CatSIM: A Categorical Image Similarity Metric

Geoffrey Z. Thompson and Ranjan Maitra

Abstract—We introduce CatSIM, a new similarity metric for binary and multinary two- and three-dimensional images and volumes. CatSIM uses a structural similarity image quality paradigm and is robust to small perturbations in location so that structures in similar, but not entirely overlapping, regions of two images are rated higher than using simple matching. The metric can also compare arbitrary regions inside images. CatSIM is evaluated on artificial data sets, image quality assessment surveys and two imaging applications.

Index Terms—image analysis, image segmentation, image similarity, distortion measurement, structural similarity, SSIM, Jaccard

1 Introduction

Similarity metrics are important for comparing 2D images and 3D volumes, for instance, when evaluating image processing or segmentation algorithms. Categorical images assign nominal values at each pixel/voxel, *e.g.* in tissue or land classification in medical or satellite images. Most categorical image similarity methods only provide pointwise class comparisons at each pixel/voxel [1], [2] and ignore spatial context. Such measures [3] include the Jaccard ($\mathcal{J}$) [4] or Dice ($\mathcal{D}$) [5] indices for binary problems, Cohen's κ [6] or Hamming's distance [7] for multi-class comparisons with class labels having the same meaning in both cases being compared, or the Rand ($\mathcal{R}$) [8] or Adjusted Rand indices ($\mathcal{AR}$) [9] for when they do not. These metrics can mislead, especially when the features of interest are objects or fine structures like lines that may be spatially perturbed with little visual effect but large pointwise perturbations.

There exist image similarity metrics that account for the relationships of nearby points or features. For instance, [10], [11] use geometric methods that account for spatial and intensity distortions by identifying and comparing object edges in images while [12] uses region-wise calculations and asymptotic normality arguments to arrive at a regional mutual information metric for comparing two images. The popular multiscale structural similarity (SSIM) or MS-SSIM [13], [14] for color and grayscale images accounts for spatial and intensity distortions as well as structural information across multiple scales in the image. A computationally intensive version, CW-SSIM [15] exists for grayscale and binary images. These methods have good general performance but do not always align with human assessment [16]. Such methods also do not apply to multi-class images or volumes so we propose (Section 2) methodology that adapts statistics specifically appropriate for multinary and binary data to an SSIM-like approach. Section 3 illustrates and validates our methods. The paper concludes with some discussion. An online supplement with sections, figures and tables referenced here using the prefix "S" is available.

G. Z. Thompson and R. Maitra are with the Department of Statistics, Iowa State University, Ames, Iowa 50011, USA. e-mail: {gzt,maitra}@iastate.edu.

2 Methodology

2.1 Background and Previous Work

SSIM [13] is an image quality assessment index conceptually designed to account for structural similarities in images as visualized by a human rater. The basic version of the index combines summary statistics on sliding $N\times N$ aligned patches of the images ($\boldsymbol{X}$ and $\boldsymbol{Y}$) being compared. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be aligned patches from $\boldsymbol{X}$ and $\boldsymbol{Y}$, with $\mu_{\boldsymbol{x}}$ and $\mu_{\boldsymbol{y}}$ being the averaged values in each patch, $\sigma^2_{\boldsymbol{x}}$ and $\sigma^2_{\boldsymbol{y}}$ the variances, and $\sigma_{\boldsymbol{xy}}$ the covariance. The SSIM is calculated from the luminance, contrast and structural similarity functions, $l(\boldsymbol{x},\boldsymbol{y})$, $c(\boldsymbol{x},\boldsymbol{y})$ and $s(\boldsymbol{x},\boldsymbol{y})$ as follows. We define $l(\boldsymbol{x},\boldsymbol{y}) = f(\mu_{\boldsymbol{x}},\mu_{\boldsymbol{y}};C_1)$ and $c(\boldsymbol{x},\boldsymbol{y}) = f(\sigma_{\boldsymbol{x}},\sigma_{\boldsymbol{y}};C_2)$ where

$$f(\theta,\phi;k) = \frac{2\theta\phi + k}{\theta^2+\phi^2+k}, \qquad s(\boldsymbol{x},\boldsymbol{y}) = \frac{\sigma_{\boldsymbol{xy}} + C_3}{\sigma_{\boldsymbol{x}}\sigma_{\boldsymbol{y}} + C_3},$$

with C_1, C_2, and C_3 small constants. The three functions take values in [0,1], with 1 attained for identical patches. The means and the standard deviations (SDs) of the two patches influence $l(\boldsymbol{x},\boldsymbol{y})$ and $c(\boldsymbol{x},\boldsymbol{y})$ while $s(\boldsymbol{x},\boldsymbol{y})$ compares the covariance to their individual SDs. Averaging each function over all possible $N\times N$ patches in $\boldsymbol{X}$ and $\boldsymbol{Y}$ provides $l(\boldsymbol{X},\boldsymbol{Y})$, $c(\boldsymbol{X},\boldsymbol{Y})$ and $s(\boldsymbol{X},\boldsymbol{Y})$. Multiplying these yields $\text{SSIM}(\boldsymbol{X},\boldsymbol{Y}) = l(\boldsymbol{X},\boldsymbol{Y})c(\boldsymbol{X},\boldsymbol{Y})s(\boldsymbol{X},\boldsymbol{Y})$.

MS-SSIM [13] computes the SSIM at multiple scales after downsampling and combining the results from each scale:

$$\text{MS-SSIM}(\boldsymbol{X},\boldsymbol{Y}) = l_M(\boldsymbol{X},\boldsymbol{Y})^{\alpha_M}\prod_{j=1}^{M} c_j(\boldsymbol{X},\boldsymbol{Y})^{\beta_j} s_j(\boldsymbol{X},\boldsymbol{Y})^{\gamma_j}$$

where j indexes the scale at which $c(\boldsymbol{X},\boldsymbol{Y})$ and $s(\boldsymbol{X},\boldsymbol{Y})$ are computed, M is the highest scale after $M-1$ iterations of re-scaling, and α_j, β_j, and γ_j are constants on the M-dimensional simplex. [13] use $\alpha_j = \beta_j = \gamma_j$ with values empirically specified. The metric aims to capture local variation and structural similarities between images on several scales in a way that mimics the human eye.

CW-SSIM [15] uses the product of functions of the magnitude and phase of the complex wavelet coefficients of images downsampled over 6 levels. The magnitudes feed into a SSIM index that is 1 only when they match for both images. The metric uses a function of the dot product of the

phases that is 1 only when both images are aligned. CW-SSIM is said to tie to computational models for vision [15].

The SSIM and MS-SSIM methods can be made to work for binary images (by considering 0-1 to be part of the continuum) while CW-SSIM is formulated for binary or grayscale images. However, neither extends to nominal multi-class images. CW-SSIM can also not handle image data with missing observations (such as pixels/voxels outside a mask). Also, while not a major limitation, neither methods are currently implemented for 3D volumes. So we develop a SSIM metric for multi-class images and volumes.

2.2 Development of CatSIM

The SSIM philosophy can be developed for multinary images by defining suitable binary or multinary analogs of the stages of the MS-SSIM algorithm. In Section 2.2.1, we introduce luminance, contrast, and structural similarity functions that respect the multinary nature of the data. In Section 2.2.2, we define the method of downsampling to other scales. Finally, in Section 2.3, we specify how the results are combined across scales to produce the final metric.

2.2.1 Index Functions

We first define statistics for a categorical image or image patch $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and $x_i \in \{1, 2, \ldots, K\}$. Let $\boldsymbol{m_x}$ be the vector of the proportions of each class in the patch $\boldsymbol{x}$, and $S_{\boldsymbol{x}}$ a categorical variance measure [17]. That is, define

$$p_i = \frac{1}{n}\#\{x_j = i,\, j \in 1, 2, \ldots, n\} \quad \text{for } i = 1, 2, \ldots, K$$

$$\boldsymbol{m_x} = \{p_i\}_{i=1}^{K}, \quad S_{\boldsymbol{x}} = \frac{1 - \sqrt{\sum_{i=1}^{K} p_i^2}}{1 - 1/\sqrt{K}} = \frac{1 - \|\boldsymbol{m_x}\|_2}{1 - 1/\sqrt{K}}.$$

These quantities are related to the Gini-Simpson index [18], [19] but are scaled similarly to the SD and constrained to be in $[0, 1]$. For two aligned $N_u \times N_v (\times N_w)$ sections of K-class image (volume) patches $\boldsymbol{x}$ and $\boldsymbol{y}$, with $\boldsymbol{m_x}$, $\boldsymbol{m_y}$, $S_{\boldsymbol{x}}$, and $S_{\boldsymbol{y}}$, we construct luminance, contrast, and structural similarity functions analogous to SSIM for the continuous case. Specifically, we insert categorical analogues in place of the means and covariances (in SSIM) to get:

$$l^c(\boldsymbol{x}, \boldsymbol{y}) = \frac{2\boldsymbol{m_x}^\top \boldsymbol{m_y} + C_1}{\boldsymbol{m_x}^\top \boldsymbol{m_x} + \boldsymbol{m_y}^\top \boldsymbol{m_y} + C_1}$$

$$c^c(\boldsymbol{x}, \boldsymbol{y}) = \frac{2S_{\boldsymbol{x}} S_{\boldsymbol{y}} + C_2}{S_{\boldsymbol{x}}^2 + S_{\boldsymbol{y}}^2 + C_2}, \quad s^c(\boldsymbol{x}, \boldsymbol{y}) = v(\boldsymbol{x}, \boldsymbol{y}),$$

where C_1, C_2, are small scalar constants chosen for numerical stability when the denominator approaches zero, and $v(\boldsymbol{x}, \boldsymbol{y})$ is an inter-rater agreement measure chosen based on the characteristics of the image. For instance, the $\mathcal{J}$ and $\mathcal{D}$ indices are appropriate choices for $v(\boldsymbol{x}, \boldsymbol{y})$ in binary problems where presence is more important than absence [20]. Accuracy or κ is more appropriate if labels have meaning while $\mathcal{R}$ or $\mathcal{AR}$ is more appropriate for images with labels that are arbitrarily assigned. (Both κ and $\mathcal{AR}$ can take negative values so are truncated to be in $[0, 1]$.)

