École de technologie supérieure
Department of Software Engineering
and Information Technology
Neuro-iX

Presentation of Five Papers

Brain Segmentation: From Manual Methods to Traditional Automated Approaches, Leading to Transformer-Based Models

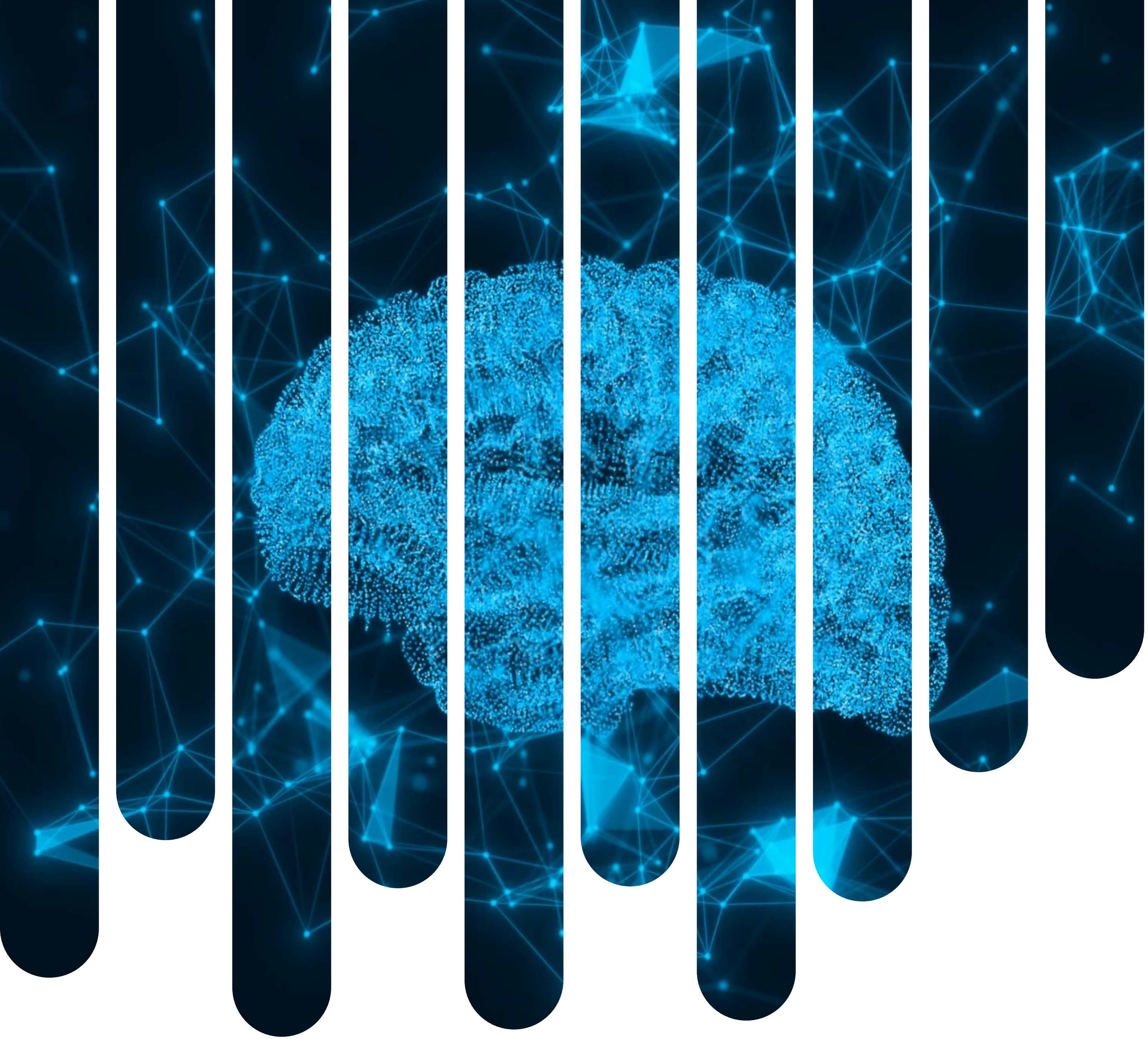
Prepared by:
Ahmed REKIK

Supervised by:
M.Sylvain BOUIX

Course MTR871: Directed Readings

Master's in Information Technology

Session Winter 2025



Contents

1

General Introduction

Presentation of the Five Papers

Context and Objective:

- These five papers explore different approaches in medical imaging and image processing
- Focusing on segmentation techniques and advanced models to enhance image analysis.

Manual Segmentation
(Paper 1)

- Reference approach to neuroanatomy and medical imaging.
- Experts manually segment the structures of interest on MRI images, guaranteeing high anatomical accuracy.

Traditional Methods
(Papers 2 & 3)

- Use of pre-segmented anatomical atlases to automatically annotate new MRI images.
- Anatomically validated approach, useful for accurate analysis.

