Published in final edited form as:

Proc SPIE Int Soc Opt Eng. 2012 February 23; 8314: . doi:10.1117/12.910918.

Generalized Statistical Label Fusion using Multiple Consensus Levels

Zhoubing Xu^{*,a}, Andrew J. Asman^a, and Bennett A. Landman^{a,b}

^aElectrical Engineering, Vanderbilt University, Nashville, TN, USA 37235

^bBiomedical Engineering, Johns Hopkins University, Baltimore, MD 21218

Abstract

Segmentation plays a critical role in exposing connections between biological structure and function. The process of label fusion collects and combines multiple observations into a single estimate. Statistically driven techniques provide mechanisms to optimally combine segmentations; yet, optimality hinges upon accurate modeling of rater behavior. Traditional approaches, e.g., Majority Vote and Simultaneous Truth and Performance Level Estimation (STAPLE), have been shown to yield excellent performance in some cases, but do not account for spatial dependences of rater performance (i.e., regional task difficulty). Recently, the COnsensus Level, Labeler Accuracy and Truth Estimation (COLLATE) label fusion technique augmented the seminal STAPLE approach to simultaneously estimate regions of relative consensus versus confusion along with rater performance. Herein, we extend the COLLATE framework to account for multiple consensus levels. Toward this end, we posit a generalized model of rater behavior of which Majority Vote, STAPLE, STAPLE Ignoring Consensus Voxels, and COLLATE are special cases. The new algorithm is evaluated with simulations and shown to yield improved performance in cases with complex region difficulties. Multi-COLLATE achieve these results by capturing different consensus levels. The potential impacts and applications of generative model to label fusion problems are discussed.

Keywords

Segmentation; Statistical Label Fusion; STAPLE; COLLATE; Consensus Levels; Manual Labeling

1. INTRODUCTION

Segmentation plays a critical role in both scientific and clinical inquiry. Particularly, segmentation is critical for clinical diagnosis, treatment and assessment as well as understanding the complex relationships between biological structure and function. For some classes of problems, fully-automated segmentation techniques have been shown to provide consistent, high-throughput and accurate segmentations. Despite this success, extensions of fully automated techniques to new classes of problems can be difficult given

*zhoubing.xu@vanderbilt.edu; <http://masi.vuse.vanderbilt.edu>; Medical-image Analysis and Statistical Interpretation Laboratory, Department of Electrical Engineering, Vanderbilt University, Nashville, TN, USA 37235.

imaging and anatomical variability. As a result, fully-manual voxel-by-voxel labeling by experienced anatomical experts has long been considered the gold standard segmentation technique. Unfortunately, even experts often suffer from intra- and inter-rater variability (e.g. on the order of 10–15% by volume for certain classes of problems [1, 2]). A logical extension of manual labeling is to collect segmentations (i.e. “labels”) from multiple sources (i.e. “raters”), and then combine these labels to form a single estimate of the underlying segmentation [3, 4]. In this technique, called statistical label fusion, the research objective becomes developing algorithmic and statistical techniques that optimally combine labels from multiple raters given the stochastic models by which these raters make mistakes.

The simplest way to fuse labels from human observers is through a majority vote [5]. In this technique, the mode label is selected at every voxel. However, majority vote fails to assess both the likelihood of the estimate and the quality of the observed segmentations. As an alternative, statistically driven techniques [6, 7] (e.g. Simultaneous Truth and Performance Level Estimation (STAPLE) [6]) were proposed to optimally combine observed segmentations through incorporating a global estimate of rater performance into the estimation process. Yet, optimality hinges upon accurate modeling of rater behavior, and these techniques do not intrinsically include spatially varying models of rater performance. Recently, we proposed COnsensus Level, Labeler Accuracy and Truth Estimation (COLLATE) [8] to account for existence of “consensus regions” and “confusion regions”. The central tenant of COLLATE is that certain voxels are harder to label than other voxels; hence, more difficult voxels should be weighted more highly in order to get an accurate estimates of rater performance. This technique was shown to provide significant benefits in certain labeling scenarios over the traditional STAPLE model of rater behavior.