Our index can accommodate missing values, as in the case of imaging a volume with known boundaries (*e.g.*, a mask in medical imaging applications), as in Sections 3.1.3 or 3.3.1, by the pair-wise deletion of corresponding points in the sliding window calculations at each level.

2.2.2 Downsampling and combining across multiple scales

There are two issues to address: how to perform downsampling in a non-continuous (nominal discrete) setting and how to combine the different results across scales. The MS-SSIM algorithm down-samples by a factor of two after using a low-pass filter to reduce aliasing artifacts but can not be applied here because it disregards the structure of the data in the binary case and is meaningless for multinary images. We propose the mode of each $2\times2(\times2)$ slice (block) of pixels (with a random choice from multiple modes if they exist). The multiple scales can be combined for categorical images in a similar manner as MS-SSIM and as detailed next.

2.3 The CatSIM Algorithm

By default, we specify uniform window sizes $N_u = N_v = 11$ for 2D images and $N_u = N_v = N_w = 5$ for 3D volumes. We also set $M = 5$ levels and, as in [13], $\alpha_j = \beta_j = \gamma_j \; \forall j$. We choose γ_js to be uniform over the M-dimensional simplex. These parameters can all be set based on the application.

1) For two images $\boldsymbol{X}$ and $\boldsymbol{Y}$, the $c^c(\boldsymbol{x}, \boldsymbol{y})$ and $s^c(\boldsymbol{x}, \boldsymbol{y})$ statistics are computed over a rolling $N_u \times N_v (\times N)$ pixel (voxel) window and averaged for the entire image while $l^c(\boldsymbol{x}, \boldsymbol{y})$ is computed for the base level.
2) Downsample each image by a factor of 2 using the mode (break ties at random) of each $2\times2(\times2)$ block.
3) Repeat Steps 1 and 2 for each of M total levels.
4) Let $l_1^c(\boldsymbol{X}, \boldsymbol{Y})$, $c_j^c(\boldsymbol{X}, \boldsymbol{Y})$ and $s_j^c(\boldsymbol{X}, \boldsymbol{Y})$ be the average of $l_1^c(\boldsymbol{x}, \boldsymbol{y})$, $c_j^c(\boldsymbol{x}, \boldsymbol{y})$ and $s_j^c(\boldsymbol{x}, \boldsymbol{y})$ over all $N_u \times N_v (\times N_w)$ blocks, for $j = 1, 2, \ldots, M$. Define

$$\begin{aligned} &\text{CatSIM}(\boldsymbol{X}, \boldsymbol{Y}) \\ &\quad = [l_1^c(\boldsymbol{X}, \boldsymbol{Y})]^{\alpha_M} \prod_{j=1}^{M} c_j^c(\boldsymbol{X}, \boldsymbol{Y})^{\beta_j} s_j^c(\boldsymbol{X}, \boldsymbol{Y})^{\gamma_j}, \end{aligned} \quad (1)$$

with j indexing the level at which the luminance, contrast and similarity functions are calculated.

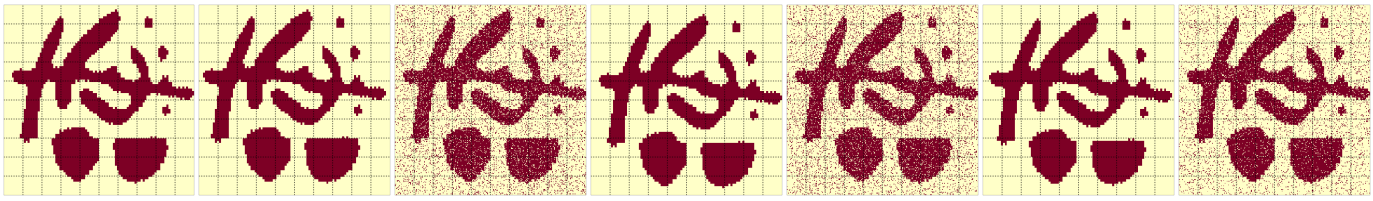
3 Illustrations and Evaluations

3.1 Illustration of CatSIM

We illustrate CatSIM's behavior on different distortions and degradations of two binary images (one with a mask and highly unequal class sizes) and one multinary image.

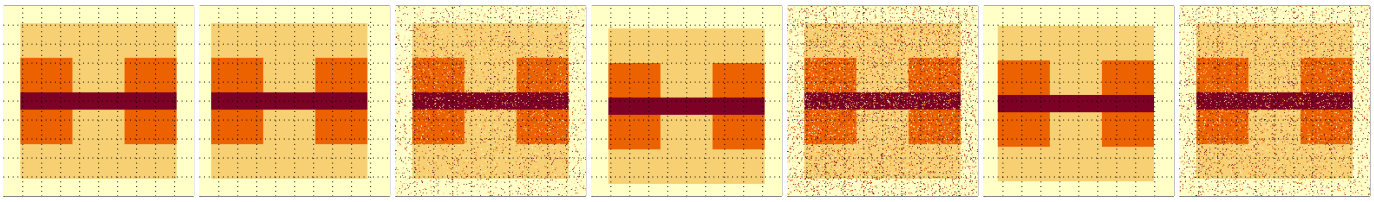
3.1.1 Expanded Besag binary image

[21] presented a 88×100 hand-drawn binary scene with intentionally awkward features. We magnified the image onto a 264×300 grid for it to be large enough to run CatSIM (κ) with five layers and to accommodate spatial translations. This expanded Besag image (EBI) (Figure 1a) was trimmed by 12 pixels off the right and bottom margins to allow for translations of the central portion. We created horizontally-shifted (by 6 pixels, Figure 1b), vertical-shifted (by 6 pixels, Figure 1d) and horizontally-and-vertically shifted (by 3 pixels in each direction, Figure 1f) versions of the EBI. Additional degraded versions (Figures 1c, 1e, and 1g) of the EBI were created by adding salt-and-pepper noise (S&P) with error rates matching those of each of the shifted images.



(a) True (b) H.-shift (c) H.-S&P (d) V.-shift (e) V.-S&P (f) H.&V.-shift (g) H.&V.-S&P

Fig. 1. A binary image from Besag (Figure 1a) to demonstrate the CatSIM metric. There are three spatial translations of the central part of the image (Figures 1b, 1d, and 1f) and images with salt-and-pepper noise added to match their error rates (Figures 1c, 1e, and 1g).



(a) True (b) H.-shift (c) H.-S&P (d) V.-shift (e) V.-S&P (f) H.&V.-shift (g) H.&V.-S&P

Fig. 2. A constructed multicategory image (Figure 2a) to demonstrate the CatSIM metric. There are three spatial translations of the central part of the image (Figures 2b, 2d, and 2f) and images with salt-and-pepper noise added to match their error rates (Figures 2c, 2e, and 2g).

TABLE 1
CatSIM (κ) and other metrics for different distortions of the EBI of Figure 1.

Image	CatSIM 5 levels	CatSIM 1 level	CatSIM (whole)	$\mathcal{AR}$	κ	CW-SSIM
H Shift	0.594	0.464	0.763	0.627	0.763	0.831
H - S & P	0.515	0.092	0.769	0.630	0.771	0.783
V Shift	0.569	0.449	0.751	0.610	0.751	0.752
V - S & P	0.516	0.090	0.756	0.613	0.759	0.780
H & V Shift	0.658	0.561	0.827	0.720	0.827	0.834
H & V - S & P	0.557	0.110	0.832	0.725	0.834	0.810

Tables 1 and S1 illustrate the CatSIM, CW-SSIM and other space-agnostic metrics on the different cases of Figure 1. Each pair (shifted and matching S&P-degraded original) of figures have similar values for $\mathcal{AR}$ and κ. On the other hand, CW-SSIM rates the horizontal- and horizontal-and-vertically-shifted versions higher than their matching noisy counterparts but lower for the vertically-shifted version. However, the translated images are almost visually indistinguishable from the original, since these are minor spatial perturbations, while the noisy versions are more discordant, and these factors should be reflected in a metric that mimics the visual system. We compute three different versions of CatSIM: the default with five different scales, the default but with only the first scale (no downsampling), and one computing the index for the entire image at once on one scale (rather than averaging results from a sliding window). CatSIM rates the spatially-shifted images differently from the S&P-degraded images. The difference is stark when considering only one level, to the point that it cannot pick up the structural similarity at that scale. How much this matters in the final index depends on the weights chosen for each level. CatSIM calculated over the whole image (rather than a sliding window) is expectedly not much different from κ, which it is based on. Interestingly and like 1-level CatSIM, MS-SSIM (Table S1) applied to 0-1 images assumed to be in continuous space, regards the degraded images very poorly with rates that are not justified visually. In summary, the 5-level CatSIM (with default settings) provides the most consistent representation of the scene.

3.1.2 *Four-class image*

Figure 2a is our example of a 220×220 four-class image, with several distinct spatial regions and vertically and horizontally symmetric, around the middle, in each dimension. MS-SIM and CW-SSIM are not easily extended to the multinary case, so we only study the behavior of CatSIM and the space-unaware metrics to spatial shifts (Figures 2b, 2d 2f) and matching S&P degradations (Figures 2c, 2e. 2g) of the image obtained in the same manner as in Section 3.1.1.