Transformer-Based Models
(Papers 4 & 5)

- Transformer models apply self-attention to images by dividing them into patches or hierarchical blocks, enabling a detailed analysis of spatial relationships.
- They capture complex connections between anatomical structures and reduce reliance on atlases.

2

Manual Segmentation: Reference Standard

(Paper 1)

Context & Methodology

Manual Brain Structure Segmentation

- Reference method in neuroanatomy, ensuring validation of automated segmentation.
- **Critical for deep learning**, serving as a ground truth for training and evaluating segmentation models.

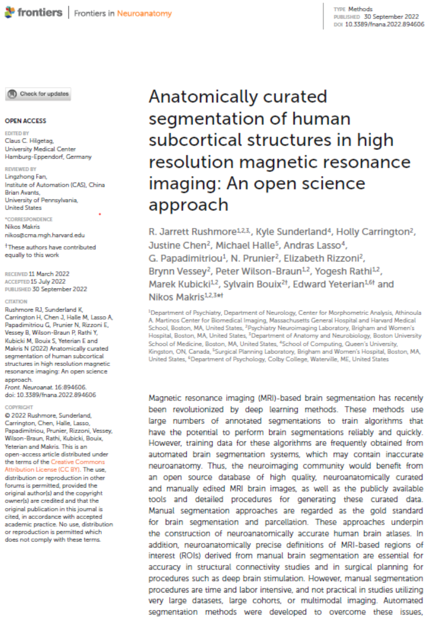
Methodology

Dataset:

- High-resolution T1-weighted MRI scans from the Human Connectome Project (50 subjects, 25M/25F).
- Definition of brain structures according to rigorous criteria.

3D Slicer Segmentation Tool:

- NeuroSegmentation module for precise manual editing.
- Use of advanced tools (histogram, anatomical marking).
- Work assisted by a graphics tablet and stylus for improved precision.

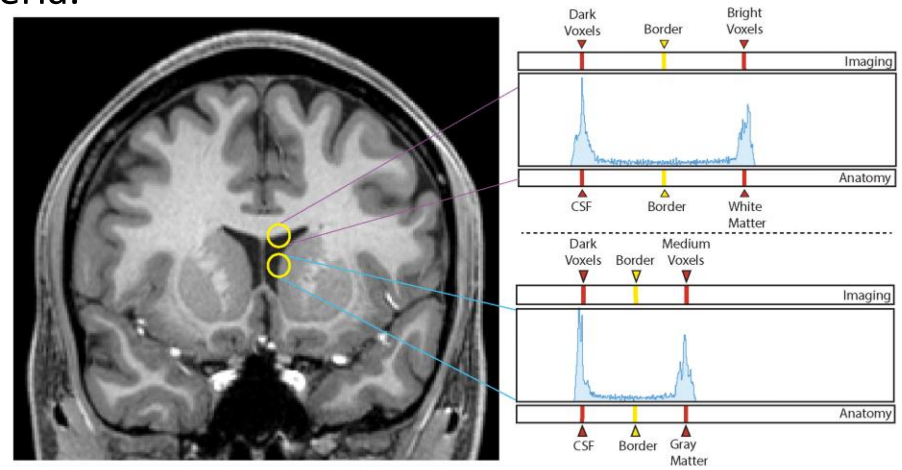

frontiers | Frontiers in Neuroanatomy

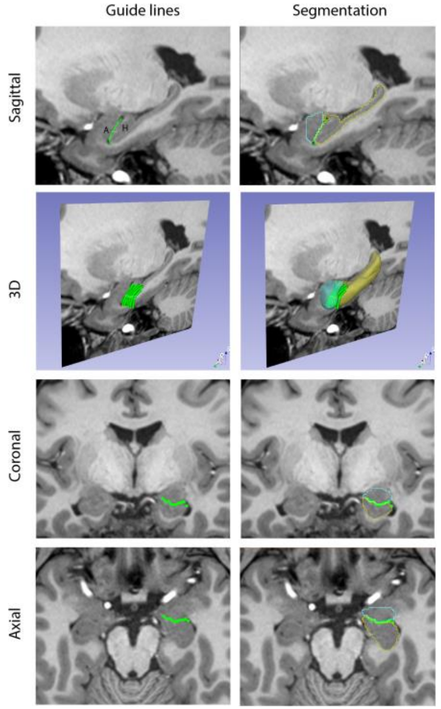
Anatomically curated segmentation of human subcortical structures in high resolution magnetic resonance imaging: An open science approach

R. Jarrett Rushmore[1,2,3], Kyle Sunderland[4], Holly Carrington[2], Justine Chen[2], Michael Halle[5], Andras Lasso[4], G. Papadimitriou[1], N. Prunier[2], Elizabeth Rizzoni[2], Brynn Vessey[2], Peter Wilson-Braun[1,2], Yogesh Rathi[1,2], Marek Kubicki[1,2], Sylvain Bouix[2†], Edward Yeterian[1,6†] and Nikos Makris[1,2,3*†]

