Twofold binary image consensus for medical imaging meta-analysis

C. Lopez-Molina[1,3], J. Sanchez Ruiz de Gordoa[2],
V. Zelaya-Huerta[2], and B. De Baets[3]

[1] Dpto. Automatica y Computacion, Universidad Publica de Navarra,
Pamplona, Spain
carlos.lopez@unavarra.es
[2] NavarraBiomed, Servicio Navarro de Salud/Osasunbidea
Pamplona, Spain
[3] KERMIT, Dept. Data Analysis and Mathematical Modelling
Ghent University, Gent, Belgium

Abstract. In the field of medical imaging, ground truth is often gathered from groups of experts, whose outputs are generally heterogeneous. This procedure raises questions on how to compare the results obtained by automatic algorithms to multiple ground truth items. Secondarily, it raises questions on the meaning of the divergences between experts. In this work, we focus on the case of immunohistochemistry image segmentation and analysis. We propose measures to quantify the divergence in groups of ground truth images, and we observe their behaviour. These measures are based upon fusion techniques for binary images, which is a common example of non-monotone data fusion process. Our measures can be used not only in this specific field of medical imagery, but also in any task related to meta-quality evaluation for image processing, e.g. ground truth validation or expert rating.

Keywords: Data fusion, Twofold Consensus Ground Truth, Meta-analysis, Medical Imagery, Immunohistochemistry (IHC)

1 Introduction

Data fusion pursues rather different goals in very disparate contexts. An common goal is to produce a reduced (compact) representation of a certain amount of data objects. Whichever specific technique the fusion is based upon, and whichever data objects are to be fused, reduction is the main goal in most fusion processes. However, fusion can lead the way to some other subsidiary goals just as interesting as reduction. For example, the result of a fusion process can be used as starting point to study the data to be processed, including its individual and group characteristics. Otherwise said, it can be used for data analysis, specifically to generate metadata (data about data).

The application of data fusion techniques to produce metadata is certainly not novel; in this regard, a relevant example is the standard deviation. The

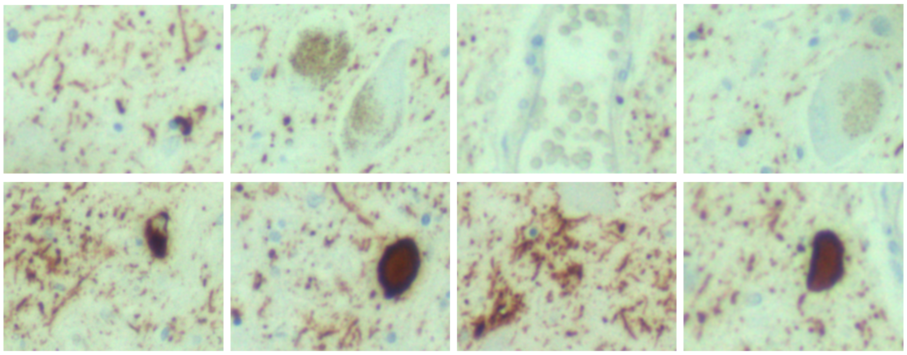


Fig. 1. Subimages extracted from Immunohistochemistry (IHC) images. The upper row displays regions without artifacts affected by tau protein, while the lower row displays artifacts or regions in which the presence of such protein is evident.

arithmetic mean can be seen as procedure to fuse scalar data into a compact representation with minimal loss of information, such loss being measured as the sum of the squared distance to the original values. At the same time, it is also a key to compute the standard deviation, which is a dispersion measurement. Even when dealing with non-Gaussian distribution of values, the standard deviation is used as a feature in data meta-analysis. We believe that principles similar to those by the mean and standard deviation can be ported to scenarios in which monotonicity plays no role. That is, we believe that fusion of non-standard data can also be taken as starting point to produce metadata in non-motonote universes.

In this work we elaborate on images in the context of neurology and neuropathology. This work is part of a research effort on immunohistochemistry (IHC) images for the measurement of deposits of tau protein in patients affected by Progressive Supranuclear Palsy (PSP). In this research effort, expert neuropathologists analyze microscope images of brain tissue and perform manual labelling of the areas affected by tau protein (see Fig. 1). Such binary labelling is further used to perform quantitative measurements with interest for posthumous analysis and disease profiling. Since this process is extremely time-consuming, automatic segmentation methods are being proposed to alleviate the workload of the pathologists. These methods shall be designed to produce results similar to those by expert humans. A key problem found in the evaluation and tuning of such automatic methods is the fact that pathologists often feature severe differences of opinion and/or precision. From a computational point of view, they generate different binary images which shall be taken as ground truth (that is, perfect solutions) for automatic segmentation algorithms. Of course, the multiplicity of ground truth solutions severely hinder the evaluation (and training) of such algorithms. Understanding and evaluating set-based or multivalued ground truth is hence a priority for our applied developments.