Tables 2 and S2 list the index values. As in the EBI example, $\mathcal{AR}$ and κ values are similar for each pair of the shifted and its S&P-matched twin. The case for the whole-image CatSIM values is similar. However, the default 5-level and 1-level CatSIM indices clearly distinguish between images that are minor spatial perturbations over images degraded with added noise. The 1-level CatSIM index is unnecessarily harsh on the S&P-degraded images, being very poor at

TABLE 2
CatSIM (κ) and other metrics for different distortions of the four-class image of Figure 2.

Image	CatSIM 5 levels	CatSIM 1 level	CatSIM (whole)	$\mathcal{AR}$	κ
H Shift	0.816	0.686	0.906	0.828	0.906
H - S & P	0.610	0.105	0.906	0.842	0.907
V Shift	0.652	0.462	0.827	0.723	0.827
V - S & P	0.548	0.079	0.825	0.717	0.827
H & V Shift	0.814	0.533	0.869	0.777	0.869
H & V - S & P	0.565	0.093	0.874	0.790	0.875

recognizing similarity with the original. The 5-layer version (with default parameters) recovers this similarity because the downsampling smooths out the added noise.

3.1.3 *Highly-imbalanced binary image*

Our next illustration is on a binary image with a mask and with disproportionate class sizes, as happens, say, in activation detection with functional Magnetic Resonance Imaging (fMRI) where no more 3% voxels are expected to be activated. Our ground truth (Figure 3a) is a 256×256 version of the modified 128×128 Hoffman activation phantom [22] that has a small proportion (3.98%) of truly activated in-brain pixels [23]. The in-brain pixels form a mask, which also renders a method such as CW-SSIM inapplicable.

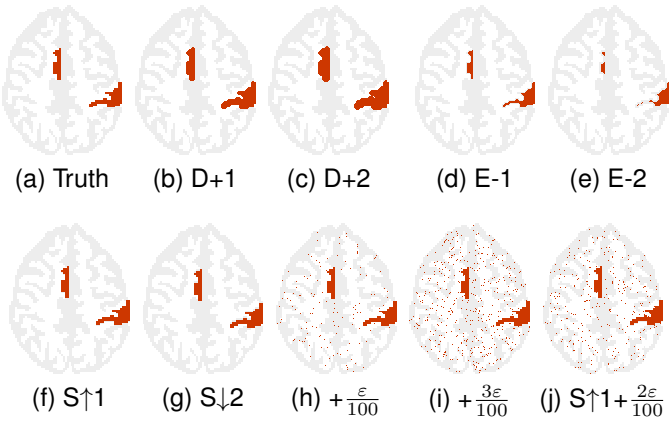


Fig. 3. (a) The in-brain portion of the modified Hoffman activation phantom of [23] and distortions of its activated (dark-shaded) regions: dilation by (b) 1 and (c) 2 pixels, erosion by (d) 1 and (e) 2 pixels, (f) upward shift by 1 pixel, (g) downward shift by 2 pixels. Other distortions add noise by randomly activating (h) 1%, (i) 3% and (j) 2% inactivated in-brain pixels, the last after also shifting the activated region upward by 1 pixels.

We prepared nine distorted versions of the image using mathematical morphology and other operations. Specifically, we dilated the activated regions by one (D+1) and two (D+2) pixels (with 1 or 2 inactivated pixels orthogonally adjacent to an activated pixel changing status), eroded the activated regions by one (E-1) and two (E-2) pixels (essentially deactivating any activated pixel orthogonally adjacent to an inactivated one), the activation regions shifted up one (S↑1) and down two (S↓2) pixels, the original image with 1% ($+\frac{\varepsilon}{100}$) and 3% ($+\frac{3\varepsilon}{100}$) of the inactive pixels randomly activated, and the original image with the activated regions shifted up by one pixel and adding 2% noise (S↑1+2%).

The similarity indices for these distorted images (Table 3) all agree with our expectation that an increase of distortion - a larger shift or more noise - will have a lower similarity to the base image. The accuracy (Acc.) is a poor measure in this context because the important feature to capture is the difference in activation, and it reports a very high agreement for all of the distorted images (because the pixels in all the images are largely of one class). The Jaccard and Dice indices deteriorate as desired with increasing distortion and both proposed methods, CatSIM with $\mathcal{J}$, or CatSIM($\mathcal{J}$) and CatSIM with Cohen's κ, do as well, though CatSIM(κ) is not specifically designed for this behavior and does not capture the difference as well as the other indices. Compared to

TABLE 3
Similarity indices for the distorted and degraded images of Figure 3.

	$\mathcal{J}$	Dice	Acc.	CatSIM ($\mathcal{J}$)	CatSIM (κ)
Dilated (D+1)	0.75	0.86	0.99	0.74	0.94
Dilated (D+2)	0.61	0.76	0.97	0.59	0.88
Eroded (E-1)	0.68	0.81	0.99	0.61	0.90
Eroded (E-2)	0.43	0.60	0.98	0.33	0.72
Shift Up (S↑1)	0.83	0.91	0.99	0.75	0.94
Shift Down (S↓2)	0.68	0.81	0.99	0.52	0.77
Noise ($+\frac{\varepsilon}{100}$)	0.81	0.90	0.99	0.63	0.80
Noise ($+\frac{3\varepsilon}{100}$)	0.57	0.73	0.97	0.45	0.58
Shift+Noise (S↑1$+\frac{2\varepsilon}{100}$)	0.58	0.73	0.97	0.42	0.63

$\mathcal{J}$, CatSIM($\mathcal{J}$) penalizes noise in the images more than for minor perturbations that do not affect the basic spatial extent of the activated region. We now study the impact of the (five) layers in the calculation of CatSIM($\mathcal{J}$).

Table 4 illustrates the utility of the five layers in the calculation of the index. The first level, without any down-

TABLE 4
CatSIM ($\mathcal{J}$) values for each layer for different types of distortions. Layer 1 is the image itself while subsequent layers downsample the image in the previous layer by a factor of 2.

	Layer 1	2	3	4	5
Dilated (D+1)	0.60	0.65	0.68	0.83	1.00
Dilated (D+2)	0.43	0.48	0.53	0.62	1.00
Eroded (E-1)	0.52	0.51	0.57	0.56	1.00
Eroded (E-2)	0.28	0.32	0.34	0.38	0.35
Shift Up (S↑1)	0.72	0.59	0.72	0.75	1.00
Shift Down (S↓2)	0.55	0.59	0.58	0.59	0.35
Noise ($+\frac{\varepsilon}{100}$)	0.12	0.86	1.00	1.00	1.00
Noise ($+\frac{3\varepsilon}{100}$)	0.04	0.46	1.00	1.00	1.00
Shift+Noise (S↑1$+\frac{2\varepsilon}{100}$)	0.05	0.48	0.79	0.75	1.00

sampling, rates any added noise poorly, but higher levels smooth out that difference. Relatively large differences, such as caused by double erosion (E-2), damage the image's rating across all scales. Any larger scale than this in this application has ratings of either 1 or 0 as the activated class starts to disappear completely. The optimal number of levels depends on the size of the features in the image, the type of distortion, and the demands of the application, but using five equally weighted levels seems, in this application and generally, to strike a balance and provide good performance.

3.2 Image Quality Assessment Survey

Having ilustrated CatSIM's ability to capture structural similarity in categorical images, we now evaluate its ability to represent human visual perception. We conducted two separate surveys to compare the metric's assessment on binary and multinary images to human judgment of the image quality. We discuss the surveys and their results next.

3.2.1 *Assessment on Binary Images*

Figure 4 displays the twelve 256×256 binary images that, along with their distorted versions, were used in the survey. The first image (Figure 4, first row) is of an 8-squares checker-board pattern, and is followed by images of three

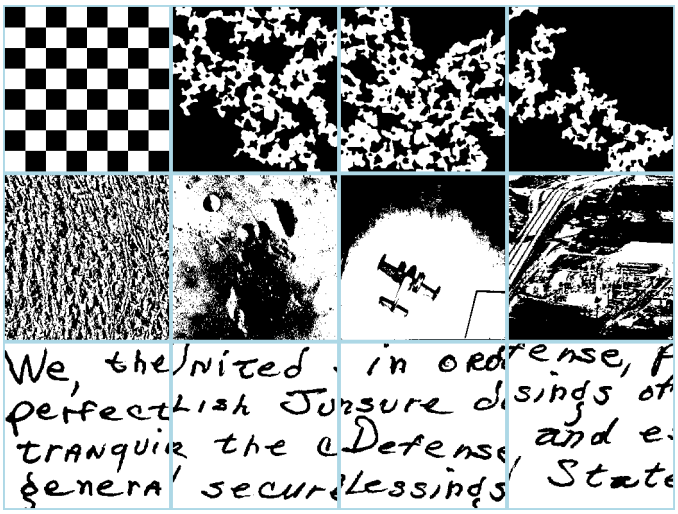


Fig. 4. The 12 undistorted 256×256 binary images from the first survey.

thresholded Gaussian processes. The next row of images (Figure 4) were obtained by thresholding monochrome images (of a texture, the lunar surface, an aerial view of, first, an airfield with an airplane, and then a highway overpass) from the USC-SIPI image database. The last set of binary images had thresholded versions of four handwriting samples from the NIST Special Database 19 [24]. Seventy-four adult volunteers were shown 30 pairs of images, with each pair comprising a randomly chosen image from Figure 4 and a distorted version. They were asked to rate the quality of each distorted image on a scale from 1 to 100, with 100 indicating perfect fidelity. For each of $12 \times 12 = 144$ ground truth-distorted image pairs, we calculated the mean opinion scores (MOS) over all the respondents shown that particular image pair. For these image pairs, we compared the MOS to CW-SSIM, the space-unaware metrics of κ, $\mathcal{AR}$, $\mathcal{J}$ and accuracy (that can be related to Peak-Signal-to-Noise-Ratio) and the CatSIM metrics with κ, $\mathcal{AR}$, $\mathcal{J}$ and accuracy.

