Standardized Registration Methods for the SATA Challenge Datasets

Brian B. Avants[1], Nicholas J. Tustison[2], Hongzhi Wang[1], Andrew J. Asman[3], Bennett A. Landman[3]
and the Insight Software Consortium[4]

[1]Penn Image Computing and Science Lab
Dept. of Radiology
University of Pennsylvania, Philadelphia, PA, 19104
[2]Dept. of Radiology and Medical Imaging,
University of Virginia, Charlottesville, VA 22903
[3]Electrical Engineering, Vanderbilt University, Nashville, TN, USA 37235
[4]http://www.insightsoftwareconsortium.org/

Abstract. The 2012 Segmentation: Algorithms, Theory and Applications (SATA) challenge suggests that even subtle variation in registration performance may impact the outcome of multi-atlas segmentation algorithms. The 2013 SATA challenge organizers therefore requested standardized registration that enables entrants to use the same mappings as input to competing algorithms. We therefore collaborated to provide, within a relatively brief window of time, over 22,000 registration results based on Advanced Normalization Tools (`http://stnava.github.io/ANTs/`). The diencephalon component of the challenge presented familiar and easily addressed data requiring only 1,600 mappings between different 3D human T1 neuroimages. The 3D multiple modality MRI "dog leg" dataset ($>$ 7,000 mappings) presented the opportunity to improve performance by using a multivariate similarity metric. The 4D cardiac (or CAP) dataset ($\approx$ 13,000 mappings) includes highly variable image quality, anatomy and field of view. We detail the ANTs variants that address the most basic brain dataset, where we used a template-based approach, to the more challenging CAP dataset which employed a more customized registration solution based on prior knowledge. The scripts, source code and a small set of example data accompany this paper and are available online. [1]

1 Introduction

Several early papers in image registration / segmentation focus on clinical applications including microscopy, nuclear medicine, tracking longitudinal change and angiography [1,2,3,4,5]. Rapid progress in registration theory and implementation followed these early developments with focus, in general, on either

[1] This work is supported by National Library of Medicine sponsored ARRA stimulus funding. Resources for this report are here (link).

similarity metrics [6], transformation models [7] or the combination of these in a general purpose method [8]. More recently, as the core technology of segmentation and registration mature, the research focus has returned to specific clinical applications such as radiation therapy [9], drug discovery [10], neurosurgery [11] and image-based biomarkers of pain [12]. Quantitative image mapping is also relied upon heavily for automated analysis of large datasets such as ADNI [13] that cannot be processed "by hand."

These new problem domains are revealing the limitations of traditional methods. Brain registration methods, for instance, often assume that a single reference space / template is adequate to combine information across a population and that there exists, roughly, a diffeomorphism between any pair of individual brains. Segmentation methods, on the other hand, often assume that intensity and object smoothness or texture encode the shape to be segmented or that a single model object is sufficient to act as a prior. These assumptions may not hold when imaging other organs or in emerging MR modalities.

Multi-atlas segmentation and registration (MASR) methods help overcome the limitations associated with mapping to a single template or using a single exemplar segmentation. These methods encode the notion that different topology or image contrast may exist between different subjects and, further, that the user cannot know these subjects *a priori*. In this case, the problem is to develop algorithms that automatically adjust for registration accuracy at both global and local scales and determine segmentation labels that incorporate confidence. MASR therefore enables applications that do not depend upon successful pairwise registration between all subjects, a goal that is currently beyond the reach of any general purpose registration tool.

The organizers of the SATA 2013 challenge collected datasets that frame, with specific examples, the issues raised above. The organizers wrote on May 10, 2013 "in order to isolate the benefits of specific label fusion algorithms we need to have a standardized registration." The ANTs development team agreed to collaborate on a set of standard registration results for the challenge. We felt that our participation was important due to the variety of the challenge data, the fact that not all entrants are biomedical image registration experts and the results of SATA 2012 which suggested that a common transformation basis set would help isolate the differences between the segmentation component in MASR. In other words, using a common set of "voters" allows a more direct comparison of the segmentation/fusion methods that are currently at the heart of the SATA challenge. We present, below, the ANTs techniques that address the unique hurdles as well as opportunities afforded by this diverse collection of images.