We propose to fuse the binary images produced by neuropathologists using the Twofold Consensus Ground Truth (TCGT, [4]). Our approach is rather different to that by other binary image fusion techniques (e.g. [2, 9]), in the sense that we avoid the statistical counting of visual items, and rather focus on the spatial interpretation of coincidences and divergences. The TCGT takes as input a set of binary images and yields a set-valued consensus based on the coincidences and divergences in the input images. The resulting set allows for a compact representation of the input set of images, and also for the quantification of some of its characteristics. In this regard, we attempt to quantify two facets of the ground truth images. Firstly, we intend to quantify heterogeneity of the set of ground truth images, since it could be related to the difficulties faced by neuropathologists in the labelling of the original image. Secondly, we aim to evaluate the dissimilarity (degree of coincidence and divergence) of a ground truth image w.r.t. a group og ground truth images. In a sense, the first question relates to the group dispersion or heterogeneity, while the second one relates to the one-to-many dissimilarity of the images. Note that, although initially designed to elaborate on binary edge images, the TCGT can be ported to scenarios in which binary images hold different semantics.

The remainder of this work is organized as follows. In Section 2 we introduce the idea of weak and strong consensus, together with the Twofold Consensus Ground Truth. The usefulness of this concept is explained in Section 3, in which we develop the application for the meta analysis of IHC ground truth. Finally, Section 4 features some conclusions and future lines of research.

2 Twofold consensus ground truth

2.1 Preliminary Notations

In this work we consider images to have some fixed dimensions $\mathcal{M} \times \mathcal{N}$, so that $\Omega = \{1, \ldots, \mathcal{M}\} \times \{1, \ldots, \mathcal{N}\}$ represents the set of positions in an image. The set of all binary images is denoted $\mathbb{B}$, and can be dually seen as the set of mappings $\Omega \mapsto \{0, 1\}$, or as the power set $\wp(\Omega)$. Individual binary images will be referred to with upper case (e.g. E, I), while bold-faced upper case is reserved for sets of images (e.g. $\mathbf{A} = \{A_1, \ldots, A_n\}$).

In this work we consider positive information in binary images to be represented by 1's, while negative information takes 0's. When it comes to the processing of binary images, we can use a dual signal-logical interpretation of this fact. Hence, apart from image-oriented operators, we use the classical set-theoretic operations on binary images, namely intersection ($\wedge$), union ($\vee$), and inclusion ($\subseteq$, $\subset$). The symbols $\cap$ and $\cup$ are reserved for the intersection and union of sets of images, respectively. According to the reference works on binary image morphology [1, 8], the dilation of a binary image A by some structuring element K is given by $\mathcal{D}_K(A) = \{c \in \Omega \mid c = a + b \text{ for some } a \in A \text{ and } b \in K\}$.

2.2 Strong and weak consensus on binary images

Binary images are a very common format to express the output of image processing tasks, despite being barely useful to represent visual information in human terms. This holds, for example, for object recognition or binary segmentation. The nature and shape of the information in a binary image can greatly diverge from task to task, examples being regions (for object recognition or salient region identification), lines (for boundary detection), points (for critical point detection), etc. In many of such cases there is a need to combine different images, either to fuse ground truth images [9] or to fusion different candidate images generated by different algorithms. In [4] we present a technique for binary image fusion, namely the Twofold Consensus Ground Truth (TCGT). Due to the variable understanding of the term consensus, our technique narrows down its goals to three facts, enunciated as follows:

G1. *Preserving discordances*: The consensus should represent non-unanimous features in the images.

G2. *Highlighting agreement*: The consensus must point out those aspects on which the original images agree, either positively (features appearing at all images) or negatively (those appearing at none).

G3. *Keeping original images as perfect*: As long the input images are the only source of ground truth, the result of the fusion must somehow include them. This guarantees that any automatic method performing exactly as a the sources (probably, humans) is evaluated as perfect.