Figure 5, displays the MOS with each of the metrics

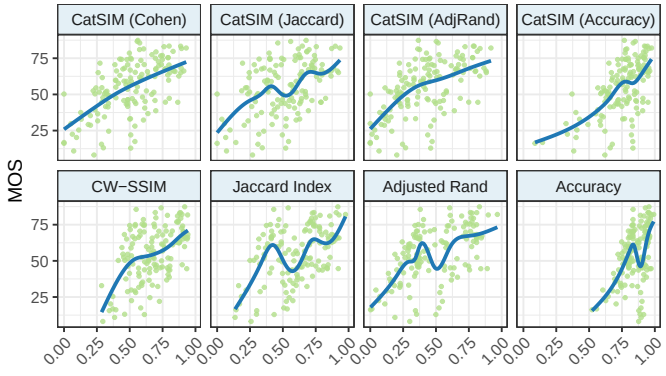


Fig. 5. Adjusted Rand index, CW-SSIM, CatSIM, Jaccard index, and accuracy compared to MOS rating for binary images in the first survey. CatSIM is computed using the default κ, $\mathcal{AR}$, $\mathcal{J}$ and accuracy.

under consideration. These metrics are all positively correlated with the MOS, with CatSIM methods using κ, $\mathcal{AR}$ and accuracy performing the best (Table 5(a)). A randomization test (see Section S2-A for details) indicated significantly

TABLE 5
Correlation with MOS for each metric and (b) p-value of test statistic for the assessment on binary images.

Method	ρ
CatSIM (Acc)	0.601
CatSIM (κ)	0.598
CatSIM ($\mathcal{AR}$)	0.580
MS-SSIM	0.578
Cohen's κ	0.577
$\mathcal{AR}$	0.557
CW-SSIM	0.510
Accuracy	0.500
CatSIM ($\mathcal{J}$)	0.470
Jaccard	0.464

(a) Correlations (ρ)

	CatSIM (κ)	CatSIM ($\mathcal{AR}$)	CatSIM (Acc)
MS-SSIM	0.311	0.424	0.410
CW-SSIM	**0.018**	0.082	0.069
Accuracy	**0.032**	**0.029**	**0.033**
Cohen's κ	0.203	0.369	0.348
$\mathcal{AR}$	0.123	0.246	0.204

(b) p-value of test statistic that MOS correlates with CatSIM metrics (column) more than with other methods (row). Bold text indicates significance at the 5% level.

higher correlations with the MOS for CatSIM (κ) against CW-SSIM and accuracy and for both CatSIM ($\mathcal{AR}$) and CatSIM (Accuracy) against accuracy (Table 5(b)). To assess relationships beyond linear association, we also fit (see Figure 5) a monotonic generalized additive model (GAM) to the MOS-values against each metric [25], [26] using generalized cross-validation (GCV) to choose the GAM smoothing parameters. While the CW-SSIM model (Table S4) had the best fit (explaining 39.1% of the deviance), CatSIM ($\mathcal{AR}$) and CatSIM (κ) were ahead of the rest with 37.5 and 37.1% deviance explained. See Section S2-B for more details.

3.2.2 *Assessment on Multinary Images*

This survey used the six multicategory images of Figure 6. Each of 614 adult volunteers were shown 11 sets of 4

Fig. 6. The undistorted multinary images used in the image ranking survey. Respondents were presented a series of panels of 4 distorted versions of each image and asked to rank their quality.

distorted images (see Section S3 for examples) along with the ground truth and asked to rank the distorted images from most to least similar to the original. (Because multinary settings require larger sample sizes, we preserved power by asking participants to rank rather than score. Additionally, ranking provides a more objective basis for comparison across subjects.) In these examples, the labels in the original and distorted images had the same meaning,

TABLE 6
Squared differences in mean rankings produced by human raters in the survey and the rankings produced by similarity metrics.

Method	Squared Difference	RMSE
CatSIM (Accuracy)	72.385	0.1933
CatSIM (Rand)	74.434	0.1961
CatSIM (κ)	77.695	0.2003
MOS	79.848	0.2030
Adjusted Rand Index	81.525	0.2052
Cohen's κ	81.548	0.2052
Rand Index	85.014	0.2096
Accuracy	85.461	0.2101
CatSIM (Adj Rand)	87.838	0.2130

so using CatSIM with (say) κ rather than $\mathcal{A}R$ or $\mathcal{R}$ is more appropriate. Table 6 reports the squared difference between the mean rankings of the sets of images by the raters and by the different metrics (the CatSIM variants and the space-unaware accuracy, $\mathcal{R}$, $\mathcal{A}R$ and κ). In this experiment, the CatSIM methods using accuracy, the Rand index, or Cohen's κ produce rankings more similar to those produced by human raters compared to methods that consider only pointwise measures of agreement.

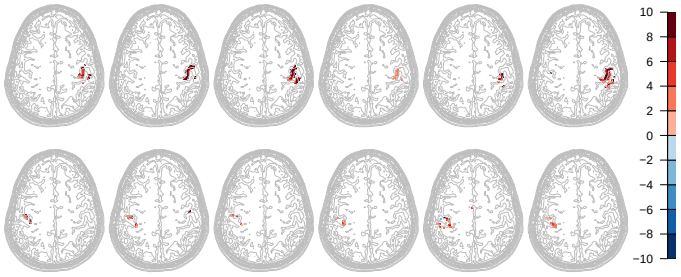
3.3 Application to Real-Data Examples

Our final evaluations are with two real-life applications, the first of which computes similarity between binary 3D image volumes and the other in 2D multinary images. Both our applications involve the use of masks in computations.

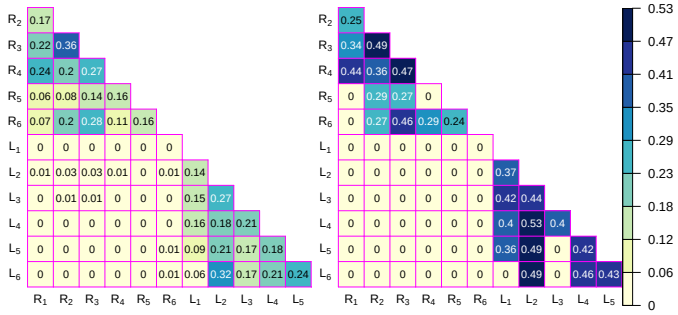
3.3.1 *Assessing Test-Retest Reliability of Activation in fMRI*

Repeatability of results across multiple fMRI [27], [28] studies is important to understand the variability of activation [29], [30] and to gauge its potential in single-subject studies [31] with a view to its adoption in clinical settings. Reliability of such activation is challenged by many factors [31], [32], not least of which is the fact that very few (no more than 3% of voxels) are expected to be truly activated. [33] introduced $\mathcal{J}$ (instead of $\mathcal{D}$) in fMRI to more finely assess reliability of activation, and a summarized version ($\ddot{\mathcal{J}}$) across K replicated studies that computes the largest eigenvalue ($\lambda_{\mathcal{J}}$) of the matrix of pairwise $\mathcal{J}$s and sets $\ddot{\mathcal{J}} = (\lambda_{\mathcal{J}} - 1)/(K - 1)$. The underlying $\mathcal{J}$ is space-unaware so here we assess whether CatSIM($\mathcal{J}$) can improve consistency of detected activation across studies.

Our data are from the replicated right- and left-hand finger-tapping experiments of [32] in which activation was detected using the AR-FAST [23] algorithm. For clarity of presentation, we restrict attention only to the six $128\times128\times128\times22$ images that had the most detected activation for each hand. Most of the activation (Figure S7) is, expectedly, in the 18th through the 21st slices encompassing the (left or right, converse to the hand used for tapping) ipsi- and contra-lateral pre-motor cortices (pre-M1), the primary motor cortex (M1), the pre-supplementary motor cortex (pre-SMA) and the supplementary motor cortex (SMA). There is wide variability of detected activation in the contra-lateral pre-M1, pre-SMA and SMA voxels. Figure 7(a) displays radiologic views of the activation in the 20th slice across the 12 experiments. Given the very small proportion of



(a) right- (top row) and left-hand (bottom) experiments.



(b) $\mathcal{J}$ (c) CatSIM($\mathcal{J}$)

Fig. 7. Activation images of the 20th slice in the finger-tapping experiments. (b,c) Graphical displays of $\mathcal{J}$ and CatSIM($\mathcal{J}$) values for each 3D volume pair, with R_i or L_i indicating ith right- or left-hand experiment.

expected activated voxels, we only use $\mathcal{J}$ and CatSIM($\mathcal{J}$). Figure 7(b) displays the pairwise $\mathcal{J}$ between the 12 3D activation image volumes. We see highest similarity between the second and the third activation images. Mild similarity between detected activation in some of the left-hand and right-hand activation maps is also reported. On the other hand, Figure 7(c) shows that such commonality of activation is likely from stray voxels and not structurally supported, as also seen by careful inspection of the activation images in Figure S7. Further, Figure 7(c) shows that there is greater reliability in the activation detected in the left-hand experiments which show greater specificity as the right-hand-dominant male shows greater focus in carrying out a left-hand task. In general, Figure 7(c) shows a wider range of values than Figure 7(b), allowing for greater discrimination. Further, $\ddot{\mathcal{J}}$, calculated using the pairwise $\mathcal{J}$s for each of the right- and left-hand experiments were both 0.189 but 0.297 and 0.361 when the pairwise CatSIM($\mathcal{J}$) values were substituted for $\mathcal{J}$ in the calculation of the summarized coefficient. While the higher values for the left-hand experiments are as expected for a right-hand-dominant subject, the generally low values of CatSIM($\mathcal{J}$) (or $\mathcal{J}$) illustrates the challenge of reliable activation detection in single-subject fMRI.