2 Methods

Three MRI datasets constitute the driving biological problems for the 2013 challenge: T1-weighted whole head images of the diencephalon (mid-brain), multimodality T2w and T2FS canine leg images and a 4D cardiac atlas dataset. The

SATA organizers requested that we provide transformed label sets and intensity images for all training↔testing and training↔training pairs. The registration goal was to map the images with “reasonable and consistent” performance and an open-source system.

We processed all data with Advanced Normalization Tools (git commit e3ba1c527fa63cc524bae7912b415d77adadc165) based on ITKv4. The processing was organized with a combination of bash or perl scripts and run through standard Sun Grid Engine distributed computing. The core methods are available in ANTs programs `antsRegistration`, `ImageMath` (utilities), `N4BiasFieldCorrection` (bias correction) and `Atropos` (segmentation). These core tools are described in archived publications [14,15,16]. However, we employed several new additions to ANTs registration that are as recent as 2013 and which are based upon our work for version 4 of the Insight ToolKit.

2.1 The (Fire) ANTs System

ANTs fundamentals involve pairwise registration of images defined in physical space. ANTs registrations, as of 2013 with the Fire beta, are defined by an initial transformation (possibly the identity) of the “fixed” and/or “moving” images. This is followed by a series of registration stages typically with increasing degrees of freedom. Each stage begins with k similarity metric definitions (most often $k = 1$) and is followed by parameters defining the multi-resolution strategy. Subsequent stages may be added in the same style. Finally, ANTs takes options controlling optimization details and pre/post-processing. Application-specific choices should be made for the transformation, similarity metrics and multi-resolution strategy. A variety of options are exercised in SATA processing.

2.2 Diencephalon/Brain Data (SATA-B)

> “... the mid-brain and canine leg datasets should not be terribly difficult ...” —*SATA Organizers*

We employed an efficient “pseudo-geodesic” processing strategy for this data to avoid explicitly computing ≈1,600 image registration instances, as shown in Figure 1. As was shown in previous work (technical report), this strategy improves upon single template labeling while approaching (not equaling) the more computationally expensive all-pairs registration required by standard MASR.

We first derive a custom whole-head template for this dataset and extract the template cerebrum using MASR and LPBA40 labels [17]. We then extract the cerebrum from the individual SATA-B images based on existing ANTs software (`extractBrainAllSubjects.pl`) for template-based brain mapping. We then re-registered the individual cerebrum images to the template cerebrum (`runSyNRegistrations.pl`). The subject n to template mapping composes with the inverse of the subject m to template mapping to create the final mapping between subject n and subject m. The affine and deformable components of the

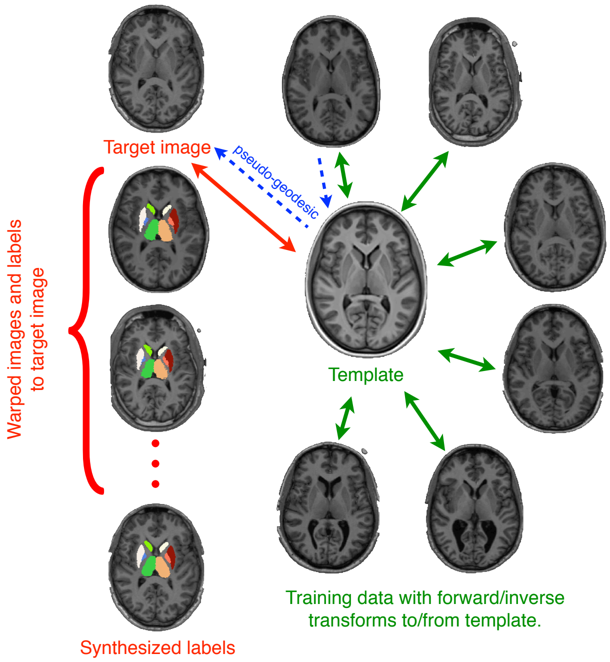
Target image
pseudo-geodesic
Warped images and labels to target image
Template
Training data with forward/inverse transforms to/from template.
Synthesized labels

Fig. 1: SATA-B registrations map every image to every other image through the optimal template space (available online). This requires only n registrations rather than n^2.