The TCGT is supported by two different consensus operators, namely the *strong* and *weak consensus.*

Definition 1 *The strong consensus image of a set of binary images* $\mathbf{I} = \{I_1, \ldots, I_k\}$ *is the binary image* $s_T(\mathbf{I})$ *defined as*

$$s_T(\mathbf{I}) = \mathcal{D}_T(I_1) \wedge \mathcal{D}_T(I_2) \wedge \ldots \wedge \mathcal{D}_T(I_k) \ , \tag{1}$$

where $\mathcal{D}_T(I_i)$ *denotes the dilation of image* I_i *using the structuring element* T*.*

Definition 2 *The weak consensus image of a set of binary images* $\mathbf{I} = \{I_1, \ldots, I_k\}$ *is the binary image* $w_T(\mathbf{I})$ *defined as*

$$w_T(\mathbf{I}) = \mathcal{D}_T(I_1) \vee \mathcal{D}_T(I_2) \vee \ldots \vee \mathcal{D}_T(I_k) \ , \tag{2}$$

where $\mathcal{D}_T(I_i)$ *denotes the dilation of image* I_i *using the structuring element* T*.*

The strong and weak consensus of a set of images materialize as the tightest and loosest agreement that can be reached given a set of binary images $\mathbf{I}$. In this sense, they resemble the upper and lower bounds of interval-valued data, or the boundaries of rough sets [6]. Note that their result is influenced by a structuring element T. This element is used, in the present context, to consider the variable position of the same objects when delineated by different experts. The characteristics of T must fit the conditions of the specific problem. For example, if we consider a spatial tolerance of 7 pixels, T might be a disk with radius 7. If the task allows for no tolerance at all, then a radius 1 disk can be used to perform no dilation in the generation of the strong and weak consensus.

2.3 The Twofold-Consensus Ground Truth

From goals G1-G3, is is evident that the consensus must be expressed as a set or multivalued object. Otherwise, it could not allocate the different images we attempt to fuse (as required in G3). The consensus shall not be an image in $\wp(\Omega)$, but a subspace in $\wp(\Omega)$. We seek the set of images which (a) contain all of the positive information in which all ground truth images agree on and (b) does not include positive information not featured by any ground truth image.

Definition 3 *The consensus of a set of binary images* $\mathbf{I}$ *is the set of images* $c_T(\mathbf{I})$ *defined as*

$$c_T(\mathbf{I}) = \{B \in \mathbb{B} \mid B \subseteq w_T(\mathbf{I}) \textit{ and } s_T(\mathbf{I}) \subseteq \mathcal{D}_T(B)\} \ . \tag{3}$$

The consensus set satisfies some practical properties, which we review in Section 2.4. Also, it has some interesting theoretical properties:

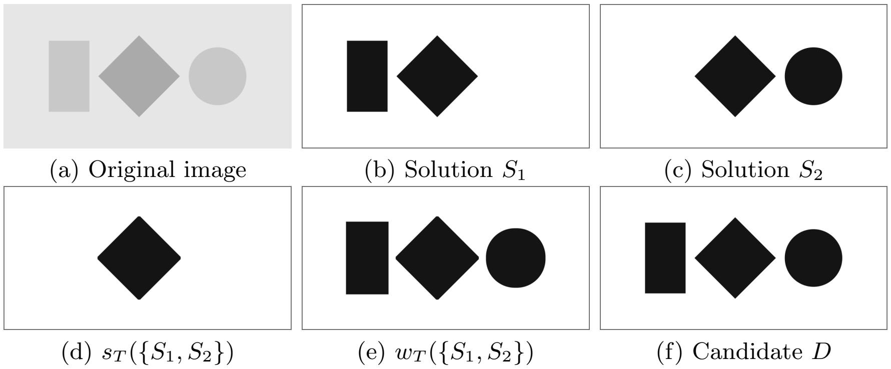
(i) For any $\mathbf{I} \in \wp(\mathbb{B})$, it holds that $\mathbf{I} \subseteq c_T(\mathbf{I})$. This guarantees goal G3.
(ii) For any $\mathbf{I} \in \wp(\mathbb{B})$, it holds that $c_T(\mathbf{I}) = c_T(\{s_T(\mathbf{I}), w_T(\mathbf{I})\})$.
(iii) For any $\mathbf{I} \in \wp(\mathbb{B})$, it holds that $c_T(\mathbf{I}) = c_T(c_T(\mathbf{I}))$.
(iv) For any $\mathbf{I} \in \wp(\mathbb{B})$ and $B \in \mathbb{B}$, it holds that $B \in c_T(\mathbf{I})$ if and only if $c_T(\mathbf{I}) = c_T(\mathbf{I} \cup \{B\})$. Hence, the information in images within the set does not exceed that in the set itself.
(v) For any $I \in \wp(\mathbb{B})$, $c_T(\mathbf{I})$ defines a connected subspace of $\mathbb{B}$, *i.e.*, for any $B_1, B_2 \in c_T(\mathbf{I})$, there exists a sequence of images $B_1^*, \ldots, B_r^*$ in $c_T(\mathbf{I})$, so that $B_1^* = B_1$, $B_r^* = B_2$, and two consecutive images B_i^* and B_{i+1}^* only differ in one pixel.