3.3.2 *Evaluating Image Segmentation Algorithms*

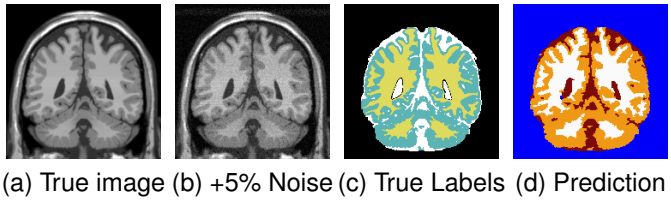
Image segmentation is important in several applications with many algorithms whose performance need to be calibrated. For example, segmenting Magnetic Resonance (MR) images into regions of gray matter, white matter, or cerebrospinal fluid is important for diagnostic purposes and important for automated image processing. We demonstrate CatSIM on a practical example using data made available by [34] who evaluate their new segmentation algorithm using

simulated datasets from BrainWeb [35]–[37], and real-data images from MRBrainS [38]. The BrainWeb interface provides multisequence (T_1-, T_2- and proton density-weighted) simulated brain images with different levels of nonhomogeneous Rayleigh noise. The MRBrainS dataset consists of multisequence (T_1-weighted, T_1-weighted inversion recovery, and T_2-weighted fluid attenuated inversion recovery) 3T MR scans of twenty subjects, manually-segmented by experts. [34] compare existing segmentation algorithms to their method which uses LSTM (long short-term memory) recurrent neural networks that account for the multi-modal nature of the data and local structure in their classification. Here we present comparisons of three of their methods – LSTM-MA (LSTM with multi-modality and adjacency constraint), SLIC-LSTM-MA (LSTM with multi-modality and super-pixel adjacency constraints), and SLIC-BiLSTM-MA (bi-directional LSTM with multi-modality and super-pixel adjacency constraints) – to k-means, support vector machines (SVM), and k-nearest neighbors (KNN) for both the MRBrainS and BrainWeb data with different amounts of Gaussian (for MRBrainS) and Rayleigh (for Brainweb) noise. Figure 8(a) shows an example baseline image, Figure 8(b) one of its noisy versions, Figure 8(c) the ground truth segmentation, and Figure 8(d) illustrates a segmentation of the noisy image obtained using SLIC-BiLSTM-MA. [34] provide five examples, three (Images C080, S099, and T075) of which are coronal, sagittal, and transverse slices from the BrainWeb data set, and two (Images 2T25 and 4T28) are transverse slices from the MRBrainS data set. They evaluate all methods using accuracy and $\mathcal{D}$. Figure 8e presents a comparison between the accuracy and the CatSIM metric (with accuracy) for these methods across the different levels of added noise. We see that the CatSIM metric gives, for almost all noise settings and across all images, the same ordering of the methods as the accuracy. However, the spread of the results is greater, meaning that, as in Section 3.3.1, we get better discrimination between the methods using CatSIM than with pointwise accuracy.

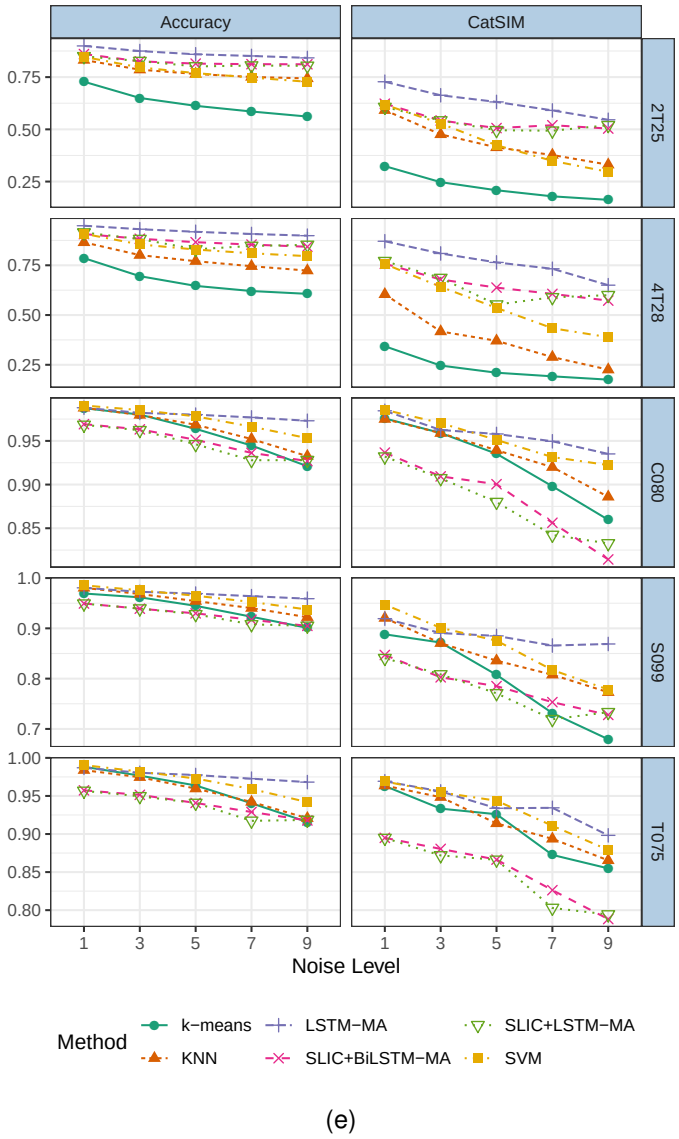
4 Discussion

We have presented a novel image similarity metric called CatSIM, implemented in an R [39] package, `catsim`, that accounts for structural similarities in binary and multinary images and extends to 3D volumes. The metric can be used with masks, which means that or can accommodate arbitrary shapes inside the images and volumes. CatSIM is more similar to human perception in ranking images and provides greater discrimination between segmentations than currently-used metrics. These findings are supported by results of experiments with artificially-created data and real data sets as well as of survey data of subjective image quality ratings. CatSIM can also flexibly deal with labels that have meaning or labels that are arbitrarily assigned.

There are a number of issues that can benefit from further attention. For instance, we could investigate the use of smoother windowing functions. Further evaluation of the weighting at different levels of the index would also be worth pursuing. Other refinements could include incorporating fuzzy class labels, different misclassification costs, or hierarchical class information within the framework.



(a) True image (b) +5% Noise (c) True Labels (d) Prediction



(e)

Fig. 8. (a) Baseline image of slice C080 from the simulated BrainWeb data set and the same image with (b) 5% Rayleigh noise added, (c) ground truth segmentation and (d) predicted segmentation of the noisy image using the SLIC-BiLSTM-MA algorithm. (e) Comparison between Accuracy and CatSIM(Accuracy) for MR images in evaluating different segmentation algorithms of MR images with different noise levels.

Acknowledgments

This research was supported in part by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health (NIH) under its Award No. R21EB016212 and by the United States Department of Agriculture (USDA)/National Institute of Food and Agriculture (NIFA), Hatch project IOW03617. The content of this paper however is solely the responsibility of the authors and does

not represent the official views of either the NIBIB, the NIH or the USDA.

REFERENCES

[1] F. B. Baulieu, "A classification of presence/absence based dissimilarity coefficients," *Journal of Classification*, vol. 6, no. 1, pp. 233–246, Dec. 1989. [Online]. Available: https://doi.org/10.1007/bf01908601

[2] J. C. Gower and P. Legendre, "Metric and euclidean properties of dissimilarity coefficients," *Journal of Classification*, vol. 3, no. 1, pp. 5–48, Mar. 1986. [Online]. Available: https://doi.org/10.1007/bf01896809

[3] A. A. Taha, A. Hanbury, and O. A. J. del Toro, "A formal method for selecting evaluation metrics for image segmentation," in *2014 IEEE International Conference on Image Processing (ICIP)*, Oct 2014, pp. 932–936.

[4] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901. [Online]. Available: https://dx.doi.org/10.5169/seals-268384

[5] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, Jul. 1945. [Online]. Available: https://doi.org/10.2307/1932409

[6] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960. [Online]. Available: https://doi.org/10.1177/001316446002000104

[7] R. W. Hamming, "Error detecting and error correcting codes," *The Bell System Technical Journal*, vol. 29, no. 2, pp. 147–160, 1950. [Online]. Available: https://dx.doi.org/10.1002/j.1538-7305.1950.tb00463.x

[8] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, Dec. 1971. [Online]. Available: https://doi.org/10.1080/01621459.1971.10482356

[9] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193–218, 1985. [Online]. Available: https://dx.doi.org/10.1007/BF01908075

[10] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, Sep. 1993.

[11] M. Prieto and A. Allen, "A similarity metric for edge images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1265–1273, Oct. 2003. [Online]. Available: https://doi.org/10.1109/tpami.2003.1233900

[12] D. B. Russakoff, C. Tomasi, T. Rohlfing, and C. Maurer, "Image similarity using mutual information of regions," in *Computer Vision - ECCV 2004*, T. Pajdla and J. Matas, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 596–607.

[13] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.

[14] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[15] M. Sampat, Z. Wang, S. Gupta, A. Bovik, and M. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2385–2401, 11 2009.

