of the problem.

B. Prior Probability of the Reference Segmentation

One of the important parts of the STAPLE algorithm is the prior ρ_{si}. In general, the prior on the reference segmentation can be a global or spatially varying. The simplest way to estimate a spatially varying prior is to use the output of the majority voting of a collection of aligned template segmentations as the prior. Using label information it is possible to find the global and the majority voting based local prior as follows:

$$\rho_{si}^G = \rho_s^G = \frac{\sum_{i,j:D_{ij}=s}}{NJ}$$
$$\rho_{si}^L = \frac{\sum_{j:D_{ij}=s}}{J} \tag{13}$$

where ρ_{si}^G and ρ_{si}^L are global and spatially varying prior, respectively. It is also possible to combine these two priors to get a general form for the prior:

$$\rho_{si} = (1-\lambda)\rho_s^G + \lambda \rho_{si}^L \tag{14}$$

where $0 \le \lambda \le 1$ is a weighting factor that expresses the relative importance of global versus local prior information.

C. MAP Formulation of the Performance Parameters

In the cases that a priori information about the performance of the raters is available; the maximum a posteriori (MAP) estimation of the parameters is desired. Suppose that $P(\boldsymbol{\theta}) = \prod_{j,s,s'} P(\theta_{js's})$ and γ are the probability map of the prior on the performance parameters and the weight of the prior term, respectively. Thus, the MAP formulation of the fusion problem can be written as:

$$Q_{MAP}(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = Q'(\boldsymbol{\theta}|\boldsymbol{\theta}^t) + \gamma \log(P(\boldsymbol{\theta})) = Q'(\boldsymbol{\theta}|\boldsymbol{\theta}^t) + \gamma \sum_j \sum_s \sum_{s'} \log(P(\theta_{js's})) \tag{15}$$

Similar to [22], using the Beta distribution for the rater j and labels s and s' with parameters $\alpha_{js's}$ and $\beta_{js's}$, the prior probabilities are set to $P(\theta_{js's}) = \frac{1}{Z}(\theta_{js's})^{\alpha_{js's}-1}(1-\theta_{js's})^{\beta_{js's}-1}$ where Z is the normalization constant. The model is based on this fact that if a given structure has not been delineated on a local block, it does not mean that the rater has necessarily a poor segmentation performance in that region. In other words, we assume that the performance of the raters is a priori high which can be represented by setting parameters of beta distribution of diagonal and off-diagonal performance parameters to reflect frequently correct and incorrect decision making, respectively. [22]. Finally, optimization of Eq. 15 with the constraint $\sum_{s'} \theta_{js's} = 1$, leads to the following MAP solution:

$$\theta_{js's}^{t+1} = \frac{\sum_{i:D_{ij}=s'} W_{si}^t \omega_{ij} + \gamma\left(\alpha_{js's} - 1 + \theta_{js's}^{t+1} \sum_{n' \neq s'} \frac{\beta_{jn's}-1}{1-\theta_{jn's}^{t+1}}\right)}{\sum_i \zeta_{jsi}^{t+1} \omega_{ij} (\theta_{js's}^{t+1})^{(\omega_{ij}-1)} W_{si}^t + \gamma \sum_{n'} (\alpha_{jn's} + \beta_{jn's} - 2)} \tag{16}$$

The fixed point solution can be achieved using the previously described approach.

D. Markov Random Field Model for Spatial Consistency

The above model suggests that the estimated probability of the true segmentation is independent of the neighboring voxels. However, in some cases, there are strong anatomical relations between spatially related voxels in the target image. To improve the accuracy of the estimation of the reference standard segmentation of the target image, it is possible to use Markov random field (MRF) model to model this relation in a local neighborhood around each voxel. To model this relation, one can use the approach in [18] to update the estimated ground truth:

$$\overline{W}_{si}^t \leftarrow \frac{1}{Z}\exp\left(\ln \rho_{si} + \sum_j \omega_{ij} \ln(\theta_{jD_{ij}s}^t) + \sum_j \ln(\zeta_{jsi}^t) + \sum_m \sum_n J_{sn} \overline{W}_{nm}^t\right) \tag{17}$$

where J_{sn} describes the spatial compatibility of the labels s and n and, Z is the normalization constant and

$m \in \{1, \ldots, N\}$. At each iteration $\overline{W}_{si}^t$ is normalized using the constraint $\sum_s \overline{W}_{si}^t = 1$. This equation is computed iteratively until convergence and the final $\overline{W}_{si}^t$ is the modified version of estimated ground truth W_{si}^t using the mean field estimation.

In the next section, we describe our approach for the calculation of weights using the intensity information of the target image and the templates.

E. Intensity Similarity as the Measure of Reliability of the Decision

In the previous sections, we have developed a general framework for the fusion of decisions when every decision is associated with a weight. As we discussed in Section I, it is known that when a template is correctly aligned to the target image around a voxel, the corresponding warped manual segmentation is not affected by the registration error. Hence, it is possible to use intensity similarity of the templates and the target image as an indicator of the reliability of the decisions. To consider this knowledge in the fusion process, we define the weight as an increasing function of the local intensity similarity of the target image and the template. In other words, we assign higher weight to a decision at a voxel when the intensity similarity between the target image and the template is higher around the voxel. Let us assume that the templates are registered to the target image $\mathbf{G} = \{G_1, \ldots, G_i, \ldots, G_N\}$ where G_i is the intensity of the target image at voxel i. Let $\mathbf{I} = \{\mathbf{I}_1, \ldots, \mathbf{I}_i, \ldots, \mathbf{I}_N\}$ and $\mathbf{I}_i = \{I_{i1}, \ldots, I_{ij}, \ldots, I_{iJ}\}$ where I_{ij} is the intensity of the template j at voxel i.

1) Weight Calculation: The intensity similarity between the target image and templates can be estimated using a function of Mean square error (MSE) or normalized cross correlation (NCC). It is known that $-1 \leq NCC \leq 1$. However, there are two main constraints in our approach: Weights should be between $[0-1]$, and higher intensity similarities should lead to higher weights. Therefore, to generate the weights we use the following approach to map the local NCC values to the $[0-1]$:

$$\begin{aligned} &\varphi_{ij} = \varphi(\mathcal{N}_{G_i}, \mathcal{N}_{I_{ij}}) = \\ &\frac{1}{|\mathcal{N}|} \sum_{k \in \mathcal{N}} \frac{(\mathcal{N}_{G_i}(k) - \mu_{\mathcal{N}_{G_i}})(\mathcal{N}_{I_{ij}}(k) - \mu_{\mathcal{N}_{I_{ij}}})}{\sigma_{\mathcal{N}_{G_i}} \sigma_{\mathcal{N}_{I_{ij}}}} \\ &\omega_{ij} = \frac{1}{1 + e^{-s(\varphi_{ij} - b)}} \end{aligned} \tag{18}$$

In this equation, μ and σ are used for the representation of the mean and standard deviation in the region defined by $\mathcal{N}_i$ around the voxel i. s and b are scale and shift parameters that can be optimized for each specific problem. It is easy to see that ω_{ij} takes the values between zero and one.