2.4 Visual properties of the set consensus

The set $c_T(\mathbf{I})$, which we refer to as TCGT in the remainder of this work, has interesting visual properties related to the information in the images in $\mathbf{I}$.

The first property is that of *information combination.* This property refers to the ability to combine information from different ground truth images, meaning that the resulting set selectively picks information from each image. An example can be found in Fig. 2. Considering the original image in Fig. 2(a), two humans have created the ground truth images S_1 and S_2 in Figs. 2(b)-(c). The strong and weak consensus of the set of images are included in Figs. 2(d)-(e). The candidate image in Fig. 2(f), which is a selective combination of the images S_1 and S_2, actually belongs to their TCGT (i.e., $D \in c_T(\{S_1, S_2\})$). This illustrates how the TCGT is able to implicitly produce derived information from the combination of divergent solutions. Otherwise said, images which are not in the original set, but similar to (or composed of parts of) them, are included in the TCGT.

Although the example in Fig. 2 is intentionally simplistic, we can observe that, in the definition of the set-valued consensus, we construct something much more powerful than a closed list of images. There is an actual, yet implicit, knowledge construction process.


(a) Original image (b) Solution S_1 (c) Solution S_2

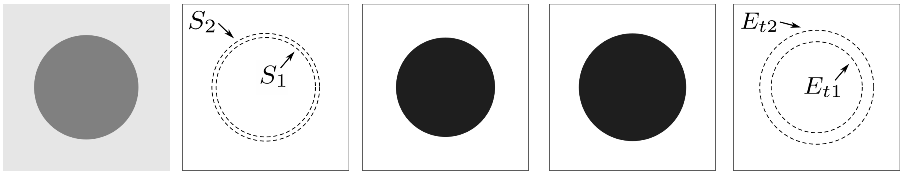
(d) $s_T(\{S_1, S_2\})$ (e) $w_T(\{S_1, S_2\})$ (f) Candidate D

Fig. 2. Example of *information fusion* using the Twofold Consensus Ground Truth. We have (a) an image, (b,c) two hand-made segmentations from it, (d,e) the strong and weak consensus images and (f) a candidate image. The candidate image belongs to $c_T(\{S_1, S_2\})$, although it does not match any of the original images. The structuring element T used for the dilation is a disk of radius 5.

The example in Fig 2 involves the presence or absence of information in a binary image. However, it is also interesting to analyze the alterations in such information, may they be due to contamination, errors or simple interpretation. Regarding this, an interesting property of the TCGT is the *smart tolerance for spatial displacements*.

The TCGT of a set of images includes images containing objects that do not coincide exactly with those delineated by humans in the generation of the ground truth. Moreover, it implicitly discriminates variations as acceptable/unacceptable not only based upon their magnitude (how different), but also upon their congruence of that variation with the existing variations in the original images in **I**. That is, the acceptance of an object depends upon *the amount of spatial variation*, but also upon *its direction*.

Figure 3 includes a binary image with two ground truth solutions (images S_1 and S_2 in Fig 3(b)). Note that only the boundaries of the regions are drawn, so that they can be comfortably compared. In order for an image to be part of the TCGT, the object it features must be in between those of S_1 and S_2. Any image featuring a circle-like region will belong to the TCGT of $\{S_1, S_2\}$ as long as its boundaries are confined between those of S_1 and S_2. Hence, it is not only the fact that distorted solutions (in this case, reduced or enlarged circles) do belong to the TCGT. That distortion is not only measured in terms of *distance to the existing solutions*, but also in terms of congruency w.r.t. the divergences already existing in the TCGT. In this case, a solution created as a slight enlargement of the circle in S_2 (as E_{t1}), or a slight decrease of S_1 (as E_{t2}) are not included in the TCGT. However, greater distortions can be considered within the TCGT, as long as still confined in between the limits of S_1 and S_2.