Despite the initial success of COLLATE on the fusion of human raters, only a derivation of the binary consensus level case (subsequently referred to as Binary-COLLATE) has actually been presented. Herein, we extend the generality of COLLATE by reformulating the generative model of human behavior and derive COLLATE using any number of consensus levels, subsequently referred to as Multi-COLLATE. In empirical situations where complex and highly variable structure is annotated, the model by which a human would label these images would almost certainly be more complex than the simple Binary-COLLATE model. Through extension to multiple consensus levels, the derivation opens a near infinite number of possible techniques for modeling human labeling behavior, ranging from complex to simple (see Figure 1). Additionally, we show that the commonly used fusion techniques, such as Majority Vote, STAPLE, and Binary-COLLATE are special cases of the newly presented generative model. We evaluate the benefits of using multiple consensus levels on two simulations, one modeling where raters simply include/exclude regions and on using a four consensus level human rater model.

2. THEORY

The following derivation closely follows with the derivations presented in [6, 8].

2.1 Problem Definition

Consider an image with N voxels with the task of determining the correct label for each voxel in that image. Consider a collection of R raters that provide an observed delineation of all N voxels exactly once. The set of labels, $\mathbf{L}$, represents the set of possible values that a rater can assign to all N voxels. Let $\mathbf{D}$ be an $N \times R$ matrix that indicates the label decisions of the R raters at all N voxels where each element $D_{ij} \in \{0, 1, \dots, L-1\}$. Let $\mathbf{T}$ be the N -vector latent true segmentation that we seek to estimate, where $T_i \in \{0, 1, \dots, L-1\}$. In addition to the traditional human-rater label fusion model, we augment the true segmentation, $\mathbf{T}$, with the “consensus level vector”, $\mathbf{C}$, where $C_i \in \{0, 1, \dots, F-1\}$. The *consensus level* of a voxel indicates the overall difficulty of that voxel compared to the other voxels, where $C_i = 0$ indicates that there is no consensus (or high confusion) about voxel i and $C_i = F-1$ indicates that voxel i is in full consensus. A demonstration of this concept can be seen in Figure 1. Lastly, the R raters’ performance are characterized by $\boldsymbol{\theta}$, where each element, θ_j , is an $L \times L$ confusion matrix where the s', s element in the matrix quantifies the probability that rater j will assign label s' to a voxel when the true label is s in a region of no consensus. For reference, the perfect rater would have a confusion matrix equal to the identity matrix.

2.2 Derivation of Multi-COLLATE

Accurately estimating and understanding the stochastic model that governs the way in which raters make mistakes is essential to accurate and robust label fusion. In addition to augmenting the true segmentation with the consensus level vector, we introduce a novel generative model that estimates the stochastic model of rater behavior

$$f(D_{ij}|T_i=s, C_i=\varsigma, \boldsymbol{\alpha}, \theta_j^{(k-1)}) \equiv \frac{\alpha_{js\varsigma} \delta(D_{ij}, s) + \theta_{js's}}{\alpha_{js\varsigma} + 1} \quad (1)$$

where $\delta(\cdot, \cdot)$ is the kronecker delta function, and $\alpha_{js\varsigma} \in \mathbb{R}^+$ is a free parameter that controls the impact of consensus level ς on the performance level parameters, θ_j , for rater j and label s . To illustrate this, a value of $\alpha_{js\varsigma} = 0$ indicates that a voxel will have a full impact on θ_j and a value of $\alpha_{js\varsigma} = \infty$ indicates that this voxel will have no impact on θ_j . In order to maintain the theoretical interpretation of increasing consensus, we enforce that $\alpha_{js\varsigma} > \alpha_{js\varsigma'}, \forall \varsigma > \varsigma'$.

Now, as with [6, 8] we implement an Expectation-Maximization (EM) algorithm to estimate the true segmentation and the performance level parameters. First, in the E-step, we estimate the probability that each voxel has each of the possible labels and is a member of each of the possible consensus levels

$$W_{si\varsigma}^{(k-1)} \equiv f(T_i=s, C_i=\varsigma | \mathbf{D}, \boldsymbol{\alpha}, \boldsymbol{\theta}^{(k-1)}) = \frac{f(C_i=\varsigma | T_i=s) f(T_i=s) \prod_j f(D_{ij}|T_i=s, C_i=\varsigma, \boldsymbol{\alpha}, \theta_j^{(k-1)})}{\sum_{\varsigma'} \sum_{s'} f(C_i=\varsigma' | T_i=s') f(T_i=s') \prod_j f(D_{ij}|T_i=s', C_i=\varsigma', \boldsymbol{\alpha}, \theta_j^{(k-1)})} \quad (2)$$

where $W_{si\varsigma}$ is the estimated probability that voxel i has associated label s and consensus level ς for the k^{th} iteration of the EM algorithm. The generative model presented in (1) is then appropriately plugged into (2).