[1]Department of Psychiatry, Department of Neurology, Center for Morphometric Analysis, Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Boston, MA, United States, [2]Psychiatry Neuroimaging Laboratory, Brigham and Women's Hospital, Boston, MA, United States, [3]Department of Anatomy and Neurobiology, Boston University School of Medicine, Boston, MA, United States, [4]School of Computing, Queen's University, Kingston, ON, Canada, [5]Surgical Planning Laboratory, Brigham and Women's Hospital, Boston, MA, United States, [6]Department of Psychology, Colby College, Waterville, ME, United States

Magnetic resonance imaging (MRI)-based brain segmentation has recently been revolutionized by deep learning methods. These methods use large numbers of annotated segmentations to train algorithms that have the potential to perform brain segmentations reliably and quickly. However, training data for these algorithms are frequently obtained from automated brain segmentation systems, which may contain inaccurate neuroanatomy. Thus, the neuroimaging community would benefit from an open source database of high quality, neuroanatomically curated and manually edited MRI brain images, as well as the publicly available tools and detailed procedures for generating these curated data. Manual segmentation approaches are regarded as the gold standard for brain segmentation and parcellation. These approaches underpin the construction of neuroanatomically accurate human brain atlases. In addition, neuroanatomically precise definitions of MRI-based regions of interest (ROIs) derived from manual brain segmentation are essential for accuracy in structural connectivity studies and in surgical planning for procedures such as deep brain stimulation. However, manual segmentation procedures are time and labor intensive, and not practical in studies utilizing very large datasets, large cohorts, or multimodal imaging. Automated segmentation methods were developed to overcome these issues,





Manual Segmentation – Results & Limitations

Reliability and Limitations of Manual Segmentation

Results

- High inter- and intra-rater reliability (Dice Coefficient > 0.90 for most structures).

- Detected cerebral asymmetries:

Ex :larger nucleus accumbens on the left, hippocampus larger on the right.

- Sex-based anatomical differences: subcortical structures appear larger in males.

Limitations

- Extremely time- and labor-intensive.

- Difficult to scale for large datasets.

- Strong dependence on expert annotators.

Region of interest (ROI)	Mean Dice	SD	Min	Max
Lateral Ventricle Left	0.95	0.02	0.92	0.98
Lateral Ventricle Right	0.95	0.02	0.93	0.98
Third Ventricle	0.84	0.05	0.75	0.90
Fourth Ventricle	0.87	0.04	0.80	0.94
Nucleus Accumbens Left	0.84	0.04	0.78	0.89
Nucleus Accumbens Right	0.84	0.05	0.76	0.93
Caudate Left	0.93	0.01	0.91	0.96
Caudate Right	0.93	0.02	0.88	0.96
Putamen Left	0.93	0.02	0.91	0.99
Putamen Right	0.93	0.02	0.91	0.99
Globus Pallidus Left	0.83	0.04	0.76	0.90
Globus Pallidus Right	0.81	0.06	0.73	0.90
Brainstem	0.95	0.01	0.94	0.98
Thalamus Left	0.88	0.04	0.78	0.92
Thalamus Right	0.88	0.03	0.82	0.93
Ventral Diencephalon Left	0.88	0.02	0.84	0.92
Ventral Diencephalon Right	0.88	0.01	0.85	0.90
Inferior Horn of Lateral Ventricle Left	0.72	0.05	0.61	0.82
Inferior Horn of Lateral Ventricle Right	0.72	0.05	0.61	0.81
Hippocampal Formation Left	0.87	0.03	0.82	0.90
Hippocampal Formation Right	0.87	0.02	0.82	0.90
Amygdala Left	0.84	0.03	0.78	0.88
Amygdala Right	0.80	0.05	0.71	0.88
Fifth Ventricle	0.76	0.07	0.65	0.84
Optic Chiasm	0.74	0.15	0.54	0.95

Inter-rater reliability

Region of interest (ROI)	Mean Dice	SD	Min	Max
Lateral Ventricle Left	0.95	0.02	0.93	0.97
Lateral Ventricle Right	0.96	0.03	0.93	0.98
Third Ventricle	0.89	0.04	0.84	0.91
Fourth Ventricle	0.90	0.03	0.87	0.93
Nucleus Accumbens Left	0.87	0.04	0.82	0.89
Nucleus Accumbens Right	0.89	0.02	0.87	0.91
Caudate Left	0.93	0.03	0.91	0.96
Caudate Right	0.94	0.02	0.92	0.95
Putamen Left	0.94	0.02	0.92	0.95
Putamen Right	0.94	0.02	0.91	0.95
Globus Pallidus Left	0.81	0.05	0.77	0.86
Globus Pallidus Right	0.80	0.05	0.76	0.85
Brainstem	0.96	0.01	0.95	0.97
Thalamus Left	0.91	0.01	0.91	0.92
Thalamus Right	0.91	0.03	0.88	0.93
Ventral Diencephalon Left	0.90	0.01	0.89	0.91
Ventral Diencephalon Right	0.90	0.01	0.89	0.91
Inferior Horn of Lateral Ventricle Left	0.76	0.09	0.68	0.86
Inferior Horn of Lateral Ventricle Right	0.80	0.04	0.75	0.83
Hippocampal Formation Left	0.90	0.02	0.89	0.93
Hippocampal Formation Right	0.90	0.05	0.85	0.94
Amygdala Left	0.84	0.07	0.77	0.91
Amygdala Right	0.84	0.07	0.76	0.90
Fifth Ventricle	0.75	0.08	0.66	0.83
Optic Chiasm	0.87	0.18	0.66	0.99