(a) Original Im. (b) S_1, S_2 (c) $s_T(\{S_1, S_2\})$ (d) $w_T(\{S_1, S_2\})$ (e) E_{t1},E_{t2}

Fig. 3. Example of information fusion based on strong and weak consensus images. We have (a) an image, (b) two hand-made solutions and (c,d) the strong and weak consensus images, respectively. The candidates in (e) are E_{t1}, a slight shrink of S_1, and E_{t2}, a slight enlargment of S_2. We find $E_{t1} \notin c_T(\{S_1, S_2\})$ and $E_{t2} \notin c_T(\{S_1, S_2\})$. In figures (b) and (e) only the limits of the regions are included, for an easier visual inspection. The structuring element for the dilation is a disk of radius 3.

3 Heterogeneity measurement in immunohistochemistry imagery

3.1 Imaging in immunochemistry

Immunohistochemistry (IHC) is an imaging method for studying the localization of antigens in tissue sections (e.g., brain tissue) using antibodies. Different antibodies can be used to demonstrate normal anatomy, protein aggregates, or to indicate pathological conditions such as apoptotic cells. In the images in Fig. 1 antibodies are used against Tau, a protein normally localized in the axon of neuron cells that can be pathologically deposited in some neurodegenerative diseases such as Progressive Supranuclear Palsy. The final stage of the tissue is that in which the regions affected by Tau protein take a distinctive color. The measurement and analysis of these images relates, hence, to the localization of pixel clusters with the visual characteristics of the affected regions.

3.2 Heterogeneity measurement in IHC imagery

IHC imaging is a costly technique, specially in terms of the time consumed by experts. Depending on the expected output of the IHC image analysis, experts can take hours analyzing and labelling visible artefacts in one image. As an example, the images from which the patches in Fig. 4 are taken contain around 5 megapixels, and often feature hundreds of size-variable tau-affected regions. A detailed analysis of these images cannot be tackled in less than few hours by an expert neuropathologist. Hence, it is very interesting to create automatic procedures that can measure the amount of tau protein visible in IHC images. That is, to create algorithms to replace humans in IHC image analysis.

The first problem encountered to design specific image processing algorithms for IHC is the absence of a large number of reliably-labelled ground truth images. The reason for this absence is the amount of time required to generate them, which forces the neuropathologists to perform semi-quantitative analysess based on quick visual inspection (e.g. *mildly affected* or *very affected*). This absence of ground truth images leads to a dual problem in the context of image processing. Firstly, the absence of the ground truth makes the segmentation task to be as poorly defined as *replicating the labelling a human would perform.* Secondly, there is very few data the results of the algorithm can be tested against. In these conditions, any training or comparison effort tends to be overinfluenced by the specific conditions of the ground truth.

We intend to overcome the lack of ground truth by requiring pathologists to label small, randomly selected subregions within some images. This would cut down the amount of time required from the experts, and would give partial, yet reliable, data about the expected results. Also, this brings a subsidiary problem: different pathologists produce very different label maps for the same image. A significant part of the tau-affected artifacts is homogeneously identified as positive detections. But, there is also a large margin for heterogeneity, especially related to (a) the margins of the artifacts and (b) the interpretation of some unclear regions/artifacts. As the size or number of subregions is increased, a new source of heterogeneous decisions appears: (c) lack of attention or tedium. As a result, we have highly variable results by each expert, which is in fact a typical case of multi-valued ground truth.

Problems with multiple ground truth are not unseen in literature, and solutions range from ground truth fusion [2] to performance measure fusion [5]. For example, for the present problem we can compare the results by an algorithm to each image labelled by pathologists, then fuse those results to get an aggregated or *average* performance of an algorithm. However, our goals in this work are different, and root back to the reasons why heterogeneity appears. Questions we face when divergent solutions are produced are: Should we consider all the images in the dataset as equally important, regardless of how heterogeneous their ground truth images are? What does it mean, having a ground truth set with very high (alternatively, low) heterogeneity? Could we measure how well a ground truth fits in a set of ground truth images? Moreover, could we learn to discard those ground truth solutions that are too different from other ground truth images? We intend to use the TCGT to quantify the heterogeneity of a set-valued ground truth; also, to measure the dissimilarity of an ground truth image w.r.t. a set of ground truth images.