In the M-step we find the performance level parameters that maximize the conditional expectation of the log likelihood function, given in (1) and (2). Initially we perform simplifications to the problem in order to estimate the parameters

$$\begin{aligned}\theta_j^{(k)} &= \arg \max_{\theta_j} \sum_i E[\ln f(D_{ij}|T_i) \\ &= s, C_i \\ &= \varsigma, \alpha, \theta_j^{(k-1)}] | D, \alpha, \theta_j^{(k-1)}] = \arg \max_{\theta_j} \sum_{s'} \sum_{i: D_{ij}=s'} \sum_{\varsigma} \sum_s W_{si\varsigma} \ln [f(D_{ij}=s'|T_i=s, C_i=\varsigma, \alpha, \theta_j^{(k-1)})] \quad (3) \\ &= \arg \max_{\theta_j} \sum_{s'} \sum_{i: D_{ij}=s'} \sum_{\varsigma} \sum_s W_{si\varsigma} \ln \left[\frac{\alpha_{js\varsigma} \delta(D_{ij}, s) + \theta_{js's}}{\alpha_{js\varsigma} + 1} \right].\end{aligned}$$

Next, we maximize the parameters by taking the partial derivative of the equation presented in (3), setting it equal to zero and using a Lagrange multiplier with the constraint that $\sum_s \theta_{js's} = 1$.

$$\begin{aligned}0 &= \frac{\partial}{\partial \theta_{jn'n}} \left[\sum_{s'} \sum_{i: D_{ij}=s'} \sum_{\varsigma} \sum_s W_{si\varsigma} \ln \left[\frac{\alpha_{js\varsigma} \delta(D_{ij}, s) + \theta_{js's}}{\alpha_{js\varsigma} + 1} \right] \right. \\ &\quad \left. + \lambda \left(\sum_{s'} \theta_{js's} - 1 \right) \right] \quad (4) \\ &= \sum_{\varsigma} \sum_{i: D_{ij}=n'} W_{ni\varsigma} \frac{1}{\theta_{jn'n} + \alpha_{js\varsigma} \delta(D_{ij}, n)} + \lambda.\end{aligned}$$

This leaves us with a solution for θ_j that can be broken up into two components, (i) the off-diagonal component and (ii) the on-diagonal component. The off-diagonal ($\theta_{jn'n, n \neq n'}$) component can be simplified to

$$\theta_{jn'n, n \neq n'} = \frac{\sum_{\varsigma} \sum_{i: D_{ij}=n'} W_{ni\varsigma}}{-\lambda}. \quad (5)$$

The on-diagonal component, on the other hand, needs to be solved through a 1-dimensional searching procedure given two unknowns, $\theta_{jn'n, n=n'}$ and λ , and the constraint that $\sum_s \theta_{js's} = 1$.

$$-\lambda = \sum_{\varsigma} \frac{1}{\theta_{jn'n, n=n'} + \alpha_{jn\varsigma}} \left(\sum_{i: D_{ij}=n'} W_{ni\varsigma} \right). \quad (6)$$

In the interest of brevity, the discussion of the priors from the E-step, $f(C_i = \varsigma | T_i = s)$ and $f(T_i = s)$, and initialization and convergence can be found in [8], where the implementation in this presentation is essentially equivalent. Additionally, in this paper, only unique values for $\alpha_{js\varsigma}$ are considered for the individual consensus levels, thus, the notation is simplified to α_{ς} . Above, we present generalized theory for Multi-COLLATE, where any number of

consensus levels is possible and allowed. Below, we consider three specific cases where commonly used label fusion algorithms fall directly out of the proposed generative model.

2.3 Special Case 1: Majority Vote

If only one consensus level is considered (i.e. $F = 1$) and $\alpha_0 \rightarrow \infty$ then the generative model simplifies to

$$f(D_{ij}=s'|T_i=s, \alpha, \theta_j^{(k-1)}) \approx \delta(s', s). \quad (7)$$

This model of rater behavior is entirely equivalent to performing a majority vote [5, 9].

2.4 Special Case 2: STAPLE

Again, if only one consensus level is considered (i.e. $F = 1$) and $\alpha_0 = 0$ then the generative model simplifies to

$$f(D_{ij}=s'|T_i=s, \alpha, \theta_j^{(k-1)}) = \theta_{js's}. \quad (8)$$

This model of rater behavior is entirely equivalent to performing traditional STAPLE [6, 7].