Intra-rater reliability

3

Traditional Automatic Segmentation Approaches

(Papers 2 & 3)

Atlas-Based Probabilistic Segmentation

Neuron, Vol. 33, 341–355, January 31, 2002, Copyright ©2002 by Cell Press

Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain

Neurotechnique

Bruce Fischl,[1] David H. Salat,[1] Evelina Busa,[1] Marilyn Albert,[2,3] Megan Dieterich,[5] Christian Haselgrove,[5] Andre van der Kouwe,[1] Ron Killiany,[4] David Kennedy,[5] Shuna Klaveness,[5] Albert Montillo,[6] Nikos Makris,[5] Bruce Rosen,[1] and Anders M. Dale[1,7]
[1] Massachusetts General Hospital
Nuclear Magnetic Resonance Center
Rm. 2328, Building 149
13th Street
Charlestown, Massachusetts 02129
[2] Department of Neurology
Massachusetts General Hospital
Harvard Medical School
55 Fruit Street, VBK 901
Boston, Massachusetts 02114
[3] Department of Psychiatry, CNY-9
Massachusetts General Hospital
Boston, Massachusetts 02114
[4] Department of Anatomy and Neurobiology
Boston University School of Medicine
715 Albany Street
Boston, Massachusetts 02118
[5] Center for Morphometric Analysis
Neuroscience Center, MGH-East
Building 149, 13th Street
Charlestown, Massachusetts 02129
[6] Computer Science Department
University of Pennsylvania
111 Towne Building
220 South 33rd Street
Philadelphia, Pennsylvania 19104

Summary

We present a technique for automatically assigning a neuroanatomical label to each voxel in an MRI volume based on probabilistic information automatically estimated from a manually labeled training set. In contrast to existing segmentation procedures that only label a small number of tissue classes, the current method assigns one of 37 labels to each voxel, including left and right caudate, putamen, pallidum, thalamus, lateral ventricles, hippocampus, and amygdala. The classification technique employs a registration procedure that is robust to anatomical variability, including the ventricular enlargement typically associated with neurological diseases and aging. The technique is shown to be comparable in accuracy to manual labeling, and of sufficient sensitivity to robustly detect changes in the volume of noncortical structures that presage the onset of probable Alzheimer's disease.

Introduction

Neurodegenerative disorders, psychiatric disorders, and healthy aging are all frequently associated with structural changes in the brain. These changes can cause alterations in the imaging properties of brain tissue, as well as changes in morphometric properties of brain structures. Morphometric changes may include variations in the volume or shape of subcortical regions, as well as alterations in the thickness, area, and folding pattern of the cortex. While surface-based analyses that depend on models of the position and orientation of the cortical ribbon can provide an accurate assessment of cortical variability, volumetric techniques are required to detect changes in noncortical structures. For example, changes in ventricular or hippocampal volume are frequently associated with a variety of diseases (e.g., Puri et al., 1999; Killiany et al., 2000; Wolf et al., 2001). This type of analysis has commonly been accomplished by a having a trained anatomist or technician manually label some or all of the structures in the brain, a procedure that can take up to a week for high-resolution images. Here, we use the results of the manual labeling using the validated techniques of the Center for Morphometric Analysis (Caviness et al., 1989; Kennedy et al., 1989; Goldstein et al., 1999; Seidman et al., 1999) to automatically extract the information required for automating the segmentation procedure. The automated segmentation procedure requires approximately 30 min on current workstation hardware (e.g., 1 GHz Pentium III). The additional capability to run multiple processes in parallel enables the labeling of thousands of brains per day.

Typically, manual labeling of brain structures is accomplished using a variety of information including image intensities, global position within the brain, position relative to neighboring brain structures, as well as anatomical landmarks. The challenge in labeling brain structures based on MRI image intensities alone is illustrated in Figure 1, which shows the intensity histograms of nine different neuroanatomical structures defined by a manual segmentation procedure based on a typical T_1-weighted MRI image. Examining this figure, it is apparent why no global classification scheme can successfully distinguish structures from each other based only on intensity information—there is far too much overlap between the class distributions (even cortical gray matter and white matter overlap by more than 12%). While adding additional MRI sequences with differing contrast properties or different imaging modalities entirely can help separate the class distributions, spatial information is still required to disambiguate the classification problem.