[16] A. Mason, J. Rioux, S. E. Clarke, A. Costa, M. Schmidt, V. Keough, T. Huynh, and S. Beyea, "Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of mr images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 1064–1072, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8839547/

[17] E. Allaj, "Two simple measures of variability for categorical data," *Journal of Applied Statistics*, vol. 45, no. 8, pp. 1497–1516, Sep. 2017. [Online]. Available: https://doi.org/10.1080/02664763.2017.1380787

[18] E. H. Simpson, "Measurement of diversity," *Nature*, vol. 163, no. 4148, pp. 688–688, Apr. 1949. [Online]. Available: https://doi.org/10.1038/163688a0

[19] L. Ceriani and P. Verme, "The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini," *The Journal of Economic Inequality*, vol. 10, no. 3, pp. 421–443, Jun. 2011. [Online]. Available: https://doi.org/10.1007/s10888-011-9188-x

[20] M. Levandowsky and D. Winter, "Distance between sets," *Nature*, vol. 234, no. 5323, pp. 34–35, Nov. 1971. [Online]. Available: https://doi.org/10.1038/234034a0

[21] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 48, no. 3, pp. 259–279, 1986. [Online]. Available: https://doi.org/10.1111/j.2517-6161.1986.tb01412.x

[22] E. Hoffman, P. Cutler, W. Digby, and J. Mazziotta, "3-D phantom to simulate cerebral blood flow and metabolic images for PET," *Nuclear Science, IEEE Transactions on*, vol. 37, pp. 616 – 620, 05 1990.

[23] I. A. Almodóvar-Rivera and R. Maitra, "FAST adaptive smoothed thresholding for improved activation detection in low-signal fMRI," *IEEE Transactions on Medical Imaging*, vol. 38, no. 12, pp. 2821–2828, 2019.

[24] P. J. Grother and P. A. Flanagan, "NIST handprinted forms and characters, NIST special database 19." 1995. [Online]. Available: http://www.nist.gov/srd/nistsd19.cfm

[25] N. Pya and S. N. Wood, "Shape constrained additive models," *Statistics and Computing*, vol. 25, no. 3, pp. 543–559, Feb. 2014. [Online]. Available: https://doi.org/10.1007/s11222-013-9448-7

[26] N. Pya, *scam: Shape Constrained Additive Models*, 2019, r package version 1.2-5. [Online]. Available: https://CRAN.R-project.org/package=scam

[27] N. A. Lazar, *The Statistical Analysis of Functional MRI Data*. Springer, 2008.

[28] M. A. Lindquist, "The statistical analysis of fMRI data," *Statistical Science*, vol. 23, no. 4, pp. 439–464, 2008.

[29] R. L. McNamee and N. A. Lazar, "Assessing the sensitivity of fMRI group maps," *Neuroimage*, vol. 22, no. 2, pp. 920–931, 2004.

[30] R. Gullapalli, R. Maitra, S. Roys, G. Smith, G. Alon, and J. Greenspan, "Reliability estimation of grouped functional imaging data using penalized maximum likelihood," *Magnetic Resonance in Medicine*, vol. 53, no. 5, pp. 1126–1134, 2005.

[31] R. Maitra, "Assessing certainty of activation or inactivation in test-retest fMRI studies," *NeuroImage*, vol. 47, no. 1, pp. 88–97, 2009.

[32] R. Maitra, S. Roys, and R. Gullapalli, "Test-retest reliability estimation of functional MRI data," *Magnetic Resonance in Medicine*, vol. 48, no. 1, pp. 62–70, 2002.

[33] R. Maitra, "A re-defined and generalized percent-overlap-of-activation measure for studies of fMRI reproducibility and its use in identifying outlier activation maps," *NeuroImage*, vol. 50, no. 1, pp. 124–135, 2010.

[34] K. Xie and Y. Wen, "LSTM-MA: A LSTM method with multi-modality and adjacency constraint for brain image segmentation," in *2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019*. IEEE, 2019, pp. 240–244. [Online]. Available: https://doi.org/10.1109/ICIP.2019.8802959

[35] R. K. . Kwan, A. C. Evans, and G. B. Pike, "MRI simulation-based evaluation of image-processing and classification methods," *IEEE Transactions on Medical Imaging*, vol. 18, no. 11, pp. 1085–1097, Nov 1999.

[36] R. K. S. Kwan, A. C. Evans, and G. B. Pike, "An extensible MRI simulator for post-processing evaluation," in *Visualization in Biomedical Computing*, K. H. Höhne and R. Kikinis, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, pp. 135–140.

[37] D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans, "Design and construction of a realistic digital brain phantom," *IEEE Transactions on Medical Imaging*, vol. 17, no. 3, pp. 463–468, June 1998.

[38] A. M. Mendrik, K. L. Vincken, H. J. Kuijf, M. Breeuwer, W. H. Bouvy, J. de Bresser, A. Alansary, M. de Bruijne, A. Carass, A. El-Baz, A. Jog, R. Katyal, A. R. Khan, F. van der Lijn, Q. Mahmood, R. Mukherjee, A. van Opbroek, S. Paneri, S. Pereira, M. Persson, M. Rajchl, D. Sarikaya, Örjan Smedby, C. A. Silva, H. A. Vrooman, S. Vyas, C. Wang, L. Zhao, G. J. Biessels, and M. A. Viergever, "MRBrainS challenge: Online evaluation framework for brain image segmentation in 3T MRI scans," *Computational Intelligence and Neuroscience*, vol. 2015, pp. 1–16, 2015. [Online]. Available: https://doi.org/10.1155/2015/813696

[39] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: https://www.R-project.org/

Supplement to "CatSIM: A Similarity Metric for Categorical Images and Volumes"

Geoffrey Z. Thompson and Ranjan Maitra

S1. Additional results for Section 3.1

EBI Example

Table S1 provides more detailed results of the demonstration in Section 3.1.1 of how different metrics respond to geometric distortion and the addition of salt-and-pepper noise. In addition to the metrics reported there, we include the results for CatSIM using accuracy and the Rand index as the similarity metric inside it.

TABLE S1: CatSIM and other metrics for the distorted binary images of the EBI.

Metrics	Horiz. Shift	S & P Match	Vert. Shift	S & P Match	Hor. and Vert. Shift	S & P Match
CatSIM 5 levels	0.594	0.515	0.569	0.516	0.658	0.557
CatSIM 1 level	0.464	0.092	0.449	0.090	0.561	0.110
CatSIM (whole)	0.763	0.769	0.751	0.756	0.827	0.832
CatSIM (accuracy)	0.806	0.750	0.791	0.752	0.842	0.777
CatSIM (Jaccard)	0.590	0.581	0.571	0.584	0.647	0.618
CatSIM ($\mathcal{AR}$)	0.479	0.480	0.440	0.483	0.538	0.533
CatSIM (Rand)	0.734	0.723	0.714	0.727	0.771	0.758
MS-SSIM	0.670	0.135	0.659	0.130	0.700	0.170
CW-SSIM	0.831	0.783	0.752	0.780	0.834	0.810
Accuracy	0.898	0.898	0.893	0.893	0.926	0.927
Jaccard	0.720	0.734	0.708	0.723	0.788	0.799
AdjRand	0.627	0.630	0.610	0.613	0.720	0.725
Rand	0.817	0.818	0.809	0.809	0.863	0.864
Cohen's κ	0.763	0.771	0.751	0.759	0.827	0.834

Table S2 provides more detailed results of the demonstration in Section 3.1.2. In addition to the metrics reported there, we include the results for CatSIM using accuracy and the Rand index as the similarity metric inside it.

TABLE S2: CatSIM and other metrics for the 4-class image example.

Metrics	Horiz. Shift	S & P Match	Vert. Shift	S & P Match	Hor. and Vert. Shift	S & P Match
CatSIM 5 levels	0.816	0.610	0.652	0.548	0.814	0.565
CatSIM 1 level	0.686	0.105	0.462	0.081	0.533	0.092
CatSIM (whole)	0.906	0.906	0.827	0.829	0.869	0.869
CatSIM (accuracy)	0.892	0.808	0.796	0.764	0.906	0.778
CatSIM ($\mathcal{AR}$)	0.783	0.604	0.592	0.538	0.785	0.558
CatSIM (Rand)	0.879	0.799	0.770	0.748	0.886	0.765
Accuracy	0.936	0.935	0.881	0.882	0.910	0.910
AdjRand	0.828	0.842	0.723	0.723	0.777	0.784
Rand	0.926	0.933	0.881	0.883	0.904	0.909
Cohen's κ	0.906	0.907	0.827	0.831	0.869	0.871

S2. Additional Results and Details for Section 3.2.1

A. Randomization Test for Difference in Correlations

To test whether the correlation of a metric, m_1, with the MOS (mean opinion score on the survey), M, is greater than another metric, m_2, we standardize m_1, m_2, and M and perform the following randomization test:

1) For each distorted image, swap standardized elements of m_1 and m_2 with probability 0.5 to create new vectors $\tilde{m}_1$ and $\tilde{m}_2$.
2) Compute and record $corr(M, \tilde{m}_1) - corr(M, \tilde{m}_2)$.
3) Repeat n times.
4) Define the p value as the proportion of times $corr(M, m_1) - corr(M, m_2) > corr(M, \tilde{m}_1) - corr(M, \tilde{m}_2)$.

The results of this test with $n = 100,000$ are listed in Table S3.

B. Image rankings by monotonic GAM

We fit monotonic GAMs with an identity link function and Normal random components using the metrics as explanatory variables and the MOS (mean opinion score) as the response variable. The percentage of deviance explained by the model, reported in Table S4, is then a non-parametric measure of the correspondence between the image quality metrics and the subjective quality of the images.