2) Adjustment of Decisions and Weights Using Weak Regularization (Weakly Regularized LOP-STAPLE): To improve the registration accuracy of the aligned templates and to overcome parts of the registration errors due to the regularization constraints, we utilize the method of Akselrod-Ballin et al. [40] with a weak regularization where we locally search for the best matching block. To this end, for each block $\mathcal{N}_i$ centered at voxel i in the target image, we search for the most similar block in a neighborhood $\hat{\mathcal{N}}_i$ around the voxel i of each one of the aligned templates and use the adjusted weights and decisions in the LOP-STAPLE formulation:

$$\begin{aligned} &\hat{i}_j = \arg\max_{i' \in \hat{\mathcal{N}}_i} \varphi(\mathcal{N}_{G_i}, \mathcal{N}_{I_{i'j}}) \\ &\hat{\varphi}_{ij} = \varphi(\mathcal{N}_{G_i}, \mathcal{N}_{I_{\hat{i}_j j}}) \\ &\hat{\omega}_{ij} = \frac{1}{1 + e^{-s(\hat{\varphi}_{ij} - b)}} \\ &\hat{D}_{ij} = D_{\hat{i}_j j} \end{aligned} \tag{19}$$

where for the voxel i, $\hat{i}_j$ is the index of the center of the most similar block in the template j, and $\hat{\omega}_{ij}$ and $\hat{D}_{ij}$ are adjusted weight and decision at the voxel i of j-th template, respectively.

III. Results

In this section we validate our approach using both synthetic and real data. In addition, we compare our method with some of the state-of-the-art methods in the literature. First, we use synthetic data for the evaluation of our method to show the benefits of using intensity information in the fusion process.

A. Experiments on Synthetic Data

In the first step, we validated our method using synthetic data. For this purpose, we generated twenty images and their corresponding segmentations of the target image and then compared the output of the segmentation generated by STAPLE, STEPS [38], and LOP-STAPLE (Fig. 1). The templates are generated using the following procedure: Target image and the corresponding segmentation are rotated by $0°, 2°, 4°, 6°, 8°$ counterclockwise with translation of ± 1 pixel in x and y directions which add up to $5 \times 4 = 20$ templates. We also added white Gaussian noise to the templates to have SNR=40dB. Fig.1.a,b, and c show three of the generated templates and the ground truth is shown in Fig.1.d.

The segmentation errors of LOP-STAPLE, its closed form approximation, STEPS and STAPLE are shown in Fig. 1.e-h, respectively. In this simple example, we have shown the effectiveness of the intensity information in the fusion process. It can be seen that STAPLE does not achieve the correct segmentation as it is impossible to detect the correct segmentation using only the label information of the templates. However, weights generated based on the intensity similarity of the target image and templates are very low for the voxels when there is a substantial mismatch between the target image and the templates. Therefore, these voxels have smaller impact on the estimated reference segmentation and performance parameters. This means that the performance of the templates can be estimated with very high accuracy which

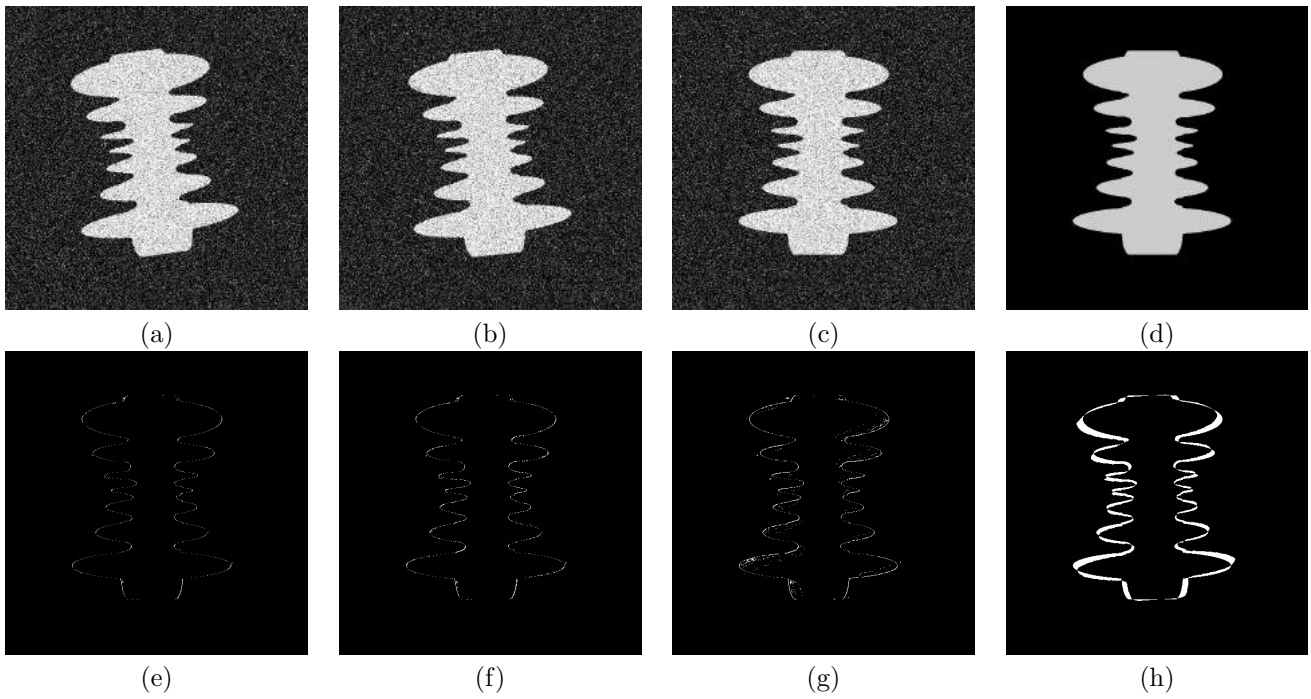

Fig. 1. Illustration of Synthetic Data Segmentation Results. Comparison of segmentation differences between the ground-truth and segmentation generated by LOP-STAPLE (e), its approximate closed form solution (f), STEPS (g), and STAPLE (h) using twenty templates. Three of the generated templates are shown in (a),(b), and (c). The target image is shown in (d). It can be seen that label information alone is insufficient to achieve excellent fusion, as illustrated by the performance of STAPLE in this example. While the STEPS algorithm, which incorporates intensity-based atlas selection, improves performance, the best performance is achieved with LOP-STAPLE, by the use of intensity information through logarithmic opinion pooling. The number of misclassified voxels using the LOP-STAPLE, its closed form approximation, and STEPS are 719, 1081, and 1676, respectively. It can be seen that the closed form approximation has lower accuracy; however, it is significantly faster than the fixed point solution.

leads to a precise segmentation of the target image. In addition, it can be seen that the closed form approximation has inferior accuracy compared to the fixed point solution; however, it is approximately 20 times faster. In this example, label information alone is insufficient to achieve excellent fusion, as illustrated by the performance of STAPLE in this example. The STEPS algorithm incorporates intensity-based atlas selection, which improves performance. The best performance is achieved with LOP-STAPLE, by the use of intensity information through logarithmic opinion pooling. Next, we compare our method to the state-of-the-art fusion algorithms using real MR images of female pelvic data. In the rest of the paper, we utilize the closed form approximation.