We propose to use the TCGT for the generation of metadata about a IHC imagery dataset. Firstly, we want to measure the heterogeneity of a set of solutions. Normally, these measures are constructed from the analysis of one-to-one distances. However, we can also exploit the fact that the TCGT explicitly materializes the coincidences and divergences in a set of binary images.

Definition 4 *Let* $\mathbf{I} = \{I_i, \ldots, I_n\}$ *be a set of binary images. The* heterogeneity *of* $\mathbf{I}$ *is given by*

$$H_T(\mathbf{I}) = 1 - \frac{|s_T(\mathbf{I})|}{|w_T(\mathbf{I})|}$$

where w_T *and* s_T *are the weak and strong consensus, as in Section 2, and* $|\cdot|$ *is the number of featured (*1*-valued) pixels in an image.*

Definition 4 has one major problem: The use of a quotient makes the measure oblivious of the number of pixels in which divergence of opinion exists. Let an extreme case be that in which $\mathbf{I}$ is a set such that $I_1, \ldots, I_{n-1}$ contain one (same) featured pixels and I_n contains one (extra) featured pixel. We have $H(\mathbf{I}) = 0.5$, despite the very subtle difference between images. This problem is partially due to the orientation of the consensus towards the featured information (assuming it is more important than the non-featured one). In this case, two pixels are more important that all of the remaining ones. Still, it feels confusing that a difference of one pixel in one image can have such great impact in the output yielded by the heterogeneity measure.

We propose an alternative version of the heterogeneity measure that solves the aforementioned problem.

Definition 5 *Let* $\mathbf{I} = \{I_i, \ldots, I_n\}$ *be a set of binary images. The* scaled heterogeneity *of* $\mathbf{I}$ *is given by*

$$H^*_T(\mathbf{I}) = \frac{|w_T(\mathbf{I}) \setminus s_T(\mathbf{I})|}{|\Omega|}$$

where w_T *and* s_T *are the weak and strong consensus, as in Section 2.*

There is a list of differences between H and H^*. The most important one is probably the reference for scaling, since they both map to $[0, 1]$ (ignoring the undefined case with $w_T((I)) = \emptyset$). However, they also feature some coincidences. If all images in $\mathbf{I}$ are equal, then $H_T(\mathbf{I}) = H^*_T(\mathbf{I}) = 0$. Also, they both reach maximum values when $s_T(\mathbf{I}) = \emptyset$, although a further analysis of such cases sheds light on a significant difference. In case of H_T, $H_T(\mathbf{I}) = 1$ if and only if $s_T(\mathbf{I}) = w_T(\mathbf{I}) = \emptyset$, except (again) for the undefined case in which all images in $\mathbf{I}$ are empty. However, for H^*_T, the maximum heterogeneity is reached when $s_T(\mathbf{I}) = \emptyset$ and $w_T(\mathbf{I}) = \Omega$.

In our interpretation, the dissimilarity of an image w.r.t. a set of images can be put in terms of the heterogeneity of a set. In fact, to the variation of the heterogeneity when a set is altered.

Definition 6 *Let* $\mathbf{I} = \{I_i, \ldots, I_n\}$ *be a set of binary images, and let* $B \in \mathbb{B}$ *be any binary image. The* dissimilarity *of* B *w.r.t.* $\mathbf{I}$ *is given by*

$$\delta_T(B, \mathbf{I}) = H_T(\{B\} \cup \mathbf{I}) - H_T(\mathbf{I}) \,,$$

where H_T *is a heterogeneity measure, as in Definition 4.*

The dissimilarity measure δ_T is affected by special cases similar to those generating unexpected outputs of H_T. Hence, we also present the scaled dissimilarity δ^*_T.