TABLE S3: Results of a randomization test of whether the correlation of MOS with the CatSIM method is greater than the comparison methods. Italics indicate $p < 0.05$. The left column indicates the CatSIM variant and the middle column indicates the metric it is compared to.

CatSIM metric	Comparison	p-value
CatSIM (κ)	MS-SSIM	0.311
CatSIM (κ)	CW-SSIM	*0.018*
CatSIM (κ)	Accuracy	*0.032*
CatSIM (κ)	Cohen's κ	0.203
CatSIM (κ)	AdjRand	0.123
CatSIM ($\mathcal{J}$)	Jaccard	0.374
CatSIM (Acc)	MS-SSIM	0.424
CatSIM (Acc)	CW-SSIM	0.082
CatSIM (Acc)	Accuracy	*0.029*
CatSIM (Acc)	Cohen's κ	0.369
CatSIM (Acc)	AdjRand	0.246
CatSIM ($\mathcal{A}R$)	MS-SSIM	0.410
CatSIM ($\mathcal{A}R$)	CW-SSIM	0.069
CatSIM ($\mathcal{A}R$)	Accuracy	*0.033*
CatSIM ($\mathcal{A}R$)	Cohen's κ	0.348
CatSIM ($\mathcal{A}R$)	AdjRand	0.204

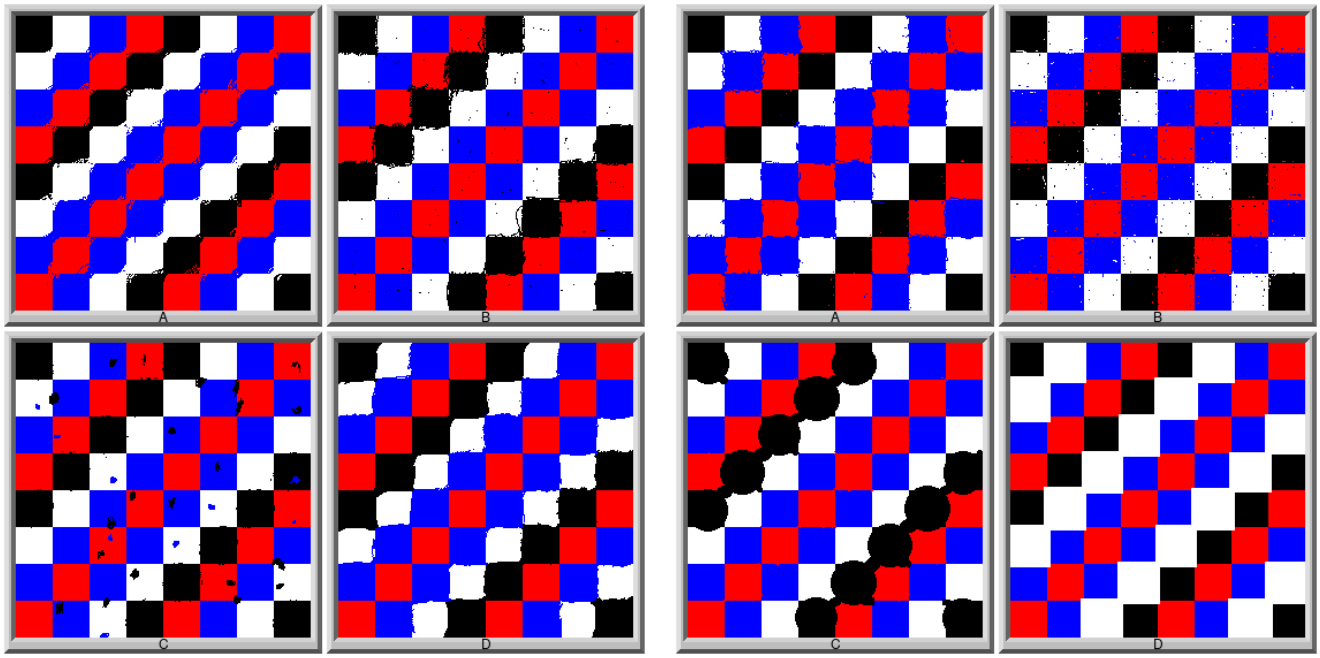
TABLE S4: The percent of deviance in MOS explained by a monotonic GAM with an identity link function and normal errors with smoothing parameter selection by generalized cross validation.

Method	% Deviance
CatSIM (Acc)	0.221
Adjusted Rand	0.323
Cohen's κ	0.329
CatSIM ($\mathcal{J}$)	0.334
Jaccard	0.348
MS-SSIM	0.356
Accuracy	0.361
CatSIM (κ)	0.371
CatSIM ($\mathcal{A}R$)	0.375
CW-SSIM	0.391

S3. Additional Details for Section 3.2.2

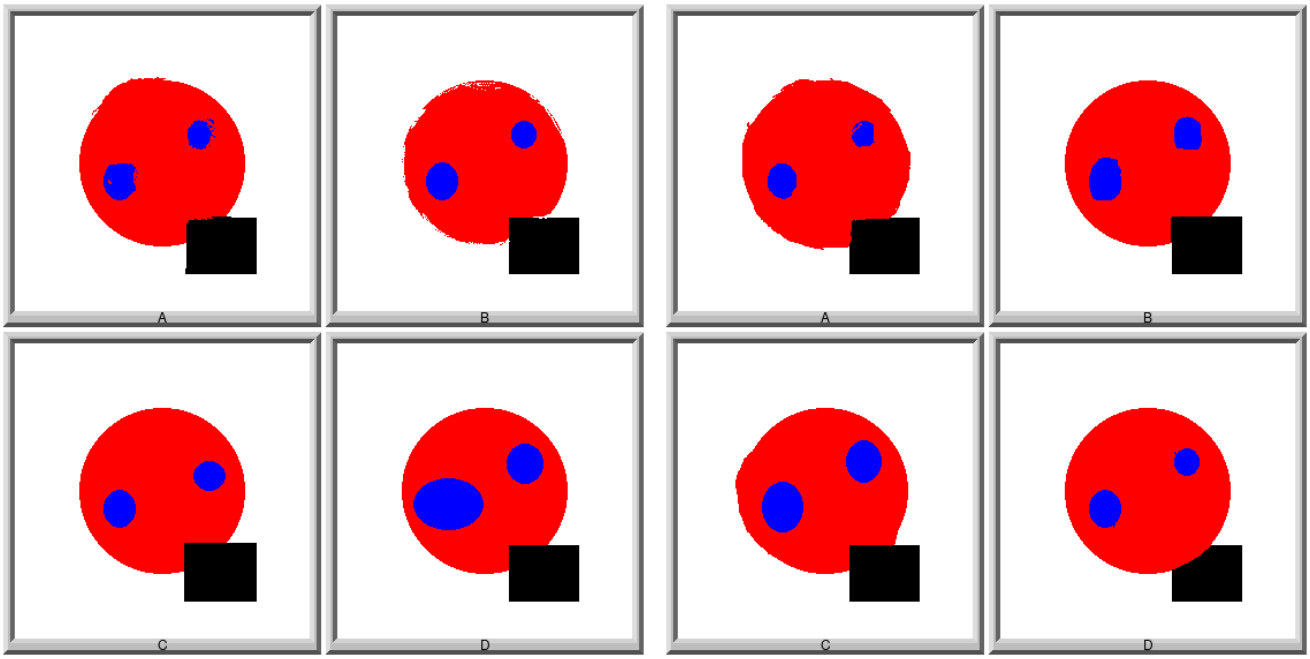
Image rankings in the categorical survey

Respondents were shown 11 sets of four distorted images. For each set of four distorted images, they were shown the true, undistorted image and asked to rank the four distorted images by their quality compared to the true image. The 11 sets were presented in a random order. Within the panels of four images, the order and labels (A, B, C, and D) were fixed. Below are the 11 sets of four distorted images shown in the survey. There were two different sets for two of the original images and one set for one of the images.



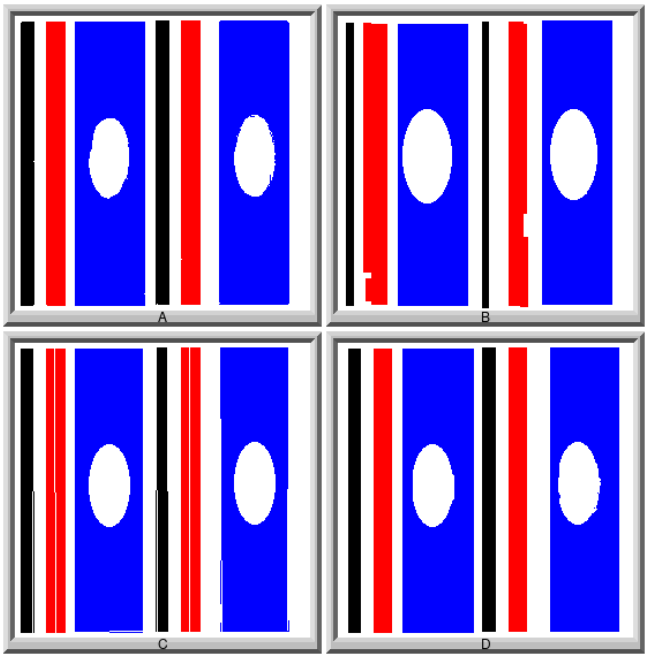
(a) Set 1-1. (b) Set 1-2.

Fig. S1: The two sets of distorted images used in the image ranking survey for image 1.