B. Imaging Sequence and Structures to Segment

We utilize publicly accessible MRI scans of 20 female pelvises to evaluate the performance of our new label fusion algorithm (http://www.crl.med.harvard.edu/software/STAPLE/index.php). In 18 of these scans, nine structures were segmented three times by three experts as described in [6] which are included in the study. Thus, for each image and for each of the nine structures, nine segmentations were available. The MR images of the pelvic floor were acquired in the axial plane using 1.5 Tesla magnet and phased-array surface coil. The imaging parameters were: T2 Turbo SE axial images with TR/TE=5000/132ms, FOV=200cm, Slice Thickness=3mm/interleaved, no gap, flip angle=180°. For each image the following structures were segmented manually: pelvic bones, symphysis, coccyx, obturator internus, levator ani, vagina, rectum, urethra, and bladder.

C. Reference Segmentation Estimation

Interactive segmentation is well known to demonstrate intra-expert and inter-expert variability. Even carefully established segmentation protocols, and disciplined experts, can generate inconsistent segmentations, especially in areas in which the image contrast is reduced such as at the boundaries between structures of interest. For example, the repeated segmentations of the experts in Fig. 2.d-f have regions which exhibit local intra-rater and inter-rater variability. Therefore, in the first step, we fuse nine manual segmentations of each scan to generate a reference segmentation of the corresponding image. We evaluate, for each reference standard, the overall agreement between the reference standard and the individual input segmentations and choose the most accurate one. Next, in Section III-E, the evaluation of our new fusion algorithm, and some of the state-of-the-art label fusion methods for the purpose of segmentation of a female pelvis MRI scan is presented.

A reference standard segmentation was estimated for each scan with each of the local MAP STAPLE with several half window size (HWS) values [21], STAPLE [18] and majority voting as the label fusion strategies. Note

that locally weighted voting and other intensity based fusion methods cannot be used in this step because the raters use the target image for the manual segmentation. All nine manual segmentations were used for the fusion step of STAPLE and local MAP STAPLE. Then, we computed pair-wise overlap between the reference standard segmentation and experts' segmentations.

We treated each expert segmentation alone as if it was a reference standard segmentation (following a leave-one-out methodology), and computed the pairwise overlap between itself and the other expert segmentations. The coefficient of variation was used as a validation measure. It is defined as the ratio between the standard deviation of the measurements and the absolute value of the average measurements, i.e. $CV = \sigma_D/|\mu_D|$. In our experiments, we computed the coefficient of variation of Dice overlap measures for each of the putative reference standard segmentations (separated for each structure) compared to the segmentation left out.

The best reference standard will have the lowest coefficient of variation of Dice overlap, indicating it is more similar to each of the input segmentations. Furthermore, assessment of the coefficient of variation [41] of Dice overlap between each expert segmentation and other expert segmentations provides a measure of the intra-expert and inter-expert variability. Fig. 2 shows a representative segmentation from an MRI of a female pelvis, the estimated segmentation from local MAP STAPLE, STAPLE and majority voting, and illustrative segmentations from each of the experts. Qualitative comparisons of the MRI and each of the segmentations indicate that expert one and three have the lowest performance. These images illustrate that portions of some structures are not delineated by these two experts, and this is due to the boundary of these structures passing through this slice, leading to incomplete visualization of the 3D extent of these structures. The segmentations generated by STAPLE and by expert number two appear very similar, and the segmentation of local MAP STAPLE appears superior to the other fusion methods. For example, it can be seen that vagina, coccyx, and bladder are delineated more precisely in the segmentation of local MAP STAPLE as compared to the segmentation of expert two. We then sought to quantitatively characterize the average performance differences. We present in Fig. 3 a summary of the average coefficients of variation across the subjects for the reference standard generated by each fusion method, and for each expert.

Fig. 3 indicates that there is inter-rater variability between the segmentations. Comparison of the repeated segmentations of the experts suggests that there is some intra-rater variability as well. The analysis in [6] indicate that the average sensitivity of the raters 1, 2, and 3 are 0.732, 0.783, and 0.650, respectively and the average specificity of them are 0.896, 0.918, and 0.941, respectively which are consistent with the evaluation using COV measure. These results also demonstrate that for HWS of 16, local MAP STAPLE provides the best reference standard estimate. We further visually inspect the fusion result of local MAP STAPLE with HWS=16 for any potential error and consider the result as the reference standard segmentation for the rest of the experiment.

D. Methods Used for Comparison and Selection of Parameters

In the next section, we use the MRI images of female pelvic floor and compare our developed method to the state-of-the-art methods in the literature. To characterize the ability of our new label fusion method to segment these images, we perform a leave-one-out study, considering each pelvis MRI in turn as a target image. For each left-out image, we align each of the remaining subjects' images on it. Next, the output transformations are applied to the generated reference standard segmentations.

Methods of [29], [30], [42] as the intensity similarity based locally weighted fusion methods, and methods of [26], [18], [24], [25] as the STAPLE based methods and methods of [39], [38] as the methods which use intensity information in the STAPLE framework are used for the comparison. For the methods of [26], [24], [25] we have used the software package given in https://masi.vuse.vanderbilt.edu/. We have used the implementation of the STEPS from sourceforge.net/projects/niftyseg/, implementation of MALF-PICSL from http://www.nitrc.org/projects/picsl_malf/, and implementation of NLS from http://www.nitrc.org/projects/jist. In addition, for the STAPLE, we have used the implementation given in http://crl.med.harvard.edu/software/STAPLE/. Also, we have used both LOP-STAPLE and STAPLE with an assigned consensus region which was repeatedly demonstrated to provide excellent performance [21]. In all of the experiments in this manuscript, we have used spatially varying prior with $\lambda = 1$. Finally, we have implemented methods of [29], [30]. In summary, we analyze the segmentation results of the following methods:

- M_0: Majority Voting
- M_1: COLLATE [24]
- M_2: SIMPLE [26]
- M_3: STAPLER [25]
- M_4: Sabuncu et al. algorithm [30]
- M_5: Artaechevarria et al. method [29]
- M_6: STAPLE [18]
- M_7: STEPS [38]
- M_8: Non-Local STAPLE [39]
- M_9: MALF-PICSL [42]
- M_{10}: LOP-STAPLE
- M_{11}: Weakly Regularized LOP-STAPLE

For some of the aforementioned methods, the intensities of the target image and the aligned templates are required to be matched using a normalization approach, otherwise comparison of the intensities of the target image and the templates would not be meaningful. To overcome this problem, the approach in [35] is used to locally normalize