SATA-B maps derive from standard `antsRegistration` parameters. The template is available in the `data/SATA-B-Template/` directory. Similarity metric summaries of performance for each subject is in Figure 4.

2.3 Dog Leg Multi-Modality Data (SATA-L)

The input modalities are bias corrected by N4 [14] before further processing (`normalization.sh`). ANTs registration employs both modalities provided for SATA-L as in Figure 2. We did not know the relative contribution of each modality to accuracy so set equal weights. The affine and deformable components of the SATA-L maps derive from a multivariate extension to standard `antsRegistration` parameters. That is, we use the standard multi-resolution affine approach combined with SyN but drive both transformation stages by the sum of mutual information across both modalities. More specifically, for affine registration, we optimize:

$$w_1 MI(I_1, J_1(\phi(x))) + w_2 MI(I_2(\psi_I(x)), J_2(\psi_J(\phi(x)))) \\ \text{Subject to: } \phi \in Diff_{\text{Affine}}, \quad (1)$$

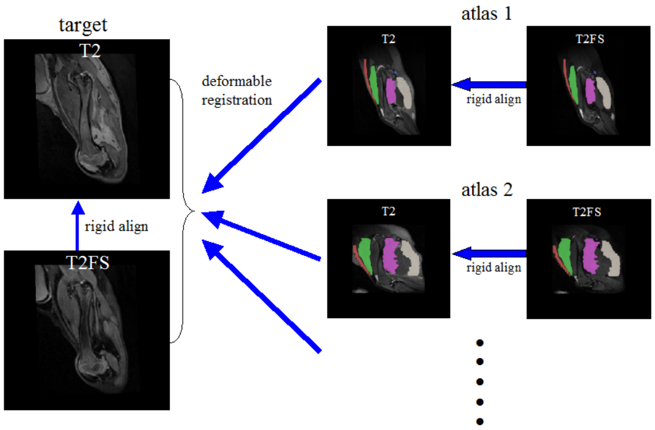

Fig. 2: The SATA-L registration is similar to SATA-B. However, the image similarity is driven by the mutual information applied to both modalities. We chose to treat both modalities equally except for one small detail—one of the modalities is rigidly registered, intra-subject, to the other before proceeding with inter-subject registration.

where $w_1 = w_2 = 0.5$ and ψ_J rigidly maps J_2 to J_1 (similarly ψ_I). The script `align.sh` shows how to compute ψ. We use the same multivariate metric for SyN deformable mapping (`registerPairsMod.sh`). Note that `antsRegistration` allows initial mappings to be provided thus reducing accumulation of interpolation error. Five percent regular subsampling of the joint intensities estimates the mutual information metric and metric derivative.

2.4 CAP Cardiac Data (SATA-C)

The CAP dataset includes a volumetric time series of the beating heart:

> " ... I wouldn't be surprised if getting 'reasonable' & 'consistent' registrations for this data is difficult or impossible." —*SATA Organizers*

A significant confound is the out-of-plane spacing which is approximately 6 times that of in-plane spacing. Other challenges include significant variability in the anatomical field-of-view and orientation, the resolution/dimensions of the images, the image intensity pattern/bias and in the shape/appearance of the myocardium (the target anatomy) with respect to the remainder of the anatomy. The dataset indeed proves to be challenging to process with a general purpose toolkit like ANTs. Our procedure is illustrated with example data in `https://github.com/stnava/LabelMyHeart`. [2]

[2] May not be the exact scripts used due to a few last minute changes stored on the distributed computing system.

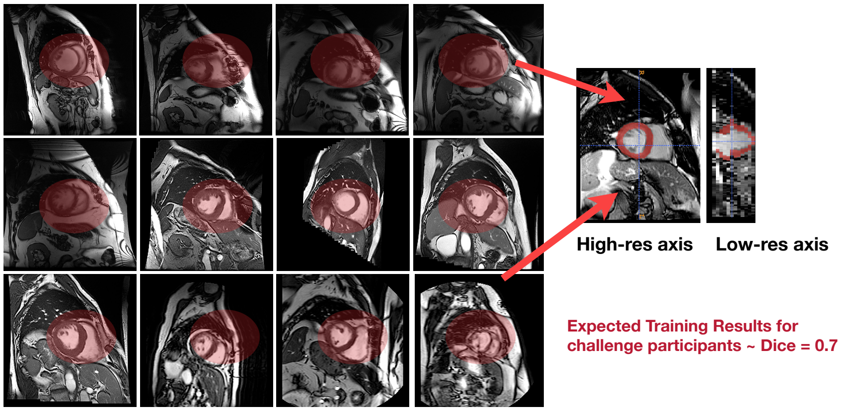
High-res axis
Low-res axis
Expected Training Results for
challenge participants ~ Dice = 0.7