The use of spatial information to aid in classification is facilitated by the construction of a *probabilistic atlas* (Collins et al., 1994; Fox et al., 1994; Mazziotta et al., 1995; Thompson et al., 1997). In this type of atlas, information regarding the statistical properties of anatomical structures is stored in a space in which coordinates have anatomical meaning as opposed to the somewhat arbitrary coordinates in a raw image, which are dependent on the position, orientation, and shape of a subject's head in the MR scanner. Spatial information can aid in classification in several ways: (1) the number of possible anatomical classes at a given global position in

[7] Correspondence: dale@nmr.mgh.harvard.edu

Principle of atlas-based segmentation :

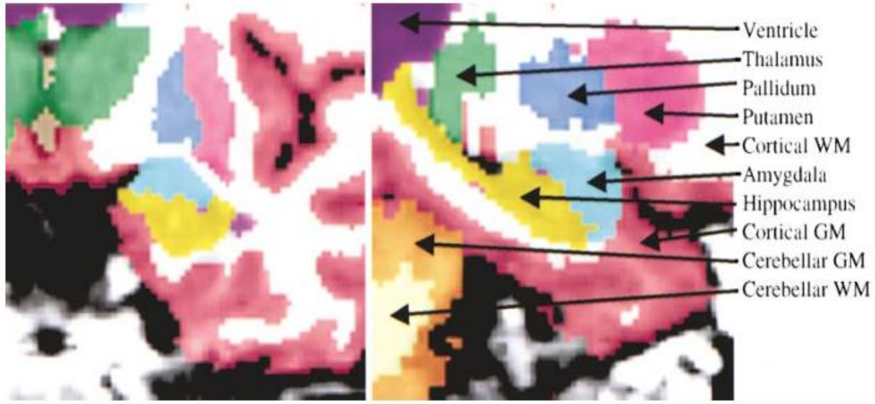
- Uses a probabilistic atlas that encodes statistical representations of anatomical structures
- Image Registration: Aligns MRI scans to the atlas reference space.

$$L = \arg\min_{L} \int (T(r) - I(Lr))^2 \, dr$$

- Voxel-Wise Labeling: Assigns tissue types based on Bayesian priors and intensity distribution models.

Methodology :

- Modelling brain structures using a Bayesian approach :
  - The objective is to compute the Maximum A Posteriori (MAP) estimate $p(W|I,L)$
- Correction using a Markov Random Field (MRF) model to account for spatial relationships between voxels, considering anatomical relationships (e.g., the amygdala is anterior to the hippocampus).



Limits :

- Registration errors can affect segmentation accuracy.
- Limited adaptability to inter-subject anatomical variability.

Multi-Atlas Label Fusion

- Uses multiple manually segmented atlases to enhance segmentation accuracy.
- Each atlas is individually registered to the target MRI, and labels are fused to generate a consensus segmentation.

Selection and preprocessing of atlases → **Atlas registration to the target image** → **Fusion of labels from the aligned atlases**


HHS Public Access
Author manuscript
IEEE Trans Pattern Anal Mach Intell. Author manuscript; available in PMC 2013 December 22.

Published in final edited form as:
IEEE Trans Pattern Anal Mach Intell. 2013 March ; 35(3): 611–623. doi:10.1109/TPAMI.2012.143.

Multi-Atlas Segmentation with Joint Label Fusion

Hongzhi Wang[1] [Member, IEEE], Jung W. Suh[2], Sandhitsu R. Das[1], John Pluta[3], Caryne Craige[4], and Paul A. Yushkevich[1] [Member, IEEE]
[1]Penn Image Computing and Science Lab, Department of Radiology, University of Pennsylvania
[2]HeartFlow, Inc
[3]Departments of Neurology and Radiology, University of Pennsylvania
[4]School of Medicine, Temple University

Abstract

Multi-atlas segmentation is an effective approach for automatically labeling objects of interest in biomedical images. In this approach, multiple expert-segmented example images, called *atlases*, are registered to a target image, and deformed atlas segmentations are combined using *label fusion*. Among the proposed label fusion strategies, weighted voting with spatially varying weight distributions derived from atlas-target intensity similarity have been particularly successful. However, one limitation of these strategies is that the weights are computed independently for each atlas, without taking into account the fact that different atlases may produce similar label errors. To address this limitation, we propose a new solution for the label fusion problem, in which weighted voting is formulated in terms of minimizing the total expectation of labeling error, and in which pairwise dependency between atlases is explicitly modeled as the joint probability of two atlases making a segmentation error at a voxel. This probability is approximated using intensity similarity between a pair of atlases and the target image in the neighborhood of each voxel. We validate our method in two medical image segmentation problems: hippocampus segmentation and hippocampus subfield segmentation in magnetic resonance (MR) images. For both problems, we show consistent and significant improvement over label fusion strategies that assign atlas weights independently.