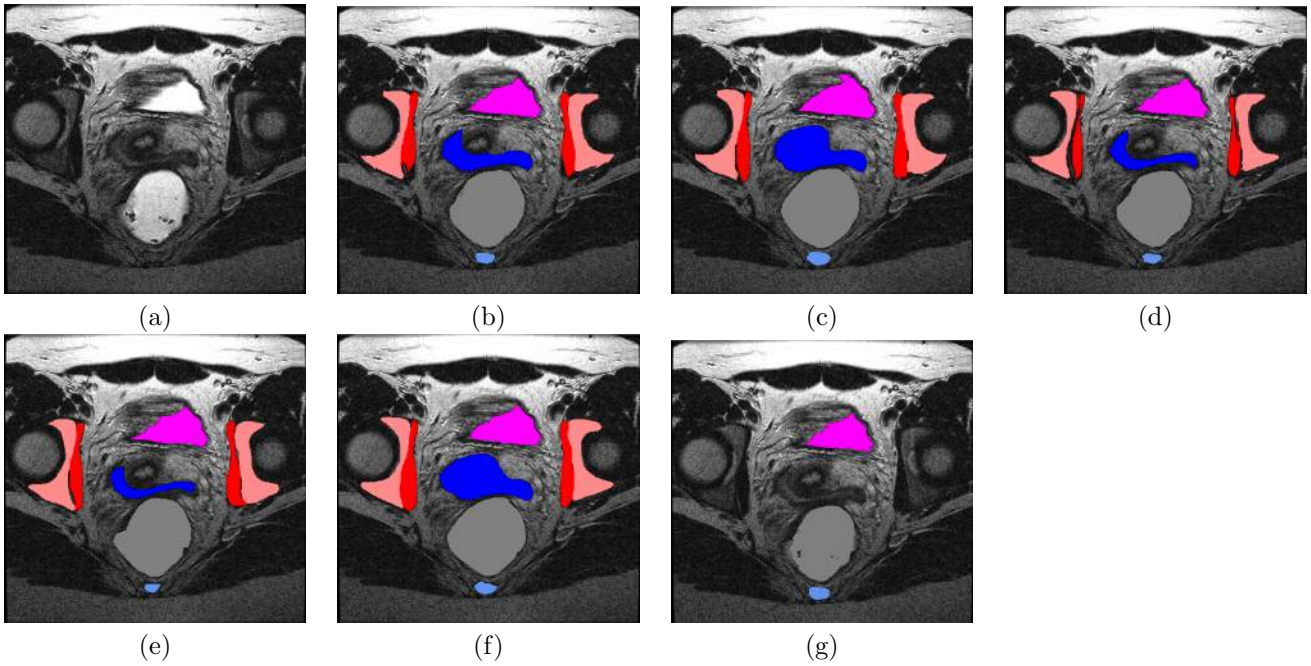

Fig. 2. Illustration of Pelvis Segmentation Fusion. (a): MRI sample slice of female pelvis, (b): estimated segmentation using local MAP STAPLE, (c): using STAPLE, and (d): using majority voting. (e,f,g) illustrate manual segmentations from each one of the experts. Legend : pink, pelvic bones; dark blue, vagina; red, obturator internus; violet, bladder; blue, coccyx; gray, rectum.

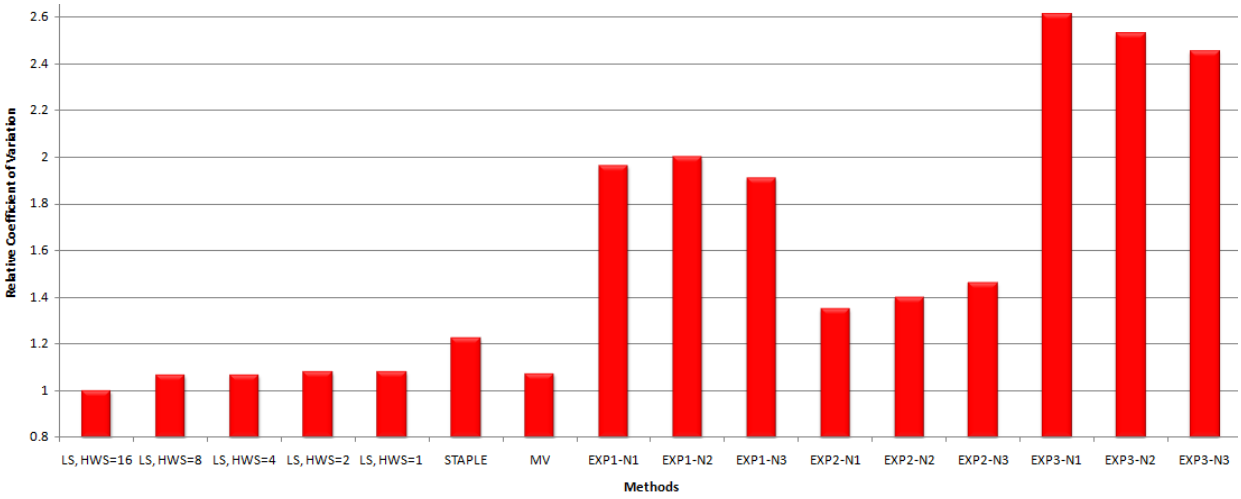

Fig. 3. Quantitative Comparison of Label Fusion of the Female Pelvis Segmentations. Comparison of the coefficient of variation of the Dice similarity coefficient (DSC) achieved by different label fusion algorithms. The lowest coefficient of variation is the best. The ratio of the coefficient of variation for each method to the lowest achieved (local MAP STAPLE with HWS=16) is reported for local MAP STAPLE (L-ST with HWS of 1, 2, 4, 8 and 16), STAPLE, Majority Voting (MV), and expert interactive segmentation (Exj-N,l corresponds to the l^{th} segmentation by expert j).

the intensities by using local mean and variance. We use this strategy for the methods of [29], [30], [39].

We have examined different non-rigid registration algorithms and based on these experiments, we have used the method of [43] for the alignment of the templates to the target image which is a robust block matching algorithm. To find the non-rigid transformation which aligns moving image to the target image, for each resolution and each iteration, first the sparse pairings between the two images are estimated using the block matching approach utilized in [44], [45]. Next, using an interpolation, a dense deformation field is generated where outliers are removed using the strategy described in [43], [40]. This dense nonrigid registration approach is very powerful when large deformation is required for alignment which is the case for our application. The output transformations were applied to the manual segmentation of the templates.

For the methods of [26], [18], [24], [25] we used their default parameters. However, other methods have some parameters related to the intensity information of the templates and the target image which need to be optimized. We have optimized the parameters for these methods by