Fig. 3: CAP methods focus the registration within a dilated version of the training image myocardium labeling. Restricting the registration focus improves robustness to extra-cardiac features.

The processing occurs in two steps. The first is *within-subject* and creates a single-subject template and label image by combining the time series MRI and labelset. We achieve this by `makeSubjectTemplate.sh` which extracts every time point, averages the time points, and registers all time points to the average with a simple SyN registration based on intensity difference. The single-subject template label is estimated by majority voting (in `ImageMath`). The second step is *between-subject* and exploits the fact that a training image is involved in every registration pairing. We use the label within the training image to focus the registration optimization in the region of the heart. This strategy effectively increases the robustness of the registration to the presence of inhomogeneity, altered field-of-view and anatomical variation not relevant to the myocardium labeling problem. Other details are in `betweenSubjectsMap.sh` with perhaps the key step being `-x [ $outnmmask.nii.gz ]` which focuses the registration optimization within the mask, as defined by morphological operations on the training image myocardium labeling, as in Figure 3. Training labels are mapped back to the 4D space by composing the between subject map with the inverse of the initial single subject template SyN mappings. A data-driven soft-evaluation of the performance is in Figure 5.

3 Discussion

The methods described above were informed by experience, the limited allotted time for processing and the methods available, contemporaneously, within ANTs, Fire beta. The ANTs system is intended to be available to the community such that our results may be reproduced/extended and improved by the community. OSX and LINUX are most actively supported with occasional support for

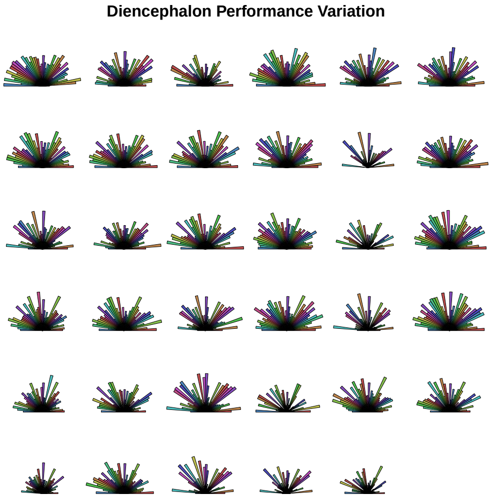

Fig. 4: We use urchin plots to visualize pairwise registration for diencephalon data. The performance is more uniform in comparison to CAP data. The distribution of the correlation metrics exhibited gaussian structure. See the Figure 5 caption for more explanation.

Windows. The processing decisions that we made in the standard registration suggest several possible outcomes and future work, discussed below.

3.1 Diencephalon/Brain Data (SATA-B)

The most familiar problem of the three presented few difficulties. We did, however, elect to extract the cerebrum from each image in order to better focus the results on relevant image features, a strategy that is similar to that used in the CAP data. Despite this approach, our pseudo-geodesic focused on the whole brain should be suboptimal for the diencephalon data. Entrants that choose to either perform pairwise registration between all subjects or focus the registration on the diencephalon should find better performance than the standard approach. We expect that entrants should find Dice overlap that is consistent with previous challenges.

3.2 Dog Leg Multi-Modality Data (SATA-L)

We employed very little problem-specific optimization for the SATA-L registration. It did, however, allow us to test whether two modalities jointly improve registration results. Based on a comparison of single-modality registration and dual modality registration in the training data, we found the dual modality similarity metric appeared to increase performance. We expect that Dice overlap will be reasonable (≈ 0.8 overall as estimated by H.W.) for the canine leg dataset.