Index Terms

multi-atlas label fusion segmentation; dependence; hippocampal segmentation

I. Introduction

Atlas-based segmentation is motivated by the observation that segmentation strongly correlates with image appearance. A target image can be segmented by referring to atlases, i.e. expert-labeled sample images. After warping the atlas to the target image via deformable registration, one can directly transfer labels from the atlas to the target image. As an extension, multi-atlas based segmentation makes use of more than one atlas to compensate

Author Manuscript

Label Fusion Methods

Majority Voting

Each atlas contributes equally

$$S(x) = \text{argmax}_l \sum_{i=1}^{n} \delta(S_i(x), l)$$

Limitation: Does not consider the reliability of the atlases (all atlases are assigned the same weight).

Adaptive Weighting

Each atlas is assigned a weight proportional to its local similarity with the target image.

<u>LWGaussian</u> :
Weights based on intensity differences.

$$m_j(x) = \frac{\exp(-D(F_j(x), T(x))/\sigma^2)}{\sum_{k=1}^{n} \exp(-D(F_k(x), T(x))/\sigma^2)}$$

Limitation: Does not account for correlated errors between atlases.

JLF - Joint Label Fusion
(Method proposed by the paper)

The main idea of Joint Label Fusion (JLF) is that some atlases may have similar errors, and these dependencies must be taken into

$$M_x(i,j) = \mathbb{E}[\delta_i(x)\delta_j(x) \quad \propto \left[\sum_{y \in \mathcal{N}(x)} |F_T(y) - F_i(y)||F_T(y) - F_j(y)|\right]^{\beta}.$$

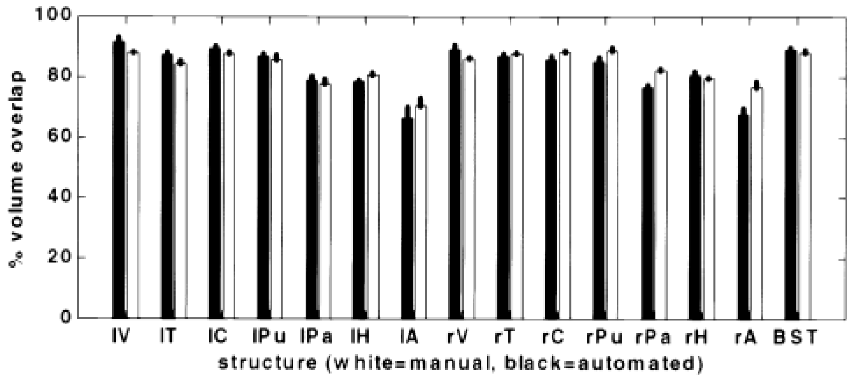
$$\mathbf{w}_x = \arg\min_{\mathbf{w}_x} \mathbf{w}_x^T M_x \mathbf{w}_x$$

$$\hat{S}_T(x) = \sum_{i=1}^{n} w_i(x) S_i(x)$$

Results and Performance of Segmentation Approaches

Evaluation of Atlas-Based Segmentation

- Automatic segmentation assigns 37 neuroanatomical labels with accuracy comparable to manual segmentation.



$$O(L1, L2) = \frac{V(L1 \cap L2)}{\frac{V(L1)+V(L2)}{2}} \times 100\%$$

- High overlap rate (~90% Dice Score) for subcortical structures such as the hippocampus and thalamus.
- Ability to detect subtle morphological changes related to neurodegenerative diseases (e.g., ventricular enlargement in Alzheimer's disease).
- Limitation: Sensitive to registration errors and anatomical variability across individuals.

Improvement with Multi-Atlas Label Fusion

- The fusion of multiple atlases reduces errors and enhances segmentation robustness.

Méthode	Average Dice Score (%)
Majority Voting (MV)	85.2
Local Weighted Voting - Gaussian (LWGaussian)	88.0
Joint Label Fusion (JLF - Proposé)	89.9

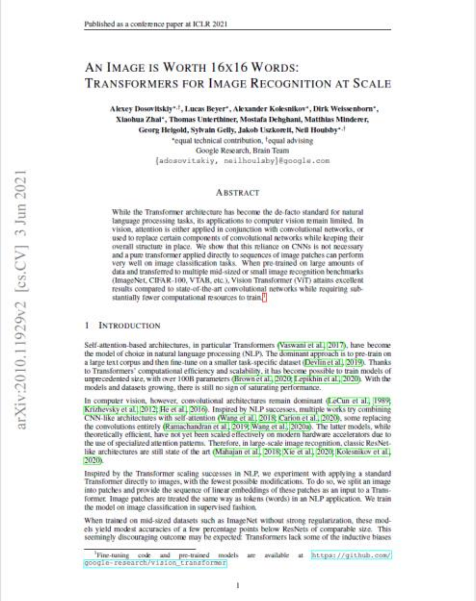
- Improved accuracy: +1.5% Dice Score for hippocampal segmentation compared to single-atlas methods.
- New adaptive weighting method accounts for correlated errors across atlases.
- Key result: Multi-atlas fusion improves segmentation accuracy while reducing dependence on a single reference atlas.