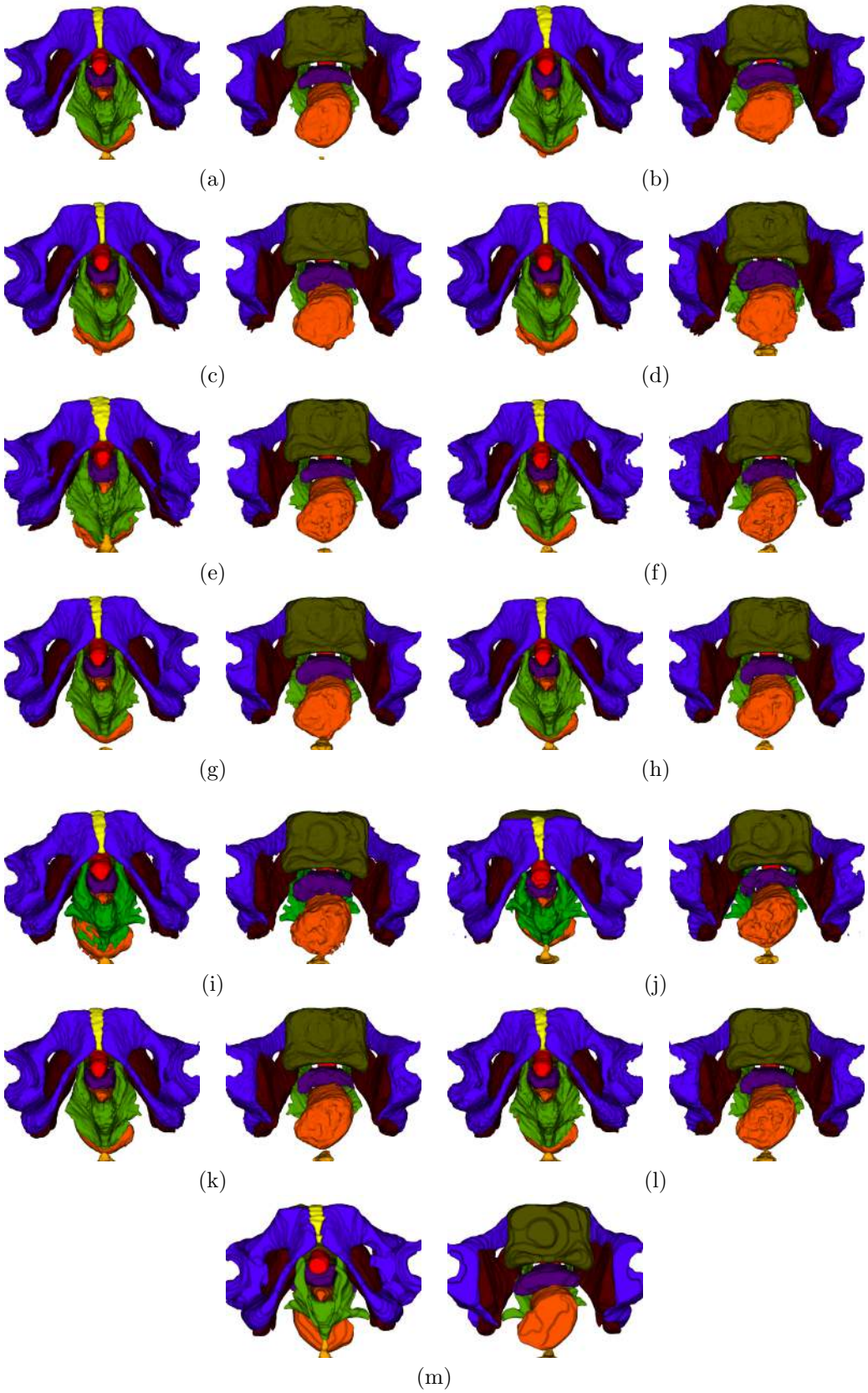

Fig. 4. 3D visualization of segmentations of nine pelvic floor structures. (a) Majority Voting, (b) COLLATE [24], (c) SIMPLE [26], (d) STAPLER [25], (e) Sabuncu et al. algorithm [30], (f) Artaechevarria et al. method [29], (g) STAPLE [18], (h) STEPS [38] , (i) Non-Local STAPLE [39], (j) MALF-PICSL [42], (k) LOP-STAPLE, (l) Weakly Regularized LOP-STAPLE, (m) Manual Segmentation. It can be seen that LOP-STAPLE and Weakly Regularized LOP-STAPLE have higher accuracy compared to the other fusion algorithms. Legend: blue violet, pelvic bones; indigo, vagina; brown, obturator internus; red, urethra; olive, bladder; dark orange, rectum; green, Levator ani; orange,coccyx; yellow, symphysis.

Method	Parameter	Value	Description
M_4	$\mathcal{N}$	2	Half window size of the block
	β	1	Power of square of intensity differences
M_5	$\mathcal{N}$	2	Half window size of the block
	σ	0.5	Standard deviation of Gaussian Kernel
M_7	$\mathcal{N}$	2	Kernel Size
	N	10	Number of selected Atlases
M_8	$\mathcal{N}$	1	Half window size of the block
	σ_i	0.2	Standard deviation of Gaussian Kernel for Euclidean distance-based decay
	σ_d	2	Standard deviation of Gaussian Kernel for intensity similarity
	$\mathcal{N}_s$	4	Half window size of the local search area
M_9	$\mathcal{N}$	2	Half window size of the block
	β	2	Power of square of intensity differences
	$\hat{\mathcal{N}}_s$	1	Half window size of the local search area
	α	0.1	Optimization Parameter
M_{10}	$\mathcal{N}$	2	Half window size of the block
	s	3	Scale parameter
	b	0.7	Shift parameter
M_{11}	$\mathcal{N}$	2	Half window size of the block
	s	3	Scale parameter
	b	0.7	Shift parameter
	$\hat{\mathcal{N}}_s$	1	Half window size of the local search area

TABLE I
PARAMETERS THAT HAVE BEEN USED FOR DIFFERENT FUSION ALGORITHMS. M_4: SABUNCU ET AL. ALGORITHM [30], M_5: ARTAECHEVARRIA ET AL. METHOD [29], M_6: STAPLE [18], M_7: STEPS [38] , M_8: NON-LOCAL STAPLE [39], M_9:MALF-PICSL [42], M_{10}: LOP-STAPLE, M_{11}: WEAKLY REGULARIZED LOP-STAPLE.

evaluating the average Dice coefficient of 9 structures in 18 datasets. The optimized parameters are described in Table I.

E. Evaluation of Segmentation Performance

In this section, we evaluate LOP-STAPLE, and Weakly Regularized LOP-STAPLE as our new developed methods and methods of majority voting, COLLATE, SIMPLE, STAPLER, Sabuncu et al., Artaechevarria et al. STAPLE, STEPS, Non-Local STAPLE, and MALF-PICSL using the female pelvic data. Some of the methods cannot be used for the multi-category segmentation. Therefore, we segment each structure separately, and use the average Dice coefficient of nine structures for the comparison of the methods. We report in Table II the average (Mean) and standard deviation (STD) of Dice scores. In addition, we have used Wilcoxon signed rank test with significance level of 0.05 for the comparison of the methods.

Since a modified registration can be employed for all of the fusion algorithms, to have a fair comparison, we statistically compare LOP-STAPLE and other implemented methods. In Table II, we report the p-value, the sum of ranks assigned to the differences with positive sign (V), and confidence interval $[C_L C_H]$ associated with each comparison. It can be seen that LOP-STAPLE is statistically superior to the other state-of-the-art methods. Also, it is clear that Weakly Regularized LOP-STAPLE is superior to LOP-STAPLE which indicates that modified registrations of the templates improves the segmentation accuracy of LOP-STAPLE. In addition, confidence intervals calculations show that with 95% confidence, the average Dice coefficients of other methods are smaller than LOP-STAPLE and the average Dice coefficients of LOP-STAPLE is smaller than Weakly Regularized LOP-STAPLE. The accurate delineation of the anatomy of the pelvic floor is critical to understanding changes in volume and shape in development, injury and disease. The LOP-STAPLE algorithm presented here is demonstrated to provide higher performance than previously described state of the art segmentation algorithms. An average Dice coefficient improvement of 3%, over eighteen pelvic MRI and nine structures, is found.

Finally, it can be seen that methods of M_7-M_{11} which use intensity information in the STAPLE framework are superior to other methods. This shows that intensity information can be used to estimate the segmentation of the reference standard with higher accuracy.

Figure 4 shows two 3D views of the automatic segmentation of a sample data using different fusion algorithms. These images show that automatic segmentation of the pelvic floor structures using our novel fusion algorithms is comparable with manual segmentation, which suggests that they can be used for clinical purposes.