3.3 CAP Cardiac Data (SATA-C)

The cardiac dataset reveals some of the limitations involved with naive pairwise registration and highlights the benefit of problem-specific strategies and multi-atlas methods. We believe that, of the 3 datasets, the biggest future performance gains may be possible in this type of data. Figure 5 suggests that, for some subjects, particular pairings provide a much better matching than the majority of pairings. We tested this hypothesis by looking at the non-gaussianity of the global correlation between the registered image pairs as calculated within the target subject space. After FDR-correction, this test suggested that 13.5% of subjects had a highly non-gaussian similarity distribution. For these individuals, a small portion of training images provide a reasonable result while the majority might be classified as failures. This accentuates the importance of MASR schemes that are robust to high failure rate input data or the need to cluster the populations before MASR. While the organizers requested that the output labels be consistent with the original 4D manual labels, the intra-rater variability in labeling the time series is unknown; at first glance, it would seem difficult to consistently label the myocardium in 3D over several dozen or more time frames. Therefore, the upper limits on accuracy remain unclear. However, we estimated that, with the standard registration, entrants should achieve up to average Dice overlap of 0.6 to 0.7 as estimated by H.W.

4 Conclusions

Open challenges provide crucial benchmarks to the biomedical image analysis field and we are honored to play a role in this process. The organizers shared excellent challenge data with the community that should yield exciting results that will advance both image registration and image segmentation technology. In particular, these datasets bring into focus the importance of multiple modality data, other organ systems and even other species, as well as efficient registration solutions and the need for MASR methods that are robust to input data with low reliability. The SATA data engaged ANTs mean squared intensity difference, cross-correlation and multivariate mutual information metrics along with the range of ANTs transformations, initialization, bias correction and masking options. We hope that the solutions we provided to the community achieved the goal set forth: “provide reasonable & consistent registration solutions” that lead to interesting SATA outcomes.

References

1. Badran, A.K., Fisher, A.C., Durrani, T.S., Paul, J.P.: An automatic matching technique for patient alignment. J Biomed Eng **13**(4) (Jul 1991) 281–286

2. Venot, A., Liehn, J.C., Lebruchec, J.F., Roucayrol, J.C.: Automated comparison of scintigraphic images. J Nucl Med **27**(8) (Aug 1986) 1337–1342
3. Venot, A., Leclerc, V.: Automated correction of patient motion and gray values prior to subtraction in digitized angiography. IEEE Trans Med Imaging **3**(4) (1984) 179–186
4. Wrigley, N.G., Chillingworth, R.K., Brown, E., Barrett, A.N.: Multiple image integration: a new method in electron microscopy. J Microsc **127**(Pt 2) (Aug 1982) 201–208
5. Adair, T., Karp, P., Stein, A., Bajcsy, R., Reivich, M.: Technical note. computer assisted analysis of tomographic images of the brain. J Comput Assist Tomogr **5**(6) (Dec 1981) 929–932
6. Wells, 3rd, W., Viola, P., Atsumi, H., Nakajima, S., Kikinis, R.: Multi-modal volume registration by maximization of mutual information. Med Image Anal **1**(1) (Mar 1996) 35–51
7. Miller, M.I., Beg, M.F., Ceritoglu, C., Stark, C.: Increasing the power of functional maps of the medial temporal lobe by using large deformation diffeomorphic metric mapping. Proc Natl Acad Sci U S A **102**(27) (Jul 2005) 9685–9690
8. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J.: Nonrigid registration using free-form deformations: application to breast mr images. IEEE Trans Med Imaging **18**(8) (Aug 1999) 712–721
9. Chan, M.K.H., Kwong, D.L.W., Ng, S.C.Y., Tong, A.S.M., Tam, E.K.W.: Experimental evaluations of the accuracy of 3d and 4d planning in robotic tracking stereotactic body radiotherapy for lung cancers. Med Phys **40**(4) (Apr 2013) 041712
10. Fox, G.B., Chin, C.L., Luo, F., Day, M., Cox, B.F.: Translational neuroimaging of the cns: novel pathways to drug development. Mol Interv **9**(6) (Dec 2009) 302–313
11. Omara, A.I., Wang, M., Fan, Y., Song, Z.: Anatomical landmarks for point-matching registration in image-guided neurosurgery. Int J Med Robot (Jun 2013)
12. Loggia, M.L., Kim, J., Gollub, R.L., Vangel, M.G., Kirsch, I., Kong, J., Wasan, A.D., Napadow, V.: Default mode network connectivity encodes clinical pain: an arterial spin labeling study. Pain **154**(1) (Jan 2013) 24–33
13. Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., Morris, J.C., Petersen, R.C., Saykin, A.J., Schmidt, M.E., Shaw, L., Siuciak, J.A., Soares, H., Toga, A.W., Trojanowski, J.Q., , A.D.N.I.: The alzheimer's disease neuroimaging initiative: a review of papers published since its inception. Alzheimers Dement **8**(1 Suppl) (Feb 2012) S1–68
14. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4itk: improved n3 bias correction. IEEE Trans Med Imaging **29**(6) (Jun 2010) 1310–1320
15. Avants, B.B., Tustison, N.J., Wu, J., Cook, P.A., Gee, J.C.: An open source multivariate framework for n-tissue segmentation with evaluation on public data. Neuroinformatics **9**(4) (Dec 2011) 381–400
16. Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C.: A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage **54**(3) (Feb 2011) 2033–2044
17. Shattuck, D.W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K.L., Poldrack, R.A., Bilder, R.M., Toga, A.W.: Construction of a 3d probabilistic atlas of human cortical structures. Neuroimage **39**(3) (Feb 2008) 1064–1080