4

Deep Learning Models: Transformer-Based Approaches

(Papers 4 & 5)

Vision Transformer (ViT) - Principle and Architecture


Published as a conference paper at ICLR 2021

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*, Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†
*equal technical contribution, †equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com

arXiv:2010.11929v2 [cs.CV] 3 Jun 2021

ABSTRACT

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]

1 INTRODUCTION

Self-attention-based architectures, in particular Transformers (Vaswani et al., 2017), have become the model of choice in natural language processing (NLP). The dominant approach is to pre-train on a large text corpus and then fine-tune on a smaller task-specific dataset (Devlin et al., 2019). Thanks to Transformers' computational efficiency and scalability, it has become possible to train models of unprecedented size, with over 100B parameters (Brown et al., 2020; Lepikhin et al., 2020). With the models and datasets growing, there is still no sign of saturating performance.

In computer vision, however, convolutional architectures remain dominant (LeCun et al., 1989; Krizhevsky et al., 2012; He et al., 2016). Inspired by NLP successes, multiple works try combining CNN-like architectures with self-attention (Wang et al., 2018; Carion et al., 2020), some replacing the convolutions entirely (Ramachandran et al., 2019; Wang et al., 2020a). The latter models, while theoretically efficient, have not yet been scaled effectively on modern hardware accelerators due to the use of specialized attention patterns. Therefore, in large-scale image recognition, classic ResNet-like architectures are still state of the art (Mahajan et al., 2018; Xie et al., 2020; Kolesnikov et al., 2020).

Inspired by the Transformer scaling successes in NLP, we experiment with applying a standard Transformer directly to images, with the fewest possible modifications. To do so, we split an image into patches and provide the sequence of linear embeddings of these patches as an input to a Transformer. Image patches are treated the same way as tokens (words) in an NLP application. We train the model on image classification in supervised fashion.

When trained on mid-sized datasets such as ImageNet without strong regularization, these models yield modest accuracies of a few percentage points below ResNets of comparable size. This seemingly discouraging outcome may be expected: Transformers lack some of the inductive biases

[1]Fine-tuning code and pre-trained models are available at https://github.com/google-research/vision_transformer

1

- ➢ Inspired by Natural Language Processing (NLP) Transformers, ViT applies self-attention mechanisms to images without relying on convolutions.

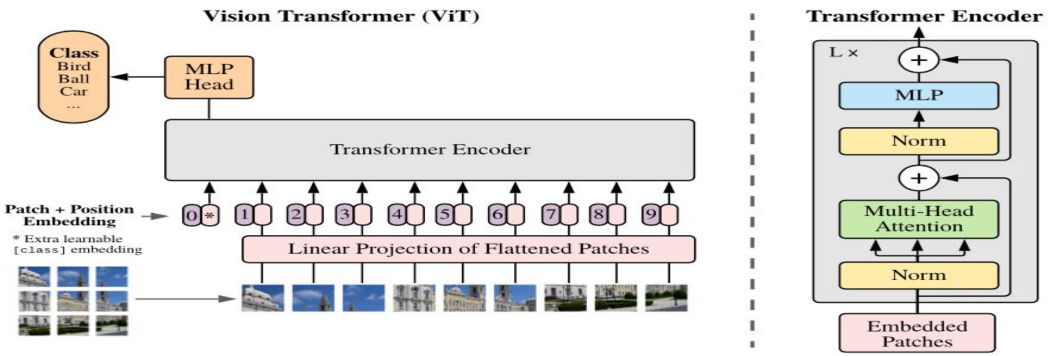
1. Splitting the image into N patches of size (P×P): $N = \dfrac{H \times W}{P^2}$.

2. **Patch Encoding**: each patch x_i is flattened and projected into an embedding vector.

3. Addition of the classification token and **positional** embeddings.

$$z_0 = \left[x_{\text{class}};\, x_p^1 E;\, x_p^2 E;\, \dots;\, x_p^N E\right]. \qquad z_0 = z_0 + E_{\text{pos}}$$

4. **Transformer Encoder**: Stacked L layers of Multi-Head Self-Attention (MSA) and Multi-Layer Perceptron (MLP).



$$z'_\ell = \mathrm{MSA}(\mathrm{LN}(z_{\ell-1})) + z_{\ell-1}, \qquad z_\ell = \mathrm{MLP}(\mathrm{LN}(z'_\ell)) + z'_\ell.$$

5. Final representation of the classification token (first position of z_L) is normalized to obtain the **prediction**.

- ➢ Limitation:
  - Performs well on large datasets (e.g., ImageNet-21k, JFT-300M), but struggles on small datasets.