We have also analyzed the sensitivity of our algorithm to the shift and scale parameters. To this end, we have changed these parameters in a broad range of acceptable values and computed the average Dice coefficient of 9 structures in 18 subjects. The results in Figure 5 show that

	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}
Pelvic bones	0.762	0.616	0.777	0.819	0.812	0.800	0.807	0.823	0.821	0.828	0.830	0.834
Vagina	0.718	0.556	0.634	0.725	0.725	0.729	0.697	0.730	0.729	0.719	0.742	0.744
Obturator Internus	0.840	0.578	0.757	0.835	0.846	0.847	0.819	0.836	0.832	0.843	0.846	0.846
Rectum	0.708	0.340	0.675	0.711	0.766	0.753	0.777	0.794	0.783	0.771	0.800	0.799
Urethra	0.762	0.497	0.639	0.691	0.784	0.797	0.679	0.746	0.722	0.783	0.783	0.786
Bladder	0.778	0.692	0.739	0.762	0.778	0.786	0.785	0.797	0.775	0.780	0.799	0.803
Levator ani	0.678	0.552	0.613	0.674	0.713	0.709	0.679	0.705	0.714	0.735	0.715	0.724
Coccyx	0.106	0.141	0.348	0.153	0.314	0.184	0.535	0.415	0.560	0.430	0.502	0.560
Symphysis	0.532	0.274	0.588	0.579	0.683	0.650	0.712	0.681	0.694	0.716	0.739	0.733
Average	0.654	0.472	0.641	0.661	0.714	0.695	0.721	0.725	0.737	0.734	0.751	0.759
STD	0.087	0.242	0.096	0.082	0.064	0.070	0.070	0.053	0.070	0.054	0.053	0.052
p<	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3	0.05	1e-3		1e-3
V	171	169.5	171	171	171	171	166.5	171	131.5	171		0
C_H	0.127	0.396	0.139	0.118	0.046	0.068	0.043	0.032	0.028	0.023		-0.005
C_L	0.065	0.143	0.073	0.059	0.025	0.039	0.017	0.017	0.001	0.01		-0.011

TABLE II

AVERAGE DICE COEFFICIENT ON THE SEGMENTATION OF 18 FEMALE PELVISES. M_0: MAJORITY VOTING, M_1: COLLATE [24], M_2: SIMPLE [26], M_3: STAPLER [25], M_4: SABUNCU ET AL. ALGORITHM [30], M_5: ARTAECHEVARRIA ET AL. ALGORITHM [29], M_6: STAPLE [18], M_7: STEPS [38] , M_8: NON-LOCAL STAPLE [39], M_9:MALF-PICSL [42], M_{10}: LOP-STAPLE, M_{11}: WEAKLY REGULARIZED LOP-STAPLE. SINCE A MODIFIED REGISTRATION CAN BE EMPLOYED FOR ALL OF THE FUSION ALGORITHMS, WE STATISTICALLY COMPARE M_{10}: LOP-STAPLE TO THE OTHER IMPLEMENTED METHODS. IN THE TABLE, WE REPORTED P-VALUE, THE SUM OF RANKS ASSIGNED TO THE DIFFERENCES WITH POSITIVE SIGN (V), AND CONFIDENCE INTERVAL $[C_L C_H]$ ASSOCIATED WITH EACH COMPARISON WHICH SHOWS THAT LOP-STAPLE IS STATISTICALLY SUPERIOR TO THE OTHER STATE-OF-THE-ART METHODS. IN ADDITION, IT CAN BE SEEN THAT MODIFIED REGISTRATIONS OF THE TEMPLATES IMPROVES THE SEGMENTATION ACCURACY OF THE FUSION ALGORITHM. CONFIDENCE INTERVALS CALCULATIONS SHOW THAT WITH 95% CONFIDENCE, THE AVERAGE DICE COEFFICIENTS OF OTHER METHODS ARE SMALLER THAN LOP-STAPLE AND THE AVERAGE DICE COEFFICIENTS OF LOP-STAPLE IS SMALLER THAN WEAKLY REGULARIZED LOP-STAPLE.

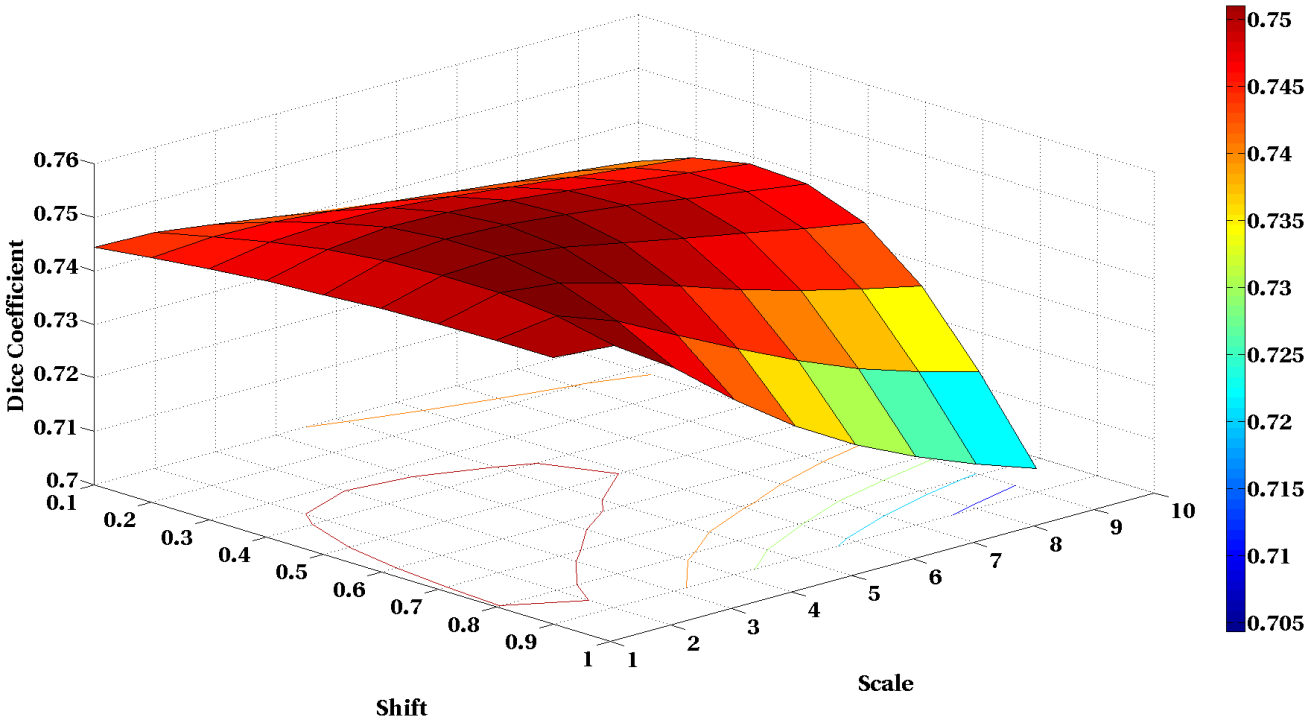

Fig. 5. Sensitivity analysis of the LOP-STAPLE algorithm to the shift and scale parameters of the reliability weights. We have evaluated performance of LOP-STAPLE for different shift and scale parameters by evaluating the average of the Dice coefficient of the 9 structures in 18 subjects. This figure shows that our algorithm is very robust and has very low sensitivity with respect to the change in the parameters in very broad range.

our algorithm has very low sensitivity to the parameters in a broad range of parameters which indicates the robustness of our algorithm.