Definition 7 *Let* $\mathbf{I} = \{I_i, \ldots, I_n\}$ *be a set of binary images, and let* $B \in \mathbb{B}$ *be any binary image. The* scaled dissimilarity *of* B *w.r.t.* $\mathbf{I}$ *is given by*

$$\delta^*_T(B, \mathbf{I}) = H^*_T(\{B\} \cup \mathbf{I}) - H^*_T(\mathbf{I}) \,,$$

where H_T *is a heterogeneity measure, as in Definition 5.*

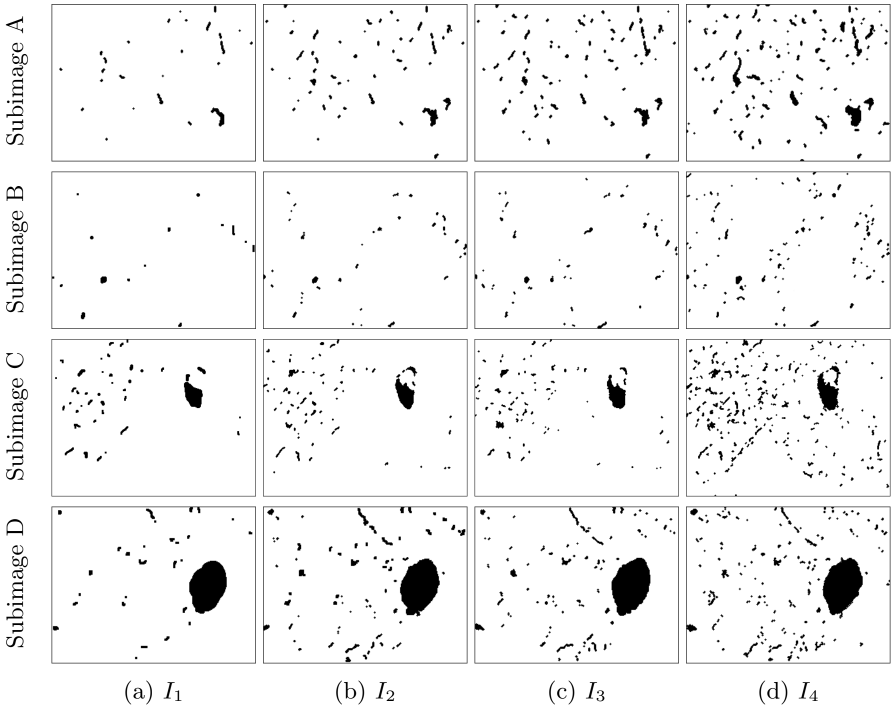


Fig. 4. Hand-labelled images produced by four neuropathologists on four of the subimages in Figure 1. Neuropathologists have been told to be mark the tau-affected areas *conservatively* (column (a)), *normally* (columns (b,c)), or *generously* (column (d)).

3.3 Case study: Measurement of tau protein

It is certainly complicated to know whether metadata is faithful to the actual facts or not [3, 7]. Given the limited amount of space available in the present work, we intend to do a small experiment to see whether the measures capture pathologists' proneness to label more or less regions. Specifically, we induce a certain bias on neuropathologists, and we check whether our measures are able to detect and quantify it.

In order to complete our experiment we have requested four different neuropathologists to label the four leftmost subimages in Fig. 1. One of the neuropathologists was requested to label the area with tau protein in a conservative manner, two other were requested to act normally, and the fourth was requested to label the featured areas in a generous manner. In this way, we expect to have two extreme ground truth images and two solutions that lie somewhere *in the middle*. Of course, pathologists do not take any kind of suggestion on how to perform their work in a normal situation, neither they have bias on the analysis. However, it is, in our opinion, a legitimate way to produce binary images whose behaviour in terms of heterogeneity and dissimilarity is predictable.

Subimage	$H(\mathbf{I})$	$H(\mathbf{I}_{2-4})$	$H(\mathbf{I}_{1-3})$	$H(\mathbf{I}_{2-3})$	$\delta(I_1,\mathbf{I}_{2-3})$	$\delta(I_1,\mathbf{I}_{2-4})$
Subimg. A	.758	.669	.484	.271	.213	.089
Subimg. B	.643	.424	.536	.235	.301	.219
Subimg. C	.747	.485	.636	.258	.377	.262
Subimg. D	.815	.639	.683	.367	.316	.177
Total	.741	.554	.585	.283	.302	.187

(a) Results using heterogeneity and dissimilarity

Subimage	$H^*(\mathbf{I})$	$H^*(\mathbf{I}_{2-4})$	$H^*(\mathbf{I}_{1-3})$	$H^*(\mathbf{I}_{2-3})$	$\delta^*(I_1,\mathbf{I}_{2-3})$	$\delta^*(I_1,\mathbf{I}_{2-4})$
Subimg. A	.272	.239	.082	.044	.038	.033
Subimg. B	.236	.154	.153	.066	.087	.082
Subimg. C	.249	.162	.147	.060	.087	.087
Subimg. D	.154	.120	.076	.040	.036	.034
Total	.228	.169	.114	.053	.062	.059

(b) Results using scaled heterogeneity and dissimilarity

Table 1. Results obtained in the quantification of heterogeneity and dissimilarity of the sets displayed in Fig. 4. For each subimage, $\mathbf{I}$ refers to all of the ground truth solutions for each image, while $\mathbf{I}_{i-j}$ refers to the images in colums from i to j, both included. The structuring element T (which is a circle with radius 5) is ommitted from the formulation in order to ease the interpretation of the table.