2.5 Special Case 3: STAPLE Ignoring Consensus Voxels

If two consensus levels are considered (i.e. $F = 2$) and $\{\alpha_0, \alpha_1\} = \{0, \infty\}$ then the generative model simplifies to the implementation seen in [7, 8], where consensus voxels are ignored and only voxels where there is disparity are considered when formulating the performance level parameters. For evaluation purposes, we refer to this algorithm as a special case of Binary-COLLATE.

3. METHODS AND RESULTS

From the generative model presented above we can derive a continuum of fusion schemes by simply adjusting the number of consensus levels (F) and the values for α for each of the corresponding consensus levels. The schemes include, but are not limited to, Majority Vote, STAPLE, STAPLE Ignoring Consensus Voxels, and Binary-COLLATE. For each of the presented simulations, we apply these standard fusion schemes and, additionally, apply a new scheme that is derived from our generative model using more than two consensus levels, referred to as Multi-COLLATE. All simulations were implemented in MATLAB (Mathworks, Natick, MA).

3.1 Controlled Rater Observation Simulation

First, a simulated intensity volume of size $100 \times 100 \times 100$ voxels was created using a 3D Gaussian distribution centered at the origin. An intensity threshold was set to extract the truth label. In order to simulate a more realistic intensity volume six additional Gaussian distributions centered about the volume were added in and around the original, “true distribution” as interference factors. To simulate raters, we constructed a set of 8 raters for which each rater decided whether or not to include or exclude each of the six interference

factors. The raters exhibited unique performance ranging from fully correct (i.e. assigning all interference factors correctly) to fully incorrect (i.e. assigning all interference factors incorrectly). The algorithms considered were Majority Vote, STAPLE, Binary-COLLATE and Multi-COLLATE. Majority Vote, STAPLE and Binary-COLLATE were implemented as described above. Note, that for situations where the appropriate α_c should equal ∞ , a value of 10^5 was actually used for algorithmic stability. For Multi-COLLATE a 3 consensus-level implementation was used with corresponding alpha values of $\{\alpha_0, \alpha_1, \alpha_2\} = \{0, 2, 10^5\}$

The results from the controlled rater observation simulation can be seen in Figure 2. The target intensity, true labels and observed labels can be seen in Figure 2A, 2B, and 2C respectively. The results of the algorithms used in this experiment can be seen in Figure 2D through 2G. Note that Multi-COLLATE is the only algorithm that converges to the correct answer. The energy in each of the 3 consensus levels utilized by Multi-COLLATE can be seen in Figure 2H, 2I and 2J. The standard model of simply ignoring consensus voxels (Binary-COLLATE) is unable to characterize the complexities of the observation set. By introducing a middle consensus level (Figure 2I) the energy is now distributed into two separate levels consisting of non-consensus voxels. The energy in consensus level 2 indicates the particular voxels for which there is significantly less confusion than the voxels seen in consensus level 1. Note that the energy at these voxels is actually probabilistically distributed between consensus level 1 and 2. This simulation, while contrived, demonstrates the ability of the Multi-COLLATE framework to detect complexities that standard fusion algorithms are not capable of detecting.

3.2 Multi-Label, Multi-Consensus Simulation

Second, we constructed a multi-label multi-consensus level simulation. In this simulation, we construct a volume of size $90 \times 90 \times 60$ with four distinct, equally-sized regions with varying numbers of labels per region. For each region, a certain level of difficulty was assigned to each region such that the raters performed consistently better in some regions and consistently worse in other regions. For each region as the amount of confusion increased, the amount of labels also increased, with 2 labels in the full-consensus region and 5 labels in the least consensus region. By constructing four separate regions with various numbers of labels, we ensure that the straightforward Binary-COLLATE model would not be able to model the complexities of the simulation. We used a voxelwise-random rater model with different estimated confusion matrices for each rater in the various regions. These unique confusion matrices allowed us to construct the various levels of consensus in each of the regions. The number of raters (or volumes) was varied between 5 and 15, and 10 Monte Carlo iterations were used for each number of volumes fused. As with the previous simulation we compared the accuracy of Majority Vote, STAPLE, Binary-COLLATE and Multi-COLLATE. The former three algorithms were constructed in the manner described in Section 2. Multi-COLLATE was constructed using 4 consensus levels with corresponding values for $\{\alpha_0, \alpha_1, \alpha_2, \alpha_3\} = \{0, 0.5, 2, 10^5\}$