IV. Conclusions

In this paper, we introduced a novel fusion framework which considers the intensity based reliability weight of the segmentation generators in the fusion process. We have formulated a new decision fusion algorithm that simultaneously assesses both intensity similarity and segmentation labels. A set of template images, consisting of multi-modal MRI and segmentations, are aligned to a target image, and the segmentation of the target image is inferred. The decision fusion algorithm provides both an estimate of the target image segmentation, and a measure of the performance of each template at providing the target image segmentation along with a measure of the confidence interval of this point estimate.

It has been common in prior work to combine intensity and label information as two independent sources of information, and to fuse them by forming an intensity weighted vote. We hypothesized that better decision fusion could be achieved by assessing the contribution of each template in comparison to a reference standard segmentation of the target image. Since the reference standard segmentation of the target image is not available, we formulated an estimation approach in which the reference standard segmentation is estimated, and then performance is assessed in comparison to this estimate of the reference standard segmentation.

Our algorithm is the first to use both intensity and label information simultaneously to compute the performance of each template at predicting the reference standard segmentation. In order to achieve this, we found it necessary to formulate the decision fusion as a type of logarithmic opinion pooling. We demonstrated theoretically and empirically that this new approach leads to a substantial increase in performance as compared to state of the art algorithms. The use of a logarithmic opinion pool does allow a tractable formulation of this problem which can be solved exactly with a fixed point solution or rapidly using a closed form approximation.

For validation, we used publicly available 18 female pelvic MR images where nine structures are manually segmented three times and using three different raters. Pelvic floor dysfunction is very common in women after childbirth and precise segmentation of pelvic floor tissues is crucial for the diagnosis and treatment of pelvic floor symptoms. However, the number of automatic pelvic floor tissue segmentation methods is very limited due to the complexity of the structures and background tissues and also their large variability.

We used local intensity similarity of the templates and the target image as an indicator of reliability of the templates and employed them to calculate the weights. We compared our method to the nine state-of-the-art fusion algorithms and showed that our novel fusion segmentation algorithm is statistically superior to other methods. We have eliminated major confounding factors such as registration, intensity normalization, and parameter optimization procedure by utilizing a similar approach for all of the fusion algorithms; however, there are factors such as the intensity similarity function which may have an impact on the performance of the fusion algorithms. Our novel multiple-template-based pelvic floor segmentation algorithm can generate highly accurate segmentation of the structures comparable with manual segmentation which suggests that it can be used for clinical purposes.

Finally, we have used intensity information to generate the weights; however, in general, any meaningful approach can be used for this purpose. Future work will focus on developing a local version of our algorithm using the concepts described in [21] to further improve the segmentation accuracy.

V. Acknowledgment

This research was supported in part by NIH grants R01 EB013248, R01 LM010033, R01 NS079788, R42 MH086984, P30 HD018655, R03 EB008680 and by a research grant from Boston Children's Hospital Translational Research Program. We would like to acknowledge the Eunice Kennedy Shriver National institute of Human Development sponsored Pelvic Floor Disorders Network as the source of the images used in this study.

References

[1] R. C. Bump and P. A. Norton, "Epidemiology and natural history of pelvic floor dysfunction," *Obstetrics and gynecology clinics of North America*, vol. 25, no. 4, pp. 723–746, 1998.

[2] R. C. Bump, A. Mattiasson, K. Bø, L. P. Brubaker, J. O. L. DeLancey, P. Klarskov, B. L. Shull, and A. R. B. Smith, "The standardization of terminology of female pelvic organ prolapse and pelvic floor dysfunction," *American journal of obstetrics and gynecology*, vol. 175, no. 1, pp. 10–17, 1996.

[3] R. Bassaly, N. Tidwell, S. Bertolino, L. Hoyte, K. Downes, and S. Hart, "Myofascial pain and pelvic floor dysfunction in patients with interstitial cystitis," *International Urogynecology Journal*, vol. 22, no. 4, pp. 413–418, 2011.

[4] L. Hoyte, M. S. Damaser, S. K. Warfield, G. Chukkapalli, A. Majumdar, D. J. Choi, A. Trivedi, and P. Krysl, "Quantity and distribution of levator ani stretch during simulated vaginal childbirth," *American journal of obstetrics and gynecology*, vol. 199, no. 2, pp. 198–e1, 2008.

[5] L. Hoyte and M. S. Damaser, "Magnetic resonance-based female pelvic anatomy as relevant for maternal childbirth injury simulations," *Annals of the New York Academy of Sciences*, vol. 1101, no. 1, pp. 361–376, 2007.

[6] L. Hoyte, W. Ye, L. Brubaker, J. R. Fielding, M. E. Lockhart, M. E. Heilbrun, M. B. Brown, and S. K. Warfield, "Segmentations of mri images of the female pelvic floor: A study of inter- and intra-reader reliability," *Journal of Magnetic Resonance Imaging*, vol. 33, no. 3, pp. 684–691, 2011.

[7] J. R. Fielding, H. Dumanli, A. G. Schreyer, S. Okuda, D. T. Gering, K. H. Zou, R. Kikinis, and F. A. Jolesz, "MR-based three-dimensional modeling of the normal pelvic floor in women," *American Journal of Roentgenology*, vol. 174, no. 3, pp. 657–660, 2000.

[8] M. E. Lockhart, J. R. Fielding, H. E. Richter, L. Brubaker, C. G. Salomon, W. Ye, C. M. Hakim, C. Y. Wai, A. H. Stolpen, and A. M. Weber, "Reproducibility of dynamic mr imaging pelvic measurements: A multi-institutional study1," *Radiology*, vol. 249, no. 2, pp. 534–540, 2008.

[9] A. A. Rodrigues Jr, R. Bassaly, M. McCullough, H. L. Terwilliger, S. Hart, K. Downes, and L. Hoyte, "Levator ani subtended volume: a novel parameter to evaluate levator ani muscle laxity in pelvic organ prolapse," *American journal of obstetrics and gynecology*, vol. 206, no. 3, pp. 244–e1, 2012.

[10] L. Hoyte, L. Schierlitz, K. Zou, G. Flesh, and J. R. Fielding, "Two-and 3-dimensional mri comparison of levator ani structure, volume, and integrity in women with stress incontinence and prolapse," *American journal of obstetrics and gynecology*, vol. 185, no. 1, pp. 11–19, 2001.

[11] K.-C. Lien, B. Mooney, J. O. DeLancey, and J. A. Ashton-Miller, "Levator ani muscle stretch induced by simulated vaginal birth," *Obstetrics and gynecology*, vol. 103, no. 1, p. 31, 2004.

[12] L. Chen, J. A. Ashton-Miller, and J. O. DeLancey, "A 3d finite element model of anterior vaginal wall support to evaluate mechanisms underlying cystocele formation," *Journal of biomechanics*, vol. 42, no. 10, pp. 1371–1377, 2009.

[13] J. M. Venuti, C. Imielinska, and P. Molholt, "New views of male pelvic anatomy: Role of computer-generated 3d images," *Clinical Anatomy*, vol. 17, no. 3, pp. 261–271, 2004.

[14] X. Li, "Semi-automatic segmentation of normal female pelvic floor structures from magnetic resonance images," Ph.D. dissertation, Cleveland State University, 2009.

[15] Z. Ma, R. N. M. Jorge *et al.*, "A shape guided CV model to segment the levator ani muscle in axial magnetic resonance images," *Medical engineering and physics*, vol. 32, no. 7, pp. 766–774, 2010.