The images produced for the experiments are included in Fig. 4. Each row in the figure corresponds to one of the four leftmost images in Fig. 1, while each column corresponds to one of the instructions given to the pathologists. Specifically, the leftmost column is the most conservative inspection of the images, while the rightmost column contains the images in which the neuropathologists was proner to label tau protein.

We have used the measures presented in Section 3.2, as recap in Table 1. The standing assumption of our experiment is that images generated under extreme biases should be identified as such by inspecting the values yielded by our measures. Table 1 displays the values gathered in different evaluations for the image sets at each of the rows of Fig. 4.

From the results in Table 1, we can confirm that our measures actually behave according to the semantics of the images. For example, in terms of heterogeneity, the values yielded by H or H^* suffer a severe increase when the set $\mathbf{I}$ includes the images I_1 or I_4, compared to when it does not. For both H and H^* the heterogeneity of $\mathbf{I}_{2-3}$ is significantly increased by adding the images I_1 or I_4, which play the role of extreme cases. This holds for all subimages and heterogeneity measures. In Table 1 we also observe that δ_T and δ^*_T identify the outlying images I_1 and I_4 w.r.t. the *neutral* images I_2 and I_3.

It is relevant to mention that, considering the very small size of the experiment, results shall be put to the test in a more complete scenario. Still, it is rather complicated to find input for metadata evaluation, and typically one must rely on either experiment-driven data (as in this case), or on questionable assumptions on the way in which ground truth data was generated.

4 Conclusions

In this work we have tackled the problem of multiple ground truth in medical imagery, specificaly in immunohistochemistry imagery of brain tissue. We have questioned the reasons on the divergences between experts when required to label such images, and proposed four different measures to quantify the heterogeneity in a set of images, as well as the 1-to-n dissimilarity of images. In order to do so, we have applied the notions and developments of the Twofold Consensus Ground Truth (TCGT), a a set-valued operator created for binary image fusion. This application intends to illustrate how fusion operators in can be used for purposes other than information aggregation or compression. In our example, the twofold consensus ground truth is used not only to fusion hand-labelled IHC images, but also to analyze the heterogeneity in a set of them, as well as to create one-to-many or many-to-many dissimilarity measures.

Despite the innovative nature of the application, we consider that our work has a solid, context-agnostic mathematical background. However, it requires more comprehensive experimental validation, considering the oriented nature of this research. Hence, the design and analysis of such a experimental setup is a key future line of research.

References

1. Chen, S., Haralick, R.: Recursive erosion, dilation, opening, and closing transforms. IEEE Trans. on Image Processing 4(3), 335–345 (1995)
2. Fernández-García, N., Carmona-Poyato, A., Medina-Carnicer, R., Madrid-Cuevas, F.: Automatic generation of consensus ground truth for the comparison of edge detection techniques. Image and Vision Computing 26(4), 496–511 (2008)
3. Lopez-Molina, C., Bustince, H., De Baets, B.: Separability criteria for the evaluation of boundary detection benchmarks. IEEE Trans. on Image Processing 25(3), 1047–1055 (2016)
4. Lopez-Molina, C., De Baets, B., Bustince, H.: Twofold consensus for boundary detection ground truth. Knowledge-Based Systems 98, 162–171 (2016)
5. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. of the IEEE International Conf. on Computer Vision. vol. 2, pp. 416–423 (2001)
6. Pawlak, Z.: Rough sets. International Journal of Computer & Information Sciences 11(5), 341–356 (1982)
7. Pont-Tuset, J., Marques, F.: Measures and meta-measures for the supervised evaluation of image segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. pp. 2131–2138 (2013)
8. Serra, J.: Image Analysis and Mathematical Morphology. Academic Press, Inc. (1983)
9. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Trans. on Medical Imaging 23(7), 903–921 (2004)