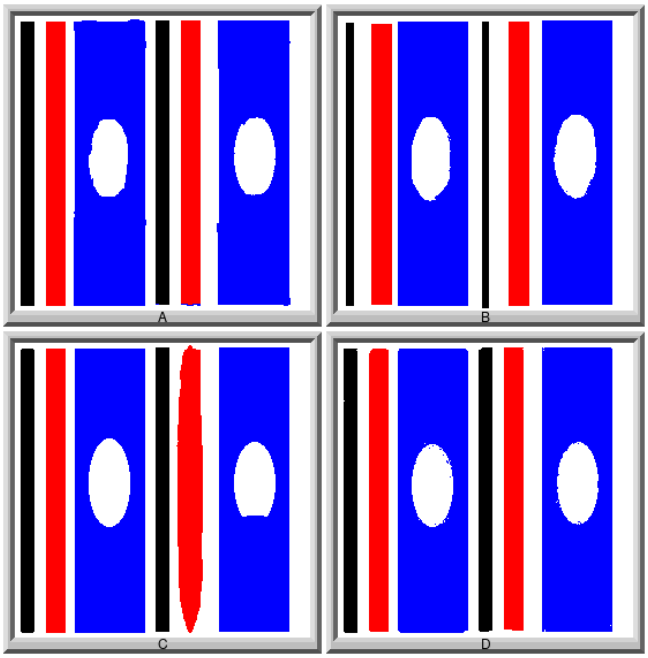


(a) Set 2-1. (b) Set 2-2.

Fig. S2: The two sets of distorted images used in the image ranking survey for image 2.



(a) Set 3-1.



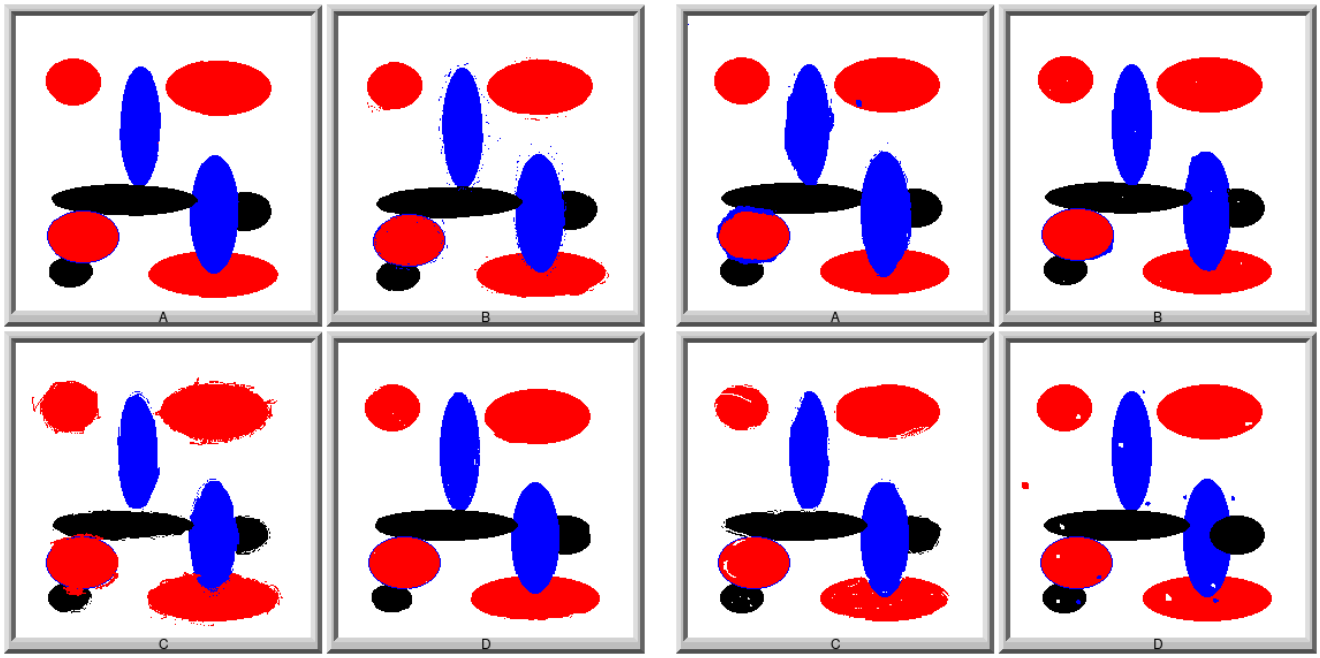
(b) Set 3-2.

Fig. S3: The two sets of distorted images used in the image ranking survey for image 3.

(a) Set 4-1.

(b) Set 4-2.

Fig. S4: The two sets of distorted images used in the image ranking survey for image 4.



(a) Set 6-1. (b) Set 6-2.

Fig. S5: The two sets of distorted images used in the image ranking survey for image 6.

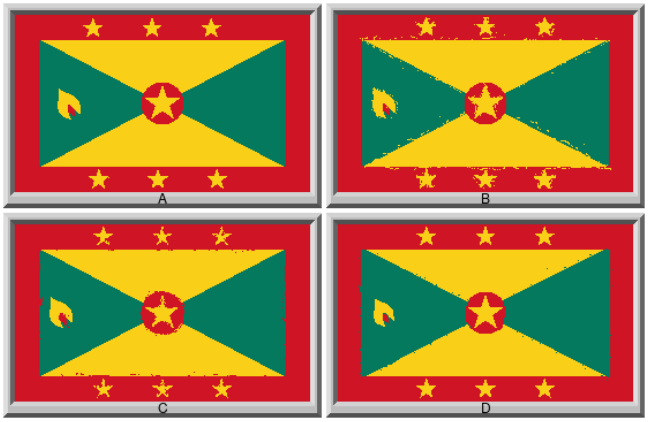


Fig. S6: The one set of distorted images used in the image ranking survey for image 5.

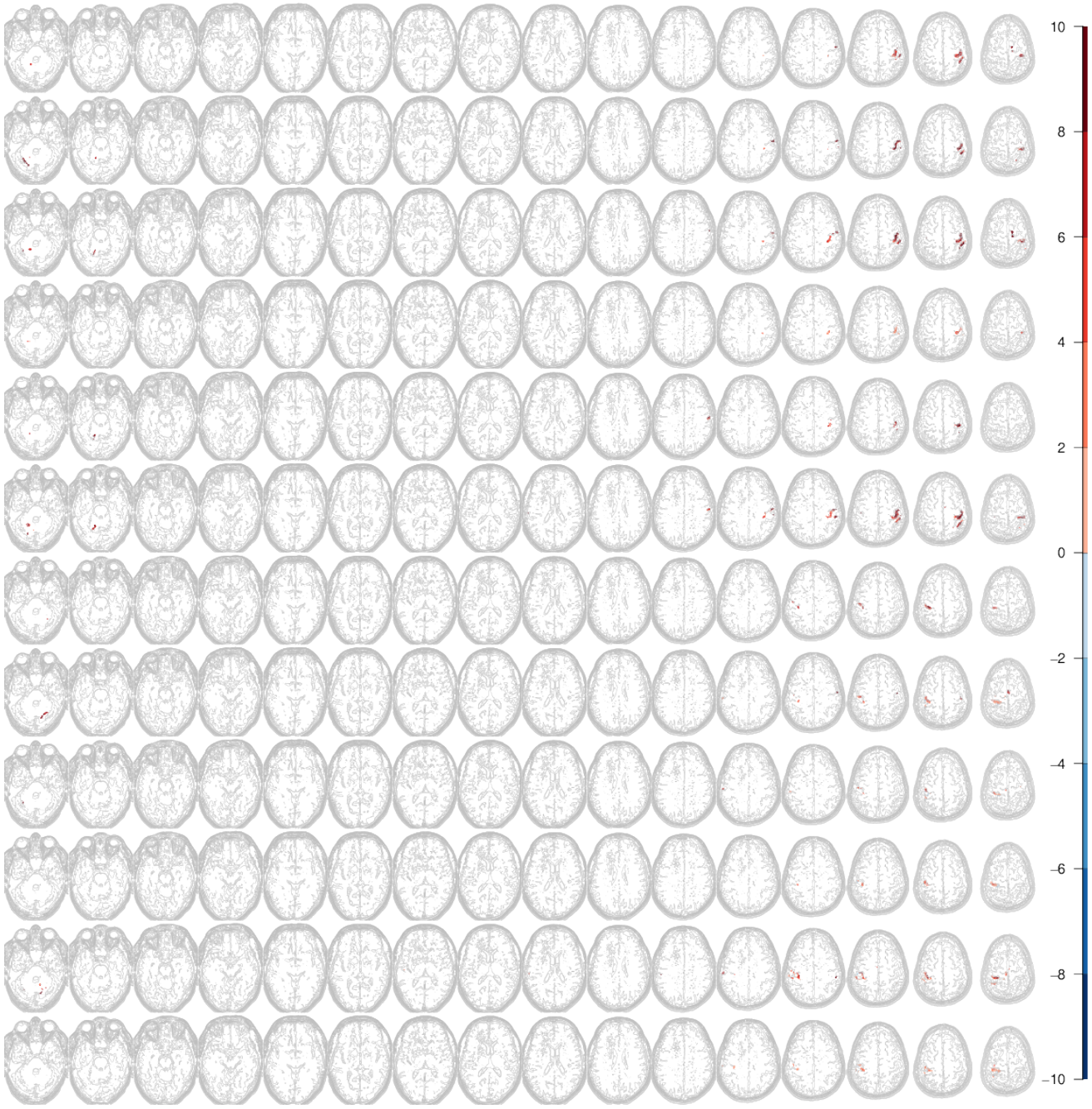


Fig. S7: Radiologic views of activation detected by AR-FAST in the top 16 slices (Slices 7-22, column-wise) and in each of the 6 right-hand (first six rows) and left-hand (last six rows) finger-tapping experiment.