[16] Z. Ma, J. M. R. S. Tavares, R. N. Jorge, and T. Mascarenhas, "A review of algorithms for medical image segmentation and their applications to the female pelvic cavity," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 13, no. 2, pp. 235–246, 2010.

[17] T. Rohlfing, D. B. Russakoff, and C. R. Maurer, Jr., "Expectation maximization strategies for multi-atlas multi-label segmentation," in *Information Processing in Medical Imaging*, vol. 2732, 2003, pp. 210–221.

[18] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 7, pp. 903–921, 2004.

[19] S. K. Warfield, K. H. Zou, and W. M. Wells, "Validation of image segmentation by estimating rater bias and variance," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 366, no. 1874, pp. 2361–2375, 2008.

[20] O. Commowick and S. K. Warfield, "A continuous STAPLE for scalar, vector and tensor images: An application to DTI analysis," *IEEE Transactions on Medical Imaging*, vol. 28, no. 6, pp. 838–846, Jun. 2009.

[21] O. Commowick, A. Akhondi-Asl, and S. K. Warfield, "Estimating a reference standard segmentation with spatially varying performance parameters: Local MAP STAPLE," *Medical Imaging, IEEE Transactions on*, vol. 31, no. 8, pp. 1593–1606, 2012.

[22] O. Commowick and S. K. Warfield, "Incorporating priors on expert performance parameters for segmentation validation and label fusion: a maximum a posteriori STAPLE," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI'10)*, ser. LNCS, vol. 6363, Sep. 2010, pp. 25–32.

[23] A. Akhondi-Asl and S. K. Warfield, "Estimation of the prior distribution of ground truth in the staple algorithm: An empirical bayesian approach," *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pp. 593–600, 2012.

[24] A. J. Asman and B. A. Landman, "Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (COLLATE)," *IEEE Transactions on Medical Imaging*, vol. 30, no. 10, pp. 1779 –1794, 2011.

[25] B. A. Landman, A. J. Asman, A. G. Scoggins, J. A. Bogovic, F. Xing, and J. L. Prince, "Robust statistical fusion of image labels," *IEEE Transactions on Medical Imaging*, vol. 31, no. 2, pp. 512–522, 2011.

[26] T. R. Langerak, U. A. van der Heide, A. N. T. J. Kotte, M. A. Viergever, M. van Vulpen, and J. P. Pluim, "Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE)," *Medical Imaging, IEEE Transactions on*, vol. 29, no. 12, pp. 2000–2008, 2010.

[27] A. J. Asman and B. A. Landman, "Formulating spatially varying performance in the statistical fusion framework," *Medical Imaging, IEEE Transactions on*, vol. 31, no. 6, pp. 1326–1336, 2012.

[28] A. Akhondi-Asl and S. Warfield, "Simultaneous truth and performance level estimation through fusion of probabilistic segmentations." *IEEE transactions on medical imaging*, 2013.

[29] X. Artaechevarria and A. Munoz-Barrutia, "Combination strategies in multi-atlas image segmentation: Application to brain MR data," *Medical Imaging, IEEE Transactions on*, vol. 28, no. 8, pp. 1266–1277, 2009.

[30] M. Sabuncu, B. Yeo, K. Van Leemput, B. Fischl, and P. Golland, "A Generative Model for Image Segmentation Based on Label Fusion," *Medical Imaging, IEEE Transactions on*, no. 99, p. 1, 2010.

[31] E. M. van Rikxoort, I. Isgum, Y. Arzhaeva, M. Staring, S. Klein, M. A. Viergever, J. P. Pluim, and B. van Ginneken, "Adaptive local multi-atlas segmentation: Application to the heart and the caudate nucleus," *Medical image analysis*, vol. 14, no. 1, pp. 39–49, 2010.

[32] J. M. P. Lötjonen, R. Wolz, J. R. Koikkalainen, L. Thurfjell, G. Waldemar, H. Soininen, and D. Rueckert, "Fast and robust multi-atlas segmentation of brain magnetic resonance images," *NeuroImage*, vol. 49, no. 3, pp. 2352–2365, 2010.

[33] P. A. Yushkevich, H. Wang, J. Pluta, S. R. Das, C. Craige, B. B. Avants, M. Weiner, and S. Mueller, "Nearly Automatic Segmentation of Hippocampal Subfields in In Vivo Focal T2-Weighted MRI," *NeuroImage*, vol. 53, no. 4, pp. 1208–1224, 2010.

[34] A. R. Khan, N. Cherbuin, W. Wen, K. J. Anstey, P. Sachdev, and M. F. Beg, "Optimal weights for local multi-atlas fusion using supervised learning and dynamic information(SuperDyn): Validation on hippocampus segmentation," *NeuroImage*, vol. 56, no. 1, pp. 126–139, 2011.

[35] H. Wang, J. Suh, S. Das, J. Pluta, M. Altinay, and P. Yushkevich, "Regression-based label fusion for multi-atlas segmentation."

[36] C. Genest and J. V. Zidek, "Combining probability distributions: A critique and an annotated bibliography," *Statistical Science*, vol. 1, no. 1, pp. 114–135, 1986.

[37] M. Rufo, J. Martín, and C. Pérez, "Log-linear pool to combine prior distributions: A suggestion for a calibration-based approach," *Bayesian Analysis*, vol. 7, no. 2, pp. 411–438, 2012.

[38] M. J. Cardoso, K. Leung, M. Modat, S. Keihaninejad, D. Cash, J. Barnes, N. C. Fox, and S. Ourselin, "STEPS: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation," *Medical Image Analysis*, vol. 17, no. 6, pp. 671 – 684, 2013.

[39] A. J. Asman and B. A. Landman, "Non-local statistical label fusion for multi-atlas segmentation," *Medical Image Analysis*, vol. 17, no. 2, pp. 194 – 208, 2013.

[40] A. Akselrod-Ballin, D. Bock, R. C. Reid, and S. K. Warfield, "Accelerating image registration with the johnson–lindenstrauss lemma: Application to imaging 3-d neural ultrastructure with electron microscopy," *Medical Imaging, IEEE Transactions on*, vol. 30, no. 7, pp. 1427–1438, 2011.

[41] A. G. Bedeian and K. Mossholder, "On the use of the coefficient of variation as a measure of diversity," *Organizational Research Methods*, vol. 3, no. 3, pp. 285–297, 2000.

[42] H. Wang, J. Suh, S. Das, J. Pluta, C. Craige, and P. Yushkevich, "Multi-atlas segmentation with joint label fusion," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 3, pp. 611–623, 2013.

[43] R. O. Suarez, O. Commowick, S. P. Prabhu, and S. K. Warfield, "Automated delineation of white matter fiber tracts with a multiple region-of-interest approach," *NeuroImage*, vol. 59, no. 4, pp. 3690–3700, 2012.

[44] S. Ourselin, R. Stefanescu, and X. Pennec, "Robust registration of multi-modal images: towards real-time clinical applications," *Medical Image Computing and Computer-Assisted Intervention£MICCAI 2002*, pp. 140–147, 2002.

[45] S. Ourselin, A. Roche, S. Prima, and N. Ayache, "Block matching: A general framework to improve robustness of rigid registration of medical images," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2000*. Springer, 2000, pp. 557–566.