Nested Transformer (NesT) – A Hierarchical Improvement


Nested Hierarchical Transformer: Towards Accurate, Data-Efficient and Interpretable Visual Understanding

Zizhao Zhang[1] Han Zhang[2] Long Zhao[2] Ting Chen[2] Sercan Ö. Arık[1] Tomas Pfister[1]

[1]Google Cloud AI [2]Google Research

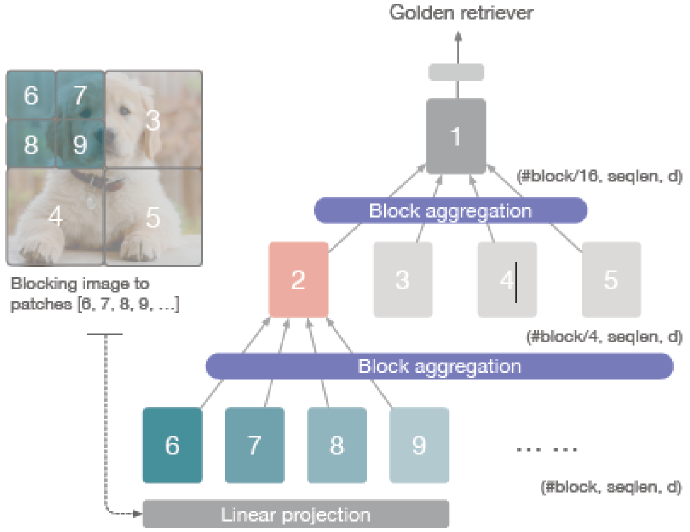
arXiv:2105.12723v4 [cs.CV] 30 Dec 2021

Abstract

Introduction

- NesT introduces a nested hierarchical organization to better capture spatial relationships.

1. Image partitioning into blocks (instead of individual patches).
2. **Local processing**: Each block is independently analyzed by a local Transformer.
3. **Hierarchical block fusion**: Gradual aggregation of blocks to capture global context.
   - Progressive aggregation using **3×3 convolutions** and **max-pooling**.
   - Ensures better integration of local and global information.
   - Each set of 4 neighboring blocks is merged into a higher-level block.
   - Gradually reduces the total number of blocks.



- Advantages over ViT:
  - Requires less data: Achieves better results on smaller datasets (CIFAR-10, ImageNet).
  - Faster training: Thanks to its nested hierarchical structure.

Results of Transformers in Segmentation

Arch. base	Method	C10 (%)	C100 (%)
Convolutional	Pyramid-164-48	95.97	80.70
	WRN28-10	95.83	80.75
Transformer full-attention	DeiT-T	88.39	67.52
	DeiT-S	92.44	69.78
	DeiT-B	92.41	70.49
	PVT-T	90.51	69.62
	PVT-S	92.34	69.79
	PVT-B	85.05*	43.78*
	CCT-7/3×1	94.72	76.67
Transformer local-attention	Swin-T	94.46	78.07
	Swin-S	94.17	77.01
	Swin-B	94.55	78.45
	NesT-T	96.04	78.69
	NesT-S	96.97	81.70
	NesT-B	**97.20**	**82.56**

NesT Performance on CIFAR-10/100

Arch. base	Method	#Params	Top-1 acc. (%)
Convolutional	ResNet-50	25M	76.2
	RegNetY-4G	21M	80.0
	RegNetY-16G	84M	82.9
Transformer full-attention	ViT-B/16	86M	77.9
	DeiT-S	22M	79.8
	DeiT-B	86M	81.8
Transformer local-attention	Swin-T	29M	81.3
	Swin-S	50M	83.0
	Swin-B	88M	83.3
	NesT-T	17M	81.5
	NesT-S	38M	83.3
	NesT-B	68M	**83.8**

Comparison on ImageNet dataset

	ViT-B/16	Swin-B	Nest-B
ImageNet Acc. (%)	84.0	86.0	**86.2**

Comparison on ImageNet benchmark with ImageNet-22K pre-training

- Data Efficiency: NesT outperforms ViT on small datasets like CIFAR-10 and CIFAR-100 due to its nested hierarchical architecture, which better captures local spatial relationships.

- Computational Cost: Unlike ViT, NesT is less resource-intensive and achieves high performance without requiring extensive pre-training on large-scale dataset

5

Conclusion

Summary of Approaches

- Manual segmentation is the reference method in neuroanatomy but is time-consuming and resource-intensive.

- Traditional methods (probabilistic atlas and multi-atlas fusion) enhance automation but are limited by anatomical variability and computational cost.

- Transformers (ViT, NesT) provide superior performance, with NesT outperforming ViT on small datasets while reducing computational costs.

Thank you for your attention!