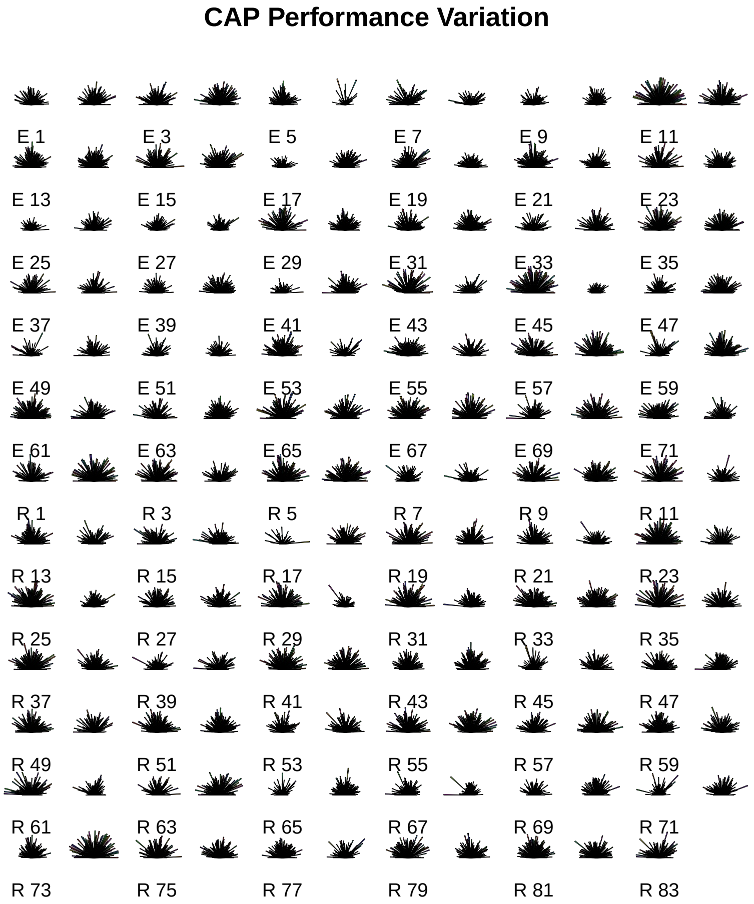

Fig. 5: We use urchin plots to visualize pairwise registration performance for CAP data. We quantified variability in all 12,865 data points by measuring the global correlation of the deformed image and the target image, i.e. both training (labeled R⋆) and testing (labeled E⋆). “Spiky” distributions of the correlation function indicate that a few pairings performed much better/worse than the majority, at least at a global level. This was verified by testing the gaussianity of the resulting correlation distribution. As an example, the sixth subject from top left (E6) has a highly non-gaussian distribution of similarity metric performance and results in a significant p-value for the non-gaussianity test ($p < 1.e\text{-}7$). Overall, 13.5% of subjects exhibited non-gaussian distributions for this function, after multiple comparisons correction by false discovery rate.