The results for the multi-label, multi-consensus level simulation can be seen in Figure 3. For all of the images seen in Figure 2A – 2F, 10 raters were used in the estimation process. The

true labels and corresponding observed labels can be seen in Figures 2A and 2B. The results of the four considered algorithms can be seen in Figures 2C – 2F. An accuracy comparison (using fraction voxels correct) of the four considered algorithms can be seen in Figure 2G. Note that as the number of volumes fused increases the improvement seen in Multi-COLLATE improves as well. This is an indication that as more data is available the ability of Multi-COLLATE to accurately estimate the distribution of energy between the various consensus levels. Additionally a two-sided t-test was performed between Multi-COLLATE and all of the other algorithms which resulted in a p-value < 0.001 for all numbers of volumes fused. The distribution of energy between the various consensus levels can be seen in Figures 2H – 2K. Note that, while certainly far from perfect, the energy is distributed between the levels in a logical manner. In other words, the most energy in consensus level 1 (the least consensus) is in the region where the raters make the most mistakes. Likewise, the energy in consensus level 4 is clearly distributed in the region where there is full consensus.

4. DISCUSSION

The ability to segment structures of interest in medical images is of the utmost importance in both clinical and scientific advancement. Neither fully-automated nor fully-manual techniques can uniformly provide robust, accurate, and efficient modeling the complex anatomical variability often seen in medical imaging. Thus, the concept of combining the talents of multiple human observers has become a commonplace for acquiring robust and accurate segmentations [6–8, 10–13]. Previous fusion techniques have relied on over-simplified models of human behavior that are not capable of modeling the diversity of human behavior seen in labeling [6–8]. Our research objective is more accurately construction of models of the complexities and idiosyncrasies with which humans observe labels on medical images.

Herein, we propose a generalized framework for modeling human behavior in which it is possible to characterize models that range from extremely simple to extraordinarily complex. This is accomplished by introducing the concept of *multiple* consensus levels, which capture the varying regional difficulty of labeling voxels. Previously this algorithm, COLLATE, was limited to the binary consensus level case, in which only simple models of rater behavior can be taken into account [8]. Here, we extend this framework to include any number of consensus levels and derive the algorithm in the general case. This derivation opens numerous doors in terms of modeling the varying degree of difficulty of labeling tasks. We note that one of the benefits of this approach is that the most commonly used fusion algorithms (e.g. Majority Vote, STAPLE and Binary-COLLATE) are special cases of the generative model. A potential area of fascinating research would be to optimize over the selection of control parameters to arrive at an optimal fusion method for a particular study design or anatomical target.

In continuing work, the performance of the COLLATE model certainly needs to be validated on larger scale empirical experiments. Unfortunately, the availability of such data is in limited supply, and this remains as necessary future work. Nevertheless, the efforts of groups to “crowd-source” the labeling problem [14, 15] presents interesting opportunities for the COLLATE model to be tested more thoroughly.

There are certainly limitations to the current implementation of the algorithm. First, the free parameters, α , are required to be manually set. Ideally, a technique for integrating these parameters into the estimation process would arise, such that these could be optimally selected based upon the data that is given. Additionally, the model that is presented is fairly restrictive in the sense that it enforces the model that raters exhibit consistent performance in the various consensus levels. In reality, this may not be the case as the clear distinction between the various levels of consensus may not exist. Second, a recently proposed algorithm, Spatial STAPLE [12], introduced the idea of using multiple, conditionally independent estimates of the performance level parameters for each rater. If raters observe labels in a manner that does not comply with the strict consensus level model, the Spatial STAPLE model may be more appropriate. Lastly, it may be ideal to combine COLLATE and Spatial STAPLE into a single technique, where the consensus levels are estimated, but conditionally independent estimates of performance is calculated in each level. Such an approach might lead to a more elegant treatment of varying task difficulty.

ACKNOWLEDGEMENTS

This project was supported by NIH/NINDS 2R01EB006136 and NIH/NINDS 1R21NS064534.

REFERENCES

1. Joe BN, Fukui MB, Meltzer CC, Huang Q, Day RS, Greer PJ, Bozik ME. Brain Tumor Volume Measurement: Comparison of Manual and Semiautomated Methods¹. *Radiology*. 1999; 212(3):811. [PubMed: 10478251]
2. Ashton EA, Takahashi C, Berg MJ, Goodman A, Totterman S, Ekholm S. Accuracy and reproducibility of manual and semiautomated quantification of MS lesions by MRI. *Journal of Magnetic Resonance Imaging*. 2003; 17(3):300–308. [PubMed: 12594719]
3. Kittler J, Hatef M, Duin RPW, Matas J. On combining classifiers. *IEEE Trans. Pattern Anal. Machine Intelligence*. 1998; 20:226–239.
4. Kearns M, Valiant LG. Learning boolean formulae or finite automata is as hard as factoring. *Harvard University Technical Report, TR-14-88*. 1988
5. Kittler J, Hatef M, Duin RPW, Matas J. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1998; 20(3):226–239.
6. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*. 2004; 23(7):903–921. [PubMed: 15250643]
7. Rohlfing T, Russakoff DB, Maurer CR. Performance-Based Classifier Combination in Atlas-Based Image Segmentation Using Expectation-Maximization Parameter Estimation. *IEEE Transactions on Medical Imaging*. 2004; 23:983–994. [PubMed: 15338732]
8. Asman AJ, Landman BA. Robust Statistical Label Fusion through Consensus Level, Labeler Accuracy and Truth Estimation (COLLATE). *IEEE Transactions on Medical Imaging*. 2011 In press.
9. Artaechevarria X, Muñoz-Barrutia A. Combination strategies in multi-atlas image segmentation: Application to brain MR data. *Medical Imaging, IEEE Transactions on*. 2009; 28(8):1266–1277.
10. Sabuncu M, Yeo B, Van Leemput K, Fischl B, Golland P. A Generative Model for Image Segmentation Based on Label Fusion. *IEEE Transactions on Medical Imaging*. 2010; 29(10): 1714–1729. [PubMed: 20562040]
11. Langerak T, van der Heide U, Kotte A, Viergever M, van Vulpen M, Pluim J. Label Fusion in Atlas-Based Segmentation Using a Selective and Iterative Method for Performance Level Estimation (SIMPLE). *IEEE Transactions on Medical Imaging*. 2010; 29(12):2000–2008. [PubMed: 20667809]

12. Asman AJ, Landman BA. Characterizing Spatially Varying Performance to Improve Multi-Atlas Multi-Label Segmentation. *Information Processing in Medical Imaging (IPMI)*. 2011; 6801:81–92.
13. Bogovic, JA.; Landman, BA.; Bazin, P-I; Prince, JL. Statistical Fusion of Surface Labels provided by Multiple Raters, Over-complete, and Ancillary Data. *SPIE Medical Imaging Conference*; San Diego, CA. 2010.
14. Stein, JA.; Asman, AJ.; Landman, BA. Characterizing and Optimizing Rater Performance for Internet-based Collaborative Labeling. Lake Buena Vista, FL: SPIE Medical Imaging; 2011.
15. Asman, AJ.; Scoggins, AG.; Prince, JL.; Landman, BA. Foibles, Follies, and Fusion: Assessment of Statistical Label Fusion Techniques for Web-Based Collaborations using Minimal Training. Lake Buena Vista, FL: SPIE Medical Imaging; 2011.

Figure 1.

Illustration of consensus levels and the algorithms that can be derived from them. The cartoon to the left represents a series of ten experiments (rows) where raters (columns) provide label decisions with carrying degrees of consensus. In the COLLATE framework, we quantify these consensus levels and integrate them into the estimation process. The brain images in the center demonstrate areas where high consensus is expected (top image) and low consensus is expected (bottom image). The right plot shows how the various algorithms estimate performance (θ) based on consensus.

Figure 2.

Results from the controlled rater observation simulation. The target intensity image and the corresponding true labels can be seen (A) and (B), respectively. The set of input observations used in this experiment can be seen in (C). The results of the four algorithms, Majority Vote, STAPLE, Binary-COLLATE and Multi-COLLATE can be seen in (D)–(G). Note that the 3 consensus-level implementation of Multi-COLLATE is the only algorithm that gets the correct answer. The 3 energy in each of the consensus levels can be seen in (H)–(J). Note that by distributing the non-consensus voxels into two separate levels, (H) and (I), the estimation process converges to the correct answer.

Figure 3.

Results from the multi-label, multi-consensus simulation. The true labels and example observations when 10 total raters were fused can be seen in (A) and (B). Note the varying levels of performance in each of the regions. The results of the 4 primary algorithms we are considering can be seen (C)–(F). Note the improved performance of Multi-COLLATE (F). An accuracy comparison using the fraction voxels correct with respect to increasing numbers of raters can be seen in (G). Multi-COLLATE produces results that are statistically significantly better for all numbers of raters. The energy in each of the consensus levels can be seen in (H)–(K). Note that, although approximate, the energy in each of the consensus levels is distributed in a manner that makes intuitive sense given